BMC Plant Biology

RESEARCH
Open Access

# Prediction of new candidate proteins and analysis of sub-modules and protein hubs associated with seed development in rice (*Oryza sativa*) using an ensemble network-based systems biology approach

M. R. P. De Silva[1], J. W. J. K. Weeraman[1], S. Piyatissa[1] and P. C. Fernando[1*]

## Abstract

**Background** Rice is a critical global food source, but it faces challenges due to nutritional deficiencies and the pressures of a growing population. Understanding the molecular mechanisms and protein functions in rice seed development is essential to improve yield and grain quality. However, there is still a significant knowledge gap regarding the key proteins and their interactions that govern rice seed development. Protein–protein interaction (PPI) analysis is a powerful tool for studying developmental processes like seed development, though its potential in rice research is yet to be fully realized. With the aim of unraveling the protein interaction landscape associated with rice seed development, this systems biology study conducted a PPI network-based analysis. Using a list of known seed development proteins from the Gene Ontology (GO) knowledgebase and literature, novel candidate proteins for seed development were predicted using an ensemble of network-based algorithms, including Majority Voting, Hishigaki Algorithm, Functional Flow, and Random Walk with Restart, which were selected based on their popularity and usability. The predictions were validated using enrichment analysis and cross-checked with independent transcriptomic analysis results. The rice seed development sub-network was further analyzed for community and hub detection.

**Results** The study predicted 196 new proteins linked to rice seed development and identified 14 sub-modules within the network, each representing different developmental pathways, such as endosperm development and seed growth regulation. Of these, 17 proteins were identified as intra-modular hubs and 6 as inter-modular hubs. Notably, the protein SDH1 emerged as a dual hub, acting as both an intra-modular and inter-modular hub, highlighting its importance in seed development PPI network stability.

**Conclusions** These findings, including the identified hub proteins and sub-modules, provide a better understanding of the PPI interaction landscape governing seed development in rice. This information is useful for achieving a systems biology understanding of seed development. This study implements an ensemble of algorithms for the analysis and showcases how systems biology techniques can be applied in developmental biology.

**Keywords** PPI network, Network-based algorithms, Sub-modules, Hub proteins, Rice, Agronomics

*Correspondence:
P. C. Fernando
pasanfernando@pts.cmb.ac.lk

[1] Department of Plant Sciences, University of Colombo, Colombo 03, Sri Lanka

## Introduction

Rice (*Oryza sativa*), widely consumed as a staple food globally, holds a significant role in plant research being the model monocot organism. The small genome size of rice [1] has enabled in-depth exploration of various biological aspects aided by technological advancement. Rice seeds play an important role in plant growth, development, and propagation [2]. It is also a major nutrient supplier on a global scale. Therefore, grain yield and quality are important for rice agriculture. The growing population has created a substantial demand for high rice yields [3]. The global population is expected to exceed 8.5 billion by 2030 [4], which requires the development of high-yielding and nutritional rice varieties as a sustainable and economical strategy.

The mechanism of rice seed development is fragile and complicated. Starting from double fertilization, it takes about 20 days to produce mature seeds [5], during which three main phases are identified: embryo morphogenesis, endosperm filling, and seed maturation. Embryo and endosperm developments are concurrently harmonized to ensure proper grain development [6]. A comprehensive understanding of the molecular processes underlying seed development is currently lacking [5, 7]. Seed development is a complicated biological process that is regulated through an intricate cross-play of several molecular sub-pathways. The knowledge of how certain proteins, including transcription factors and hormones, interact within these pathways is not complete. Especially, the information about key central protein regulators or protein hubs underlying seed development is insufficient. This is mainly due to the lack of systems biology studies that investigate molecular interactions associated with seed development. To improve rice breeding, further insights into grain quality traits and evaluation tools are needed [8]. Therefore, this work attempts to unravel the protein–protein interaction (PPI) landscape that regulates seed development in rice. This will help generate information regarding the crucial role certain proteins play in seed growth and how they are sub-modularized in biochemical pathways. This information can be used to better understand molecular pathways and interactions governing the different stages of seed development.

Proteins and their interactions are particularly important in regulating developmental processes such as seed development. Typically, these developmental phenotypes involve multiple proteins interacting in intricate pathways identified as functional modules [9, 10]. PPI networks represent the protein interactions using computational graphs, which can be used for further analysis [11, 12]. Different computational algorithms are used to analyze and unravel these networks to produce useful information. For instance, network-based candidate protein prediction algorithms are widely used to predict new protein candidates for selected molecular functions and phenotypes [10, 13]. These algorithms are considered more accurate at predicting new proteins associated with development phenotypes than other sequence or structure-based protein prediction methods, such as BLAST and iterative threading assembly refinement [13–15]. For instance, a study [16] reports that STRING PPI network-based methods consistently outperform sequence-based protein function prediction methods, such as BLAST, achieving accuracy scores over 90% for predicting bacterial virulence factors. Moreover, another study [17] performed a comprehensive evaluation of sequence-based, structure-based, and network-based protein function prediction methods, which resulted in network-based methods outperforming the other two when predicting Gene Ontology-Biological Process (GO-BP) terms for all the manually-curated UniProt Knowledgebase/Swiss-Prot proteins. Alternatively, machine learning and deep learning models like DeepFRI have been used for protein function prediction, demonstrating better accuracy [18–20]. However, such accurate models rely on experimentally elucidated protein structures, which poses a challenge for plant protein function prediction, as the majority of plant proteins are not experimentally resolved. Data scarcity is a major challenge [21], especially for deep learning models, which can be problematic when applied to plant proteins. On the other hand, PPI networks for plants are readily available in the STRING database, making PPI network-based models more applicable for plants. Furthermore, deep learning models act as black boxes [22], where the underlying biological principle for protein function prediction is unclear, whereas PPI network-based methods use the guilt-by-association principle [12, 13], where known associations to annotated proteins are used for the function prediction of unannotated proteins. Particularly, network-based methods are ideal for predicting biological processes and developmental phenotypes because, usually, such phenotypes are regulated by various proteins with different sequences and structures; therefore, sequence and structure-based prediction methods may fail to predict important proteins with lower sequence and structural similarity to original proteins. Because the network-based algorithms follow a systems biology approach of using existing protein or gene interactions to predict new candidates, they capture information from proteins with different sequences and structures during the prediction. Among network-based algorithms, community detection algorithms are crucial for analyzing complex networks, such as PPI networks [23]. These algorithms aid in understanding the network organization by identifying network modules, usually associated with specific functions or phenotypes. For instance, the Louvain community detection algorithm, which was used in this work, is a hierarchical clustering method that uses modularity optimization to identify communities [24]. The algorithm's robustness has established it as a dependable option for analyzing PPI networks.

Another important application of PPI network analysis is the identification of hub proteins, i.e., highly connected proteins within PPI networks. These are key components in maintaining corresponding biological pathways [14, 25]. The central-lethality rule, a widely accepted concept in network biology, underscores the critical role of hubs in maintaining the network architecture in dictating biological function [26]. It stresses the pivotal role of hubs over non-hubs within the network, considering their substantial contribution to network organization. Frequently, these hubs serve as drug targets and commercially important genetic engineering and breeding targets [10, 27]. Two key types of hubs stand out: inter-modular hubs, which are nodes with dense connections facilitating interactions between distinct functional modules, and intra-modular hubs, which are nodes densely interconnected within a single functional module [25].

Function-specific PPI sub-network analysis is popular in other domains, such as cancer and vertebrate development [14, 28–31]. Despite the potential, there are relatively few applications in plant development. For instance, PPI network analysis has been applied to study root development in rice and to identify genetic factors associated with parthenocarpy in bananas [10, 32]. However, a PPI network analysis specifically focused on rice seed development is currently lacking. Hence, this study employed a network-based computational approach to study the protein interactome associated with rice seed development, which may uncover useful information about novel candidate proteins, sub-modules, and hubs. To our knowledge, this study represents the first comprehensive analysis of the PPI network landscape specifically associated with seed development. This involved predicting novel candidate proteins using an ensemble of network-based algorithms. The ensemble approach, a pivotal advancement in data mining and machine learning, combines multiple models into a unified entity, enhancing prediction accuracy and generalization [33]. To our knowledge, such an ensemble approach has not been used in previous network-based candidate protein predictions for biological phenotypes in plants. The next steps involved extracting the sub-network of rice seed development, identifying sub-modules within the extracted sub-network, and uncovering the hubs associated with these sub-modules. The results from this study represent newfound knowledge regarding the systems biology point of view of seed development in rice. After further experimental validation, these results could be valuable resources for researchers and breeders seeking to make targeted improvements in rice cultivars.

## Methods

### Data retrieval and preprocessing

A comprehensive PPI network of rice was obtained from the STRING database (version 12; September 2022; https://stringdb.org). The graph was generated using the NetworkX package (version 3.0) in Python (version 3.8) with a combined score cutoff of 0.7. This cutoff is recommended by the STRING database to retain only the PPIs with high confidence [34, 35]. The STRING database compiles PPIs using different methods, such as the yeast-two-hybrid method, mass spectrometry, and computational predictions, and it is important to remove spurious interactions and retain only the high-quality ones. For this purpose, a majority of studies use the 0.7 combined score cutoff recommended by the STRING database [30, 31], which was the reason for its selection for this work. After the application of the cutoff, duplicate interactions were eliminated and protein STRING IDs were converted into their preferred names before constructing the graph.

Seed proteins (proteins annotated to rice seed development) necessary to conduct the analysis were obtained through literature mining [5, 36–40] and Gene Ontology (GO) search using the QuickGo tool (version 2022–09-16; September 2022; https://www.ebi.ac.uk/QuickGO/). The literature mining focused on studies involving proteomic analysis, high-resolution QTL mapping, and transcriptomic research centered on proteins associated with rice grain development. For GO data, only seed proteins with evidence codes such as Inferred from Biological aspect of Ancestor (IBA), Inferred from Direct Assay (IDA), Inferred from Mutant Phenotype (IMP), and Inferred from Sequence or Structural Similarity (ISS) were considered, excluding those with Inferred from Electronic Annotation (IEA) due to lack of manual review. To validate predictions, differentially expressed proteins (DEPs) associated with seed development were collected from literature mining, including studies utilizing proteomics and transcriptomics experiments to analyse rice seed proteins [41, 42].

### Seed development sub-network extraction using an ensemble of network-based algorithms

An ensemble of network-based algorithms was employed to predict new seed development protein candidates [43]. To our knowledge, this study marks the pioneering use of a diverse ensemble of network-based algorithms in rice PPI network research. Usually, only one algorithm is used for candidate protein function prediction in previous studies [10, 44]. However, in this work, four algorithms, including Majority Voting (MV) [45], Hishigaki Algorithm (HA) [46], Functional Flow (FF) [47], and Random Walk with Restart (RWR) [48], were integrated using the Rule of Sum [33] to construct the ensemble. The rationale behind this approach was to lower the impact of individual algorithm biases and errors, leading to more accurate and robust predictions [49]. These algorithms were selected based on their popularity and accessibility. For instance, RWR is currently one of the most popular network-based protein function prediction algorithms applied in different domains [44, 50–52]. Also, it was ensured that all four selected algorithms only used PPI networks as inputs and did not use other data such as protein sequence and structure. This was done to expand the number of proteins that can be analysed during the analysis as the majority of rice proteins do not have an experimentally elucidated structures. Furthermore, this ensures exclusive focus on the PPI network data, which has been proven effective when predicting GO-BP terms, without interference from other type of data [13, 16, 17, 48].

### *Majority Voting (MV) algorithm*

The MV algorithm, which was the first in the ensemble, calculates the prediction score by counting the number of seed proteins that are direct neighbors to non-seed proteins [45]. To perform this calculation Eq. (1) was used, where for a protein with n neighbors, $x_i$ represents whether neighbor i is a seed protein (x = 1), or not (x = 0). However, a notable limitation of this approach is its exclusive focus on immediate neighbors. Also, it does not make the best use of the overall network structure, and it is biased towards highly annotated functions as it relies on the frequency of annotations to predict the function of proteins.

$$\text{Prediction score} = \sum\nolimits_{i=1}^{n} x_i \tag{1}$$

### *Hishigaki algorithm (HA)*

The HA [46], the second in the ensemble, utilizes Eq. (2) to compute the prediction score. This algorithm extends the MV concept in predicting protein functions by analyzing proteins within a specified radius. It utilizes the chi-squared test to mitigate the bias for overrepresented functional annotations, unlike MV.

$$\text{Prediction score} = \frac{\left(n_{f(u)} - e_f\right)^2}{e_f} \tag{2}$$

In Eq. (2), $n_{f(u)}$ represents the count of proteins with a specific function (f) within the n-neighborhood of the protein "u". Additionally, $e_f$ stands for the expected frequency for the particular function, which is calculated using Eq. (3) below.

$$e_f = \frac{tot_f n_u}{tot_n} \tag{3}$$

In Eq. (3), $tot_f$ represents the total count of proteins with a particular function in the entire network, $tot_n$ indicates the total number of proteins in the entire network, and $n_u$ stands for the total number of neighbors for the protein "$u$".

### *Functional Flow (FF) algorithm*

The FF, the third algorithm in the ensemble, broadens the concept of guilt by association with protein groups, regardless of whether they physically interact with each other [47]. This algorithm computes the prediction score for each non-seed protein by following a set of rules iteratively, corresponding to half the network diameter (Eq. 4).

$$R_t^a(u) = \begin{cases} \infty, & if\ u\ is\ annotated\ with\ a, \\ 0, & otherwise. \end{cases} \tag{4}$$

Equation (4) defines $R_t^a(u)$ as the fluid volume of node u for function "a" at time t. Only the nodes corresponding to seed proteins have infinite fluid reserves of function "a" at time 0. Considering the amount of flow received in and exited out of each node, the reservoir of each node is recomputed for each successive time step according to Eq. (5).

$$g_t^a(u,v) = \begin{cases} 0, & if\ R_{t-1}^a(u) < R_{t-1}^a(v), \\ \min\left(w_{u,v}, R_{t-1}^a(u)\frac{w_{u,v}}{\sum_{(u,v)\in E} w_{u,y}}\right), & otherwise. \end{cases} \tag{5}$$

In Eq. (5), $g_t^a(u,v)$ stands for the flow received by protein v from protein u for the function "a" at time t, $w_{u,v}$ is the weight of the edge between u and v proteins, $R_{t-1}^a(u)$ is the reservoir of u at time t-1 for the function "a", $R_{t-1}^a(v)$ is the reservoir of v at time t-1 for the function "a", and $\sum_{(u,v)\in E} w_{u,y}$ is the sum of weights of edges connecting the protein u. There is no flow between nodes at time 0. Always the flow is downhill adhering to the capacity constraints in Eq. 5. The maximum flow allowed between two proteins is corresponding to the weight of the connecting edge.

At the end of each iteration, the reservoir at each node is finalized using Eq. (6).

$$R_t^a(u) = R_{t-1}^a(u) + \sum_{v:(u,v)\in E} \left(g_t^a(v,u) - g_t^a(u,v)\right) \tag{6}$$

The term $\sum_{v:(u,v)\in E}\left(g_t^a(v,u) - g_t^a(u,v)\right)$ of Eq. 6 represents the net flow change due to each immediate neighbor of the protein. At the end of defined iterations, each protein is annotated with the functional score calculated as the total amount of flow received by the protein. Seven iterations were required to compute the functional score, which is half the network diameter.

### *Random Walk with Restart (RWR) algorithm*

The fourth algorithm of the ensemble was RWR. This method estimates the likelihood of a "walk" along random edges connecting various nodes, ultimately reaching a specific target node from a predefined set of starting nodes [48]. The random walk length is determined by the number of iterations. The restart parameter, also known as the learning parameter, regulates the probability of the random walk "jumping" back to its starting position midwalk. The resulting node weights represent the probability of the random walk landing on them when initiated from any of the seed nodes. RWR in matrix form is displayed in Eq. (7) below.

$$w_{i+1} = \propto w_0 + (1- \propto)AD^{-1}w_i \tag{7}$$

In this context, $w_0$ signifies the initial weight vector, $w_i$ represents the weights after the $i^{th}$ iteration, $w_{i+1}$ denotes the weights after the $(i+1)^{th}$ iteration, and α represents the learning parameter for the algorithm. Additionally, A and D refer to the adjacency and degree matrices of the PPI network graph, respectively.

Here, the initial weight of corresponding nodes in the PPI network was set using the seed proteins list, ensuring a weight of 100 to prioritize the random walker's inclination to visit these seed nodes. Network propagation underwent 5 iterations with a learning parameter (α) of 0.1, focusing on local interactions [53]. The iteration count was capped at 5 to include relevant neighborhoods containing other seed development-related proteins and prevent exclusion due to longer paths [54].

### *Constructing the ensemble model using the rule of sum*

To build the ensemble model following the calculation of prediction scores for non-seed proteins using the four selected algorithms, the scores were normalized using the min–max normalization method and were compared with the list of DEPs obtained through literature mining to calculate the validation score by DEPs [41, 42]. A non-seed protein was given a score of 1 if it was a DEP; otherwise, it was assigned a score of 0. The total prediction score for each non-seed protein was obtained by adding the normalized prediction scores from the four algorithms and the DEP validation score using the rule of sum as shown in Eq. (8) below [33].

$$\begin{aligned}\text{Total prediction score} = \text{NMV} + \text{NHA} + \text{NFF} \\ + \text{NRWR} + \text{DEP}\end{aligned} \tag{8}$$

Equation (8) represents the calculation of the total prediction score for each non-seed protein, where NMV denotes the normalized MV score, NHA denotes the normalized HA score, NFF denotes the normalized FF score, and NRWR denotes the normalized RWR score.

To evaluate the performance of the ensemble model, benchmarking was conducted against the four original algorithms using tenfold cross-validation and the original seed development protein list. The performance was assessed using three key metrics: area under the precision-recall curve (AUPR), area under the receiver operator characteristic curve (AUROC), and the $F_{max}$ score. AUPR and $F_{max}$ scores are widely used metrics in protein function prediction studies because of their focus on positive annotations and better evaluation performance [17]. AUPR is the plot between the precision (Eq. 9) and the recall (Eq. 10), and AUROC is the plot between the true positive rate (Eq. 10) and the false positive rate (Eq. 11). $F_{max}$ is the maximum F1-score calculated under different thresholds, which is based on precision and recall as shown in Eq. 12, where i indicates the threshold. Furthermore, the DeLong test [55] was performed to assess the statistical significance of AUROC value comparisons.

$$Precision = \frac{True\ positives}{(True\ positives + False\ positives)} \tag{9}$$

$$True\ positive\ rate, Recall = \frac{True\ positives}{(True\ positives\ +\ False\ negatives)} \tag{10}$$

$$False\ positive\ rate = \frac{False\ positives}{(False\ positives\ +\ True\ negatives)} \tag{11}$$

$$F_{max} = max\left\{2 \times \frac{precision_i \times recall_i}{precision_i + recall_i}\right\} \tag{12}$$

### *Optimal cutoff for top candidate selection*

Initially, the resulting total prediction scores for non-seed proteins were sorted in descending order to identify the top candidates. Then, to determine the optimal top N number of predictions cutoff, the precision top-N curve [56] method was adapted and modified. Here, the DEPs, which were generated from transcriptomic experiments associated with seed development, was used when calculating the precision [41, 42]. This provides an extra validation from an independent source. This modified precision at the $N^{th}$ position calculates the percentage of DEPs within the top N predictions (Eq. 13) against the number of top predictions (N).

$$\textit{Percentage of DEPs within the top N predictions} = \frac{\textit{DEPs} \cap \textit{Top N predictions}}{N} \times 100 \tag{13}$$

In Eq. (13), the percentage of DEPs within the top N predictions was calculated by plugging in the appropriate values for N and calculating the overlap between the list of DEPs and the top N predictions. For instance, if the top 100 predictions are considered ($N=100$), this will calculate the ratio of DEPs within the top 100 predictions, which will be taken as the modified precision. The optimal cutoff was determined by selecting the value of N that corresponds to the highest ratio of DEPs within the top N predictions. This ensures that the selected cutoff captures the highest number of DEPs while being validated through an external source.

After the predictions using the ensemble model, the seed development sub-network was extracted, which included the seed proteins and candidates filtered using the optimal cutoff and their interactions. Only the proteins that are also DEPs within the selected top N candidates were included to reduce false positive predictions.

### Validation of the predictions

For computational validation of predictions, the predicted protein candidates were subjected to enrichment analysis using the functional annotation tool in the Database for Annotation, Visualization and Integrated Discovery (DAVID) web application (Version: DAVID 2021; December 2022; https://david.ncifcrf.gov/tools.jsp). The Biological Process component of Gene Ontology (GO-BP) was used and terms with a p-value below 0.05 were selected [57]. Furthermore, literature searches were conducted for predicted proteins to find their potential associations with seed development.

### Identifying sub-modules related to seed development

The Louvain community detection algorithm was used to perform sub-module analysis on the seed development sub-network [24]. This algorithm has been widely applied in biological network studies, demonstrating promising outcomes [58]. The community sub package in NetworkX (version 3.0) was used to implement the Louvain community detection algorithm. Following the sub-module identification process, isolated sub-modules that consisted of proteins less than 6 were excluded from further analysis [59]. Subsequently, functional enrichment analysis for each detected sub-module was performed using the DAVID functional annotation tool (Version: DAVID 2021). This analysis aimed to annotate each sub-module with the most relevant GO-BP term associated with seed development, with a p-value less than 0.05.

### Detection and analysis of hub proteins

To detect intra-modular hubs, which are densely connected proteins within a sub-module, a Z-score for each node was calculated [60] using Eq. (14).

$$Z_i = \frac{ks_i - \mu k_{s_i}}{\sigma_{k_{s_i}}} \tag{14}$$

The Z-score for node i is represented by $Z_i$ in Eq. (14), while $ks_i$ denotes the within-modular degree (number of interactions) of node i within the module $s_i$. Moreover, $\mu k_{s_i}$ and $\sigma_{k_{s_i}}$ represent the mean and standard deviation of within-module degree in the module $s_i$, respectively. Nodes having Z-scores greater than or equal to 1.5 were identified as intra-modular hubs [61].

Centrality measures help identify important hubs in networks, but traditional methods often miss the modular structure [62]. For example, betweenness centrality does not differentiate between nodes that are hubs within a sub-module and those linking different sub-modules. The Partition Coefficient (PC), a widely used metric for assessing a node's role in a modular community structure [63, 64], measures the ratio of a node's connections within its own module to its total connections [63]. A high PC is almost always indicative of a node that plays a stronger role as a connector between modules than as a core part of a single module. Therefore, PC was used to detect inter-modular hubs in this study. To find inter-modular hubs, PC for each node was calculated using the following formula (Eq. 15).

$$P_i = 1 - \sum_{s=1}^{N} \left( \frac{k_{s_i}}{k_i} \right)^2 \tag{15}$$

In Eq. (15), $k_i$ represents the network degree of node i, N is the total number of modules, and $k_{s_i}$ is the within-modular degree of node i within the module s. The PC value varies between 0 and 1, reflecting a protein's involvement in either intra-modular or inter-modular interactions. A PC value closer to 0 signifies a higher proportion of intra-modular connections, while a PC closer to 1 indicates a higher proportion of inter-modular connections over intra-modular connections. Inter-modular hubs were selected based on a PC value greater than 0.5 [65].

Finally, structural and functional analyses were performed for each identified hub to further unravel useful information. Through the use of web-based tools for sequence analysis, the family, superfamily, and domains present in the predicted hubs were identified. Additionally, de novo structures predicted for the candidates were searched in the AlphaFold database [66, 67]. The web tools used, their corresponding URLs and versions, and the type of analysis performed are listed in Table 1.

### The bioinformatics pipeline

The bioinformatics pipeline used for this analysis is illustrated in Fig. 1. This pipeline was constructed in Python 3.8, using NetworkX (version 3.0) and community Python packages. For network visualizations, the Cytoscape software (version 3.9.0) was used. All the data and Python codes can be accessed at: https://github.com/rashpr88/RiceSeedDevelopment.

## Results

### Data preprocessing

The global PPI network of rice is an undirected graph consisting of 25,266 nodes and 7,767,680 interactions. From this interactome, 102 seed proteins associated

**Table 1** Web tools utilized for candidate hub analysis

| Web Tool | Web URL | Web Version | Type of Analysis |
|---|---|---|---|
| InterProScan | https://www.ebi.ac.uk/interpro/ | 93.0 | Protein family |
| Pfam | https://www.ebi.ac.uk/interpro/entry/pfam/ | 35.0 | Protein family and domains |
| Panther | http://www.pantherdb.org/ | 17.0 | Protein family and class |
| ScanProsite | https://prosite.expasy.org/scanprosite/ | 2023_01 | Protein domains |
| Supfam | https://supfam.org/ | 1.75 | Protein family |
| SMART | https://smart.embl.de/ | 9.0 | Protein domain |

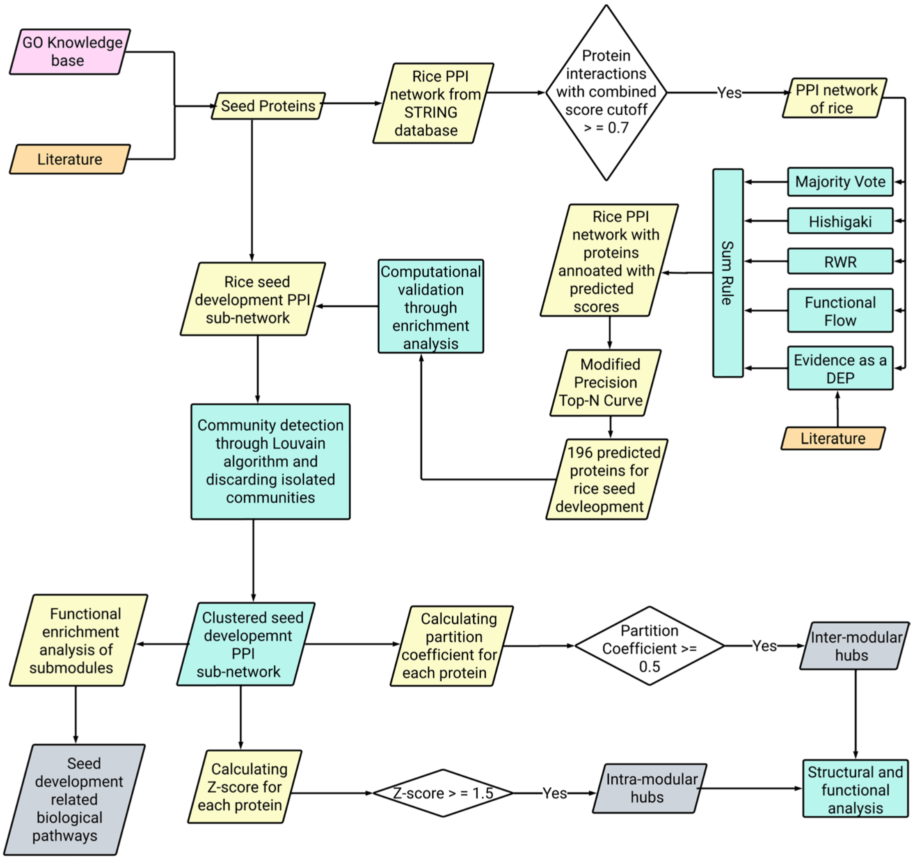
GO Knowledge base
Literature
Seed Proteins
Rice PPI network from STRING database
Protein interactions with combined score cutoff > = 0.7
Yes
PPI network of rice
Majority Vote
Hishigaki
RWR
Functional Flow
Evidence as a DEP
Literature
Sum Rule
Rice PPI network with proteins annoated with predicted scores
Modified Precision Top-N Curve
196 predicted proteins for rice seed devleopment
Computational validation through enrichment analysis
Rice seed development PPI sub-network
Community detection through Louvain algorithm and discarding isolated communities
Clustered seed developemnt PPI sub-network
Functional enrichment analysis of submodules
Seed development related biological pathways
Calculating partition coefficient for each protein
Partition Coefficient >= 0.5
Yes
Inter-modular hubs
Calculating Z-score for each protein
Z-score > = 1.5
Yes
Intra-modular hubs
Structural and functional analysis

**Fig. 1** The bioinformatics pipeline used for prediction and validation of protein candidates for seed development in rice. The hub protein and sub-pathway analysis procedures are also represented

with seed development were identified (Additional file 1: Table S1). Additionally, 640 DEPs linked to seed development were retrieved from literature mining (Additional file 1: Table S2). Filtering the PPI network with a combined score cutoff of 0.7 led to the inclusion of 95 out of the 102 seed proteins (Additional file 1: Table S1) and 285 out of the 640 DEPs. These filtered proteins were used for the predictions.

### Ensemble model construction and evaluation

The prediction performance of the ensemble model was compared against the four individual algorithms, i.e., MV, HA, RWR, and FF, while selecting the AUPR, AUROC, and $F_{max}$ score as the performance metrics. Table 2 shows the values obtained by each algorithm for each metric, and Fig. 2 represents Precision-Recall Curves and Receiver Operator Characteristic Curves for the five algorithms. According to the results, the ensemble model clearly outperformed all other algorithms for all three metrics. The AUROC results indicate that the ensemble model is well suited for distinguishing between positive and negative proteins associated with seed development compared to other individual algorithms. The DeLong test results for pairwise AUCROC comparisons (Table 3) clearly indicate that the increase in the AUROC value of the ensemble model is statistically significant at a 0.05 significance level. According to the results, only the ensemble model AUROC value shows a statistical significance in pairwise comparisons, demonstrating its superior accuracy. Moreover, the AUPR and $F_{max}$ scores indicate that the ensemble model predicts the true positive proteins associated with seed development better than others. Overall, these results clearly indicate that the ensemble model is better suited for predicting seed development proteins.

**Table 2** The area under the receiver operator characteristic curve (AUROC), the area under the precision-recall curve (AUPR), and $F_{max}$ score performance obtained by each algorithm, including Majority Voting (MV), Hishigaki Algorithm (HA), Random Walk with Restart (RWR), and Functional flow (FF), and the ensemble model

| Metric | MV | HA | RWR | FF | Ensemble |
|---|---|---|---|---|---|
| AUROC | 0.735 | 0.747 | 0.758 | 0.762 | 0.839 |
| AUPR | 0.759 | 0.780 | 0.759 | 0.758 | 0.859 |
| $F_{max}$ score | 0.706 | 0.706 | 0.750 | 0.727 | 0.800 |

### Seed development sub-network extraction

After using the ensemble of network-based algorithms for predicting novel candidates for seed development, a cutoff had to be applied to select the best candidates. Figure 3 illustrates the modified precision top-N curve generated to determine the cutoff for selecting the top predictions for further analysis.

Based on the curve analysis, the highest precision of 43.17% was achieved at 454 top-predicted proteins, which was chosen as the optimal cutoff. This resulted in 196 DEPs being selected for further analysis. According to Fig. 3, it was clear that the modified precision is sensitive to the top-N prediction value. It rapidly increases

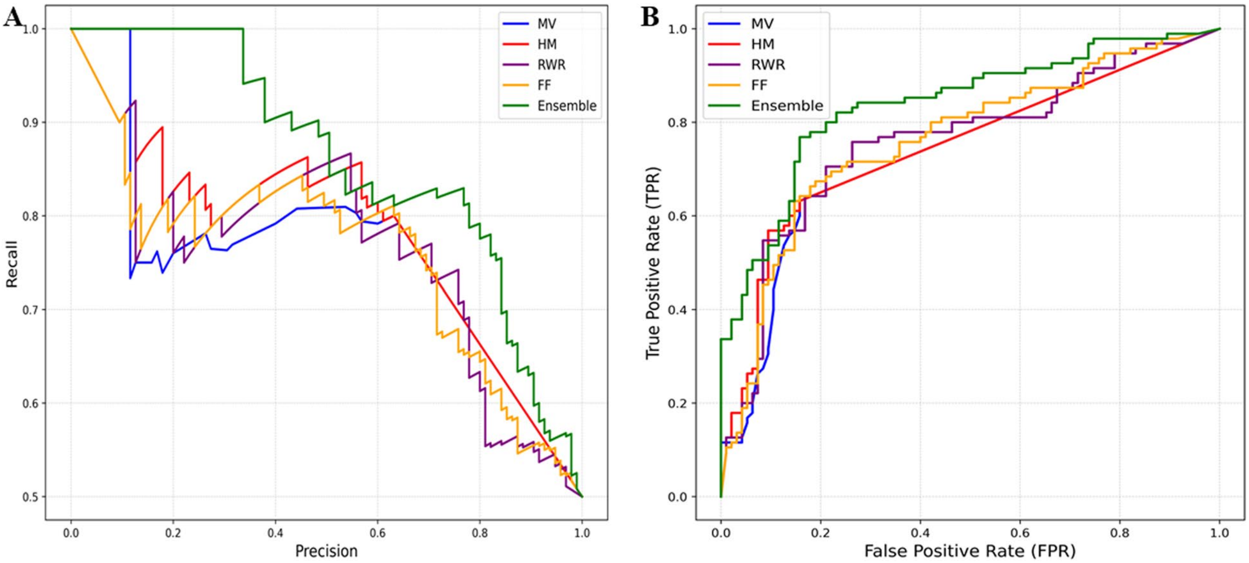

**Fig. 2** Precision-Recall Curves (**A**) and Receiver Operator Characteristic Curves (**B**) for the five algorithms, including Majority Voting (MV), Hishigaki Algorithm (HA), Random Walk with Restart (RWR), and Functional flow (FF), and the ensemble model. The ensemble model outperforms all the other algorithms on both metrics based on its higher area under the curve values

**Table 3** Pairwise comparisons of area under the receiver operating characteristic curve (AUROC) values using the DeLong test. The table presents p-values indicating the statistical significance at 0.05 level, denoted by an asterisk, for differences between AUROC of each algorithm, including Majority Voting (MV), Hishigaki Algorithm (HA), Random Walk with restart (RWR), Functional Flow (FF), and the ensemble model

| Algorithm 1 | Algorithm 2 | AUROC 1 | AUROC 2 | *p*-value |
|---|---|---|---|---|
| MV | Ensemble | 0.735401662 | 0.839445983 | 1.86E-05* |
| HA | Ensemble | 0.747202216 | 0.839445983 | 0.000229296* |
| RWR | Ensemble | 0.758227147 | 0.839445983 | 0.001056002* |
| FF | Ensemble | 0.762271468 | 0.839445983 | 0.001479837* |
| MV | HA | 0.73540166 | 0.74720222 | 0.30739538 |
| MV | RWR | 0.73540166 | 0.75822715 | 0.33336654 |
| MV | FF | 0.73540166 | 0.76227147 | 0.18088683 |
| HA | RWR | 0.74720222 | 0.75822715 | 0.62579809 |
| HA | FF | 0.74720222 | 0.76227147 | 0.53624514 |
| RWR | FF | 0.75822715 | 0.76227147 | 0.81182089 |

until 454 and decreases when the top number of predictions further increases. Therefore, selecting the top-N prediction value with the highest modified precision was crucial to the network extraction. The final seed development sub-network contained a total of 291 proteins, including 95 seed proteins directly annotated to seed development from literature, and 196 proteins predicted by the ensemble of network-based algorithms and also validated as DEPs.

### Computational validation of the predicted protein candidates

A summary of the GO-BP terms enriched among the seed proteins is provided in Additional File 1: Table S3. The top-enriched GO-BP term for seed proteins is "reproductive process" (GO:0022414). Given that seed development is a critical aspect of plant reproduction [68], this further confirms the relevance of the seed proteins selected for the analysis. The GO-BP terms from the gene enrichment analysis for predicted proteins were compared with those of seed proteins to assess their functional similarity to well-established proteins in seed development, thereby confirming the accuracy of the predictions. Table 4 presents the top 10 overlapping GO-BP terms, with "organic hydroxy compound metabolic process" (GO:1901615) being the most significant, as indicated by its p-value. Organic hydroxy compounds in rice seeds are crucial for growth, maturation, antioxidant defense, and disease resistance [69]. Phenolic acids and flavonoids protect seeds from oxidative damage by scavenging free radicals and preserving lipids, proteins, and nucleic acids, thus ensuring seed quality [70]. Plant hormones like abscisic acid (ABA) and gibberellins (GA),

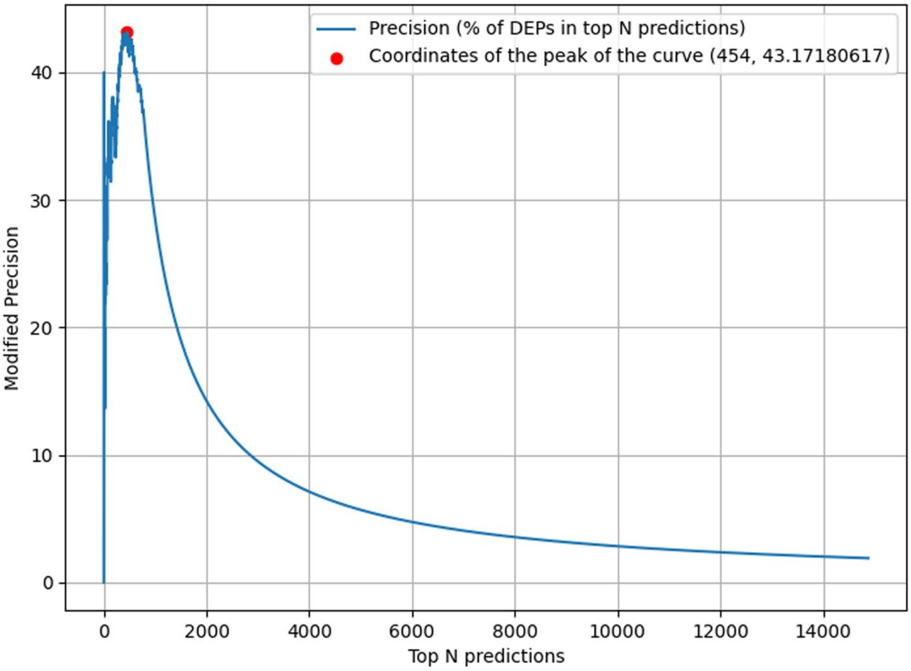

**Fig. 3** Modified precision top-N curve for protein predictions related to rice seed development, obtained from the rice protein–protein interaction (PPI) network. The x-axis represents the number of top-ranked predicted proteins (N), while the y-axis shows the modified precision, calculated as the percentage of known differentially expressed proteins (DEPs) among the top N predictions. The curve peaks at N = 454, where a maximum precision of 43.17% was achieved

**Table 4** Enriched Gene Ontology-Biological Process (GO-BP) terms that are common for both predicted proteins for seed development and original seed development proteins retrieved for the analysis in rice

| GO BP Term | *p*-value |
|---|---|
| GO:1,901,615 ~ organic hydroxy compound metabolic process | 8.30126E-12 |
| GO:0071396 ~ cellular response to lipid | 1.86946E-11 |
| GO:1,901,701 ~ cellular response to oxygen-containing compound | 2.55002E-10 |
| GO:0009755 ~ hormone-mediated signaling pathway | 5.91632E-10 |
| GO:0032870 ~ cellular response to hormone stimulus | 8.46041E-10 |
| GO:0033993 ~ response to lipid | 8.75219E-10 |
| GO:0071495 ~ cellular response to endogenous stimulus | 9.4022E-10 |
| GO:0070887 ~ cellular response to chemical stimulus | 9.91493E-10 |
| GO:1,901,700 ~ response to oxygen-containing compound | 3.14502E-09 |
| GO:0009725 ~ response to hormone | 1.15462E-07 |

both hydroxy compounds, are essential for seed growth, dormancy, and stress tolerance [71]. Likewise, the metabolism of hydroxy compounds is vital for rice seed development, emphasizing the importance of predicted proteins in facilitating this process.

Furthermore, the predicted protein candidates were validated using data from independently conducted transcriptomic experiments [41, 42]. According to this analysis, all predicted proteins were also confirmed as DEPs, exhibiting significant changes in gene expression associated with seed development. This provides an extra layer of validation for the predicted proteins. Furthermore, literature searches were conducted for the predicted proteins, and according to the results, although predicted proteins did not have direct associations with seed development, there was evidence that they were indirectly related. Table 5 displays information on such five key predicted proteins.

### Sub-modules related to seed development

After the module partitioning process of the extracted network conducted by the Louvain community detection algorithm, modules with more than five nodes were kept [59]. This resulted in fourteen sub-modules, which are depicted in Fig. 4.

Additional file 1: Table S4 presents the results of the functional enrichment analysis for sub-modules, including only the GO-BP terms with a p-value below 0.05. Each sub-module is associated with the most relevant GO-BP term related to seed development, which is listed in Table 6.

Additional file 1: Table S5 provides statistical information regarding the total number of proteins, seed proteins, predicted proteins, and the most relevant enriched GO-BP term for each of the 14 analyzed sub-modules.

When considering the enrichment results, certain sub-modules exhibit significant enrichment in seed development-specific pathways, such as endosperm development (sub-module 5), and regulation of seed growth (sub-module 13), while others are enriched for more general pathways, such as translation, cell differentiation, and glycolytic process, associated with seed development.

### Hub analysis

Hub analysis identified 17 intra-modular hubs and 6 inter-modular hubs. Figure 5 illustrates the hubs detected in the seed development sub-network.

#### *Intra-modular hubs*

Table 7 presents the 17 intra-modular hubs detected based on Z-score values. There were 9 seed proteins and 8 predicted proteins among the intra-modular hubs.

**Table 5** Information on five predicted proteins, indicating their known functions and potential functions associated with seed development

| Predicted Protein | Known Function | Potential Function in Rice Seed Development |
|---|---|---|
| ATP synthase subunit gamma, mitochondrial (OS10T0320400-01) | Participates in light-dependent photosynthesis by utilizing the proton motive force across the thylakoid membrane to synthesize ATP from ADP and inorganic phosphate [72] | May play a role in regulating biosynthetic pathways associated with seed development by enhancing energy availability and maintaining internal oxygen levels through photosynthesis [73] |
| Os03g0278900 protein (OS03T0278900-01) | Facilitates proton translocation during ATP biosynthesis [74] | Could contribute to providing the primary energy source for cellular processes during seed development [75] |
| ABI5 | Abscisic acid (ABA)-mediated transcription [76] | Could be involved in seed maturation due to its role in abscisic acid (ABA) signaling [77] |
| RPL3B | Ribosomal biogenesis [78] | Mutation of the RML1 gene encoding RPL3B resulted in reduced seed size, suggesting it could play a role in controlling seed size [78] |
| HMA4 | Transport Copper [79] | Prevents the build-up of copper in rice grains [79] |

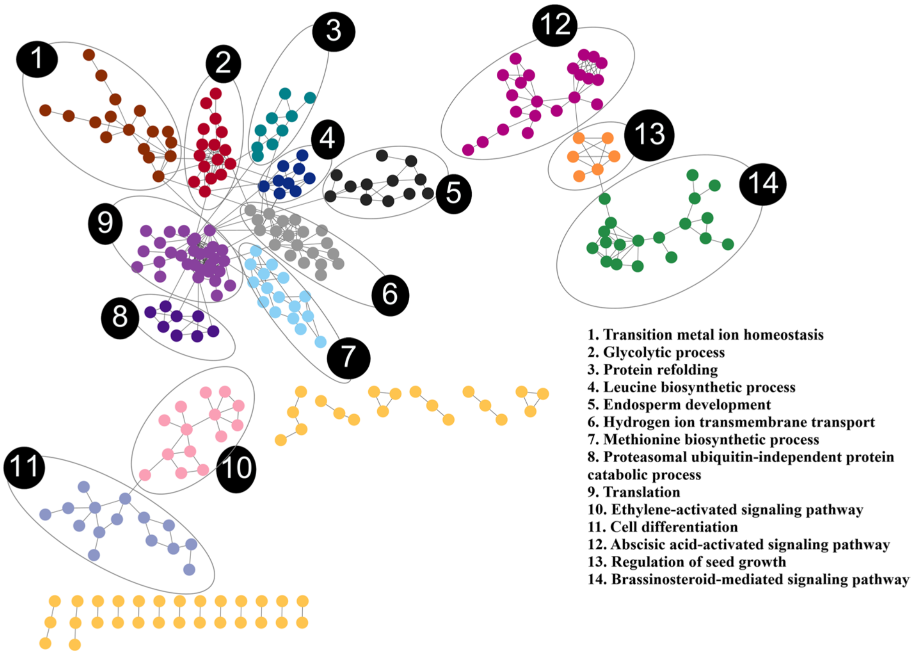
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1. Transition metal ion homeostasis
2. Glycolytic process
3. Protein refolding
4. Leucine biosynthetic process
5. Endosperm development
6. Hydrogen ion transmembrane transport
7. Methionine biosynthetic process
8. Proteasomal ubiquitin-independent protein catabolic process
9. Translation
10. Ethylene-activated signaling pathway
11. Cell differentiation
12. Abscisic acid-activated signaling pathway
13. Regulation of seed growth
14. Brassinosteroid-mediated signaling pathway

**Fig. 4** A visual representation of the seed development protein–protein interaction (PPI) sub-network of rice, with different colors assigned to the detected sub-modules, based on the Louvain community detection algorithm. The sub-modules are numbered and the most relevant Gene Ontology-Biological Process (GO-BP) term related to seed development according to enrichment analysis results are listed in the legend

These predicted proteins are also DEPs associated with seed development according to literature but do not have any other experimental evidence for seed development. They serve as excellent candidates for further wet lab validation studies related to seed development.

### *Inter-modular hubs*

Table 8 shows the 6 inter-modular hub proteins identified based on PC calculation. This included 4 seed proteins and 2 predicted proteins. The SDH1 protein was identified as both an intra and inter-modular hub.

## Discussion

### Analysis of detected seed development sub-modules and intra-modular hubs

The Louvain community detection algorithm uncovered 14 sub-modules associated with seed development, unraveling the PPI network landscape underlying rice seed growth. A comprehensive analysis of the 14 sub-modules and their intra-modular hubs can be found in Additional file 1: Table S6. Certain sub-modules, such as endosperm development (sub-module 5) and regulation of seed growth (sub-module 13), were directly associated with seed development, while others, such as transition metal ion homeostasis (sub-module 1), were indirectly influencing seed development. Seed development is a complex developmental process that has several intertwined sub-pathways. Therefore, it is expected to see a combination of general pathways, such as glycolysis (sub-module 2), cell differentiation (sub-module 11), and protein refolding (sub-module 3), and seed development-specific pathways, such as endosperm development (sub-module 5) and regulation of seed growth (sub-module 13), in the seed development sub-network. For instance, seed development is a highly energy-demanding process [75], requiring rapid cell growth [90], differentiation, and energy production. Therefore, it is anticipated that proteins associated with such general pathways have a significant impact on seed development. Furthermore, proteins that are specifically associated with processes, such as endosperm development and seed growth, are also expected to have a crucial role in regulating seed development. This PPI analysis represents the comprehensive landscape of how these different sub-pathways and associated proteins interact to regulate seed development.

Of the directly associated sub-modules related to seed development, sub-module 13 regulates rice seed growth, influencing grain size, weight, and quality. Crucial hub proteins of this sub-module include ILI5, instrumental in controlling grain length, and APG, the module hub responsible for regulating cell division and carbohydrate metabolism [92, 93]. Additionally, predicted proteins GAI and Os03g0639300 protein (OS03T0639300-01) exhibit significant interactions within the module, suggesting their potential roles in rice seed growth,

**Table 6** Most relevant Gene Ontology-Biological Process (GO-BP) terms assigned to each sub-module of the seed development subnetwork of rice

| Sub-Module | Term Identifier | Enriched GO-BP Term | Functional Description |
|---|---|---|---|
| 1 | GO:0055076 | Transition metal ion homeostasis | The pathway that regulates and maintains consistent levels of transition metal ions inside seeds [80] |
| 2 | GO:0006096 | Glycolytic process | The chemical reactions and pathways that generate pyruvate, ATP, and reducing agents. These serve as sources of carbon, energy, and reducing power for biosynthesis [81] |
| 3 | GO:0042026 | Protein refolding | The process of repairing damaged proteins into functional configurations [82] |
| 4 | GO:0009098 | Leucine biosynthetic process | The chemical processes and pathways involved in the formation of leucine, which is used to synthesize seed-storage proteins [83] |
| 5 | GO:0009960 | Endosperm development | The progression of endosperm development through nuclear division, cellularization, cell differentiation, and reserve accumulation [84] |
| 6 | GO:1902600 | Hydrogen ion transmembrane transport | The controlled movement of protons across a membrane, which is essential for importing organic and inorganic nutrients into seeds [85] |
| 7 | GO:0009086 | Methionine biosynthetic process | Biochemical processes to synthesize methionine essential for producing key metabolites [86] |
| 8 | GO:0010499 | Proteasomal ubiquitin-independent protein catabolic process | The proteasome-mediated hydrolysis of peptide bonds to break down proteins or peptides, excluding ubiquitin involvement, primarily targets damaged and short-lived regulatory proteins [87] |
| 9 | GO:0006412 | Translation | The process in cells that creates proteins, supporting cell division, growth, and reserve storage [88] |
| 10 | GO:0009873 | Ethylene-activated signaling pathway | The chain of signals triggered by ethylene reception that controls a cellular process [89] |
| 11 | GO:0030154 | Cell differentiation | The process by which an unspecialized cell acquires specific structures and functions to become a specialized cell, giving rise to all the cell types and tissues of a seed [90] |
| 12 | GO:0009738 | Abscisic acid-activated signaling pathway | The regulation of a cellular process through a series of signals triggered when the plant hormone abscisic acid (ABA) binds to a receptor [71] |
| 13 | GO:0080113 | Regulation of seed growth | Any process that controls the frequency, pace, or scale of seed growth [42] |
| 14 | GO:0009742 | Brassinosteroid-mediated signaling pathway | The series of signals triggered by the detection of brassinosteroid [91] |

warranting further investigation. The sub-module 5, which governs endosperm development, is the other sub-module directly associated with seed development. Endosperm development is crucial for rice grain development as it provides essential nutrients. This analysis revealed the presence of important proteins, such as OsFIE2 (OsJ_25971) and METB1 within the module, that are crucial for endosperm development in rice grains [94, 95].

Of the indirect associations to seed development, submodule 1 is annotated to essential transition metal ion homeostasis, crucial for enriching rice with iron (Fe) and zinc (Zn). Predicted proteins such as HMA2 within this module play pivotal roles in metal transport and contribute to grain weight [96]. Interestingly, predicted proteins HMA4 and copper chaperone emerge as module hubs (Additional file 1: Table S6), with HMA4 notably linked to copper accumulation, suggesting potential roles in grain development [79]. Copper chaperone, a member of the HMA domain-containing protein family, is essential for metal ion homeostasis and detoxification. While it shows differential expression during grain development, its specific role remains unclear due to limited literature evidence.

The sub-module 2 is annotated to the glycolytic process, which provides energy and carbon skeletons for cellular metabolism [97]. Phosphoglycerate kinase (PGK) serves as the central hub in sub-module 2, and according to literature, increased PGK expression in transgenic rice seeds raised the level of pyruvate, which elevated carotenoids levels in rice seeds [98]. Therefore, the expression of PGK may play a pivotal role in developing high-carotenoid rice seeds in combination with other necessary enzymes to address human dietary needs. Sub-module 3 of the seed development subnetwork is associated with protein refolding, which is

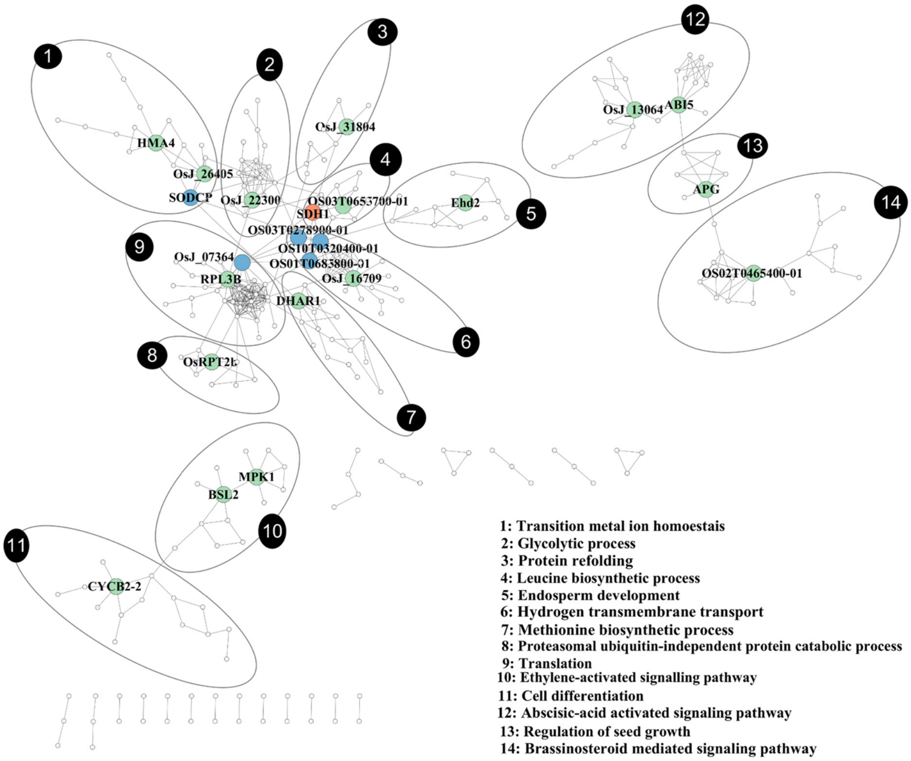
1
2
3
4
5
6
7
8
9
10
11
12
13
14
HMA4
OsJ_26405
SODCP
OsJ_22300
OsJ_31804
OS03T0653700-01
SDH1
OS03T0278900-01
OS10T0320400-01
OS01T0683800-01
OsJ_16709
OsJ_07364
RPL3B
DHAR1
OsRPT2b
Ehd2
OsJ_13064
ABI5
APG
OS02T0465400-01
MPK1
BSL2
CYCB2-2
1: Transition metal ion homoestais
2: Glycolytic process
3: Protein refolding
4: Leucine biosynthetic process
5: Endosperm development
6: Hydrogen transmembrane transport
7: Methionine biosynthetic process
8: Proteasomal ubiquitin-independent protein catabolic process
9: Translation
10: Ethylene-activated signalling pathway
11: Cell differentiation
12: Abscisic-acid activated signaling pathway
13: Regulation of seed growth
14: Brassinosteroid mediated signaling pathway

**Fig. 5** Network visualization of hub proteins (highly connected proteins within the network) in the seed development protein–protein interaction (PPI) sub-network of rice. Node size and color indicate the type of hub classification: larger green nodes, such as Ehd2, represent intra-modular hubs (proteins highly connected within a single module), identified using the Z-score method; larger blue nodes, such as SODCP, represent inter-modular hubs (proteins with high connectivity between modules), identified using the partition coefficient (PC). The SDH1 protein, which functions as both an intra- and inter-modular hub, is shown as a larger orange node. Smaller white nodes represent non-hub proteins (proteins with low connectivity). Sub-modules, identified through the Louvain community detection algorithm, are numbered, and the most significantly enriched Gene Ontology-Biological Process (GO-BP) terms derived from functional enrichment analysis are listed in the legend

crucial during rice grain development to ensure the correct folding and assembly of newly synthesized proteins in the endoplasmic reticulum, before being transported to their final destination [99]. Mitochondrial chaperonin-60 was identified as an intra-modular hub in sub-module 3. It is a heat shock protein that can stabilize and refold proteins during high temperatures and protects rice grain storage components, including starch, proteins, and RNA, against detrimental effects [100].

The sub-module 4 of the rice seed development sub-network is associated with the leucine biosynthetic process. Rice grain's essential amino acid levels, such as the leucine level, are important for its nutritional quality, but the mechanisms behind their accumulation are not yet fully understood [101]. Notably, 3-Isopropylmalate dehydrogenase, a key enzyme in leucine biosynthesis [102], acts as the intra-modular hub, supporting the module's function. However, further studies are required to unravel the potential interplay between the leucine biosynthetic pathway and rice grain development.

Hydrogen ion transmembrane transport (sub-module 6) is crucial for rice grain development as it maintains proper pH levels and ion homeostasis within cells. Various proton pumps and transporters are responsible for this process that may be associated with seed development [103]. The protein inorganic disphosphatase (OsJ_16709), which was revealed as a hub within the sub-module 6, belongs to the Inorganic Pyrophosphatase protein family. This protein plays a critical role in the metabolism of phosphate-containing compounds by hydrolyzing diphosphate in the presence of water. Although its involvement in H + ion transmembrane transport is uncertain, the hydrolysis reaction of diphosphate releases H + ions that can participate in the electrochemical gradient driving ion transport across biological membranes.

Sub-module 7 is associated with the methionine biosynthetic process, which is vital for plant growth. One of its module hubs, DHAR1, is known to increase rice grain yield and biomass when overexpressed. It also improves

**Table 7** Identified intra-modular hub proteins along with their respective network degree, module degree, Z-scores, the status as either a seed protein or a prediction, and their associated sub-module within the seed development sub-network

| Node | Network Degree | Module Degree | Z-score | Status | Sub-Module |
|---|---|---|---|---|---|
| DHAR1 | 7 | 6 | 2.74563 | Seed | 7 |
| Inorganic disphosphatase (OsJ_16709) | 14 | 14 | 2.65878 | Seed | 6 |
| CYCB2-2 | 5 | 5 | 2.6365 | Seed | 11 |
| HMA4 | 7 | 7 | 2.56946 | Predicted protein | 1 |
| Copper chaperone (OsJ_26405) | 7 | 7 | 2.56946 | Predicted protein | 1 |
| ABI5 | 8 | 7 | 2.36643 | Predicted protein | 12 |
| Alcohol dehydrogenase (OsJ_13064) | 7 | 7 | 2.36643 | Predicted protein | 12 |
| MPK1 | 6 | 6 | 2.31125 | Seed | 10 |
| Sterol delta-7-reductase (OS02T0465400-01) | 7 | 7 | 2.22752 | Predicted protein | 14 |
| RPL3B | 22 | 21 | 2.1221 | Predicted protein | 9 |
| OsRPT2b | 6 | 5 | 2.06474 | Predicted protein | 8 |
| APG | 6 | 5 | 1.73205 | Seed | 13 |
| Mitochondrial chaperonin-60 (OsJ_31804) | 4 | 4 | 1.72532 | Seed | 3 |
| BSL2 | 5 | 5 | 1.6641 | Predicted protein | 10 |
| 3-Isopropylmalate dehydrogenase (OS03T0655700-01) | 5 | 5 | 1.58114 | Seed | 4 |
| SDH1 | 10 | 5 | 1.58114 | Seed | 4 |
| Phosphoglycerate kinase (OsJ_22300) | 11 | 10 | 1.57263 | Seed | 2 |

**Table 8** The detected inter-modular hub proteins with their associated partition coefficient (PC) scores, status as a seed protein or a prediction, the module to which they belong, and the connecting modules in the seed development sub-network

| Node | PC | Status | Sub-Module | Connecting sub-modules |
|---|---|---|---|---|
| SDH1 | 0.68 | Seed | 4 | 2, 3, 4, 6, 9 |
| SODCP | 0.666667 | Seed | 1 | 1, 2, 6, 7 |
| ATP synthase subunit beta (OS01T0685800-01) | 0.592593 | Seed | 6 | 4, 6, 9 |
| ATP synthase subunit gamma, mitochondrial (OS10T0320400-01) | 0.579882 | Predicted protein | 6 | 1, 4, 5, 6, 9 |
| Os03g0278900 protein (OS03T0278900-01) | 0.56 | Predicted protein | 6 | 2, 6, 9 |
| Translation elongation factor (OsJ_07364) | 0.512397 | Seed | 9 | 4, 6, 9 |

the Ascorbic Acid and redox homeostasis [104]. The sub-module 8 of the sub-network is associated with the proteasomal ubiquitin-independent protein catabolic process. The regulation of cellular transitions is largely governed by protein turnover, with ubiquitin-independent mechanisms serving as effective means for identifying proteins for degradation [105]. Regulating the abundance of specific proteins and maintaining protein homeostasis is crucial in rice seed development, and protein turnover plays a vital role in this process. It is especially important during the transition between different growth stages. OsRPT2b, a member of the 26S proteasome regulatory subunit P45-like protein family, was revealed as an intra-modular hub in sub-module 8. This protein is crucial for the 26S proteasome regulation, responsible for the degradation of intracellular proteins in the ubiquitin–proteasome system, and exhibits variable expression levels during rice grain development [100]. However, further experimental evidence is needed to confirm its involvement in seed development.

The sub-module 9 of the network is associated with translation, which is crucial for rice grain protein synthesis. RPL3B, a predicted protein, was revealed as an intra-modular hub in this sub-module. It belongs to the ribosomal protein L3 family and its mutations impact ribosome biogenesis in rice, which results in abnormal architecture [78]. This predicted hub is an ideal candidate for further studies focusing on improving rice yield after experimental validation. The ethylene-activated signaling pathway is associated with sub-module 10 in this study. Ethylene is a gaseous hormone in plants that plays a pivotal role in multiple stages of rice grain development, such as germination, seedling growth, and grain maturation [89]. The protein MPK1 was detected as an intra-modular hub in this sub-module. It is a member of the mitogen-activated protein kinase (MAPK) family of proteins and plays a vital role in intracellular signaling pathways. This protein family responds to various stimuli and regulates essential cellular processes, including cell proliferation, differentiation, apoptosis, and stress response [106]. Through the analysis of CRISPR-edited mutants, a study has uncovered the significant role of MPK1 in rice development [107]. Although its involvement with ethylene-signaling pathways is poorly understood, it closely interacts with ethylene-related signaling proteins such as OsEIL2 and EIA1 [108], garnering further investigations into this module hub.

The cell differentiation is linked to sub-module 11, and it is essential for the growth and development of rice grains as it contributes to the architecture of the grain. The duration of cellularization in rice typically spans a period of 3 to 5 days [84]. The proper timing of cellularization in rice is dependent on the correct expression of various genes involved in the cell cycle, including CYCB2, which belongs to the cyclin protein family. In this study, CYCB2 was revealed as an intra-modular hub within the sub-module 11. Cyclins are well known for their involvement in regulating cell division and differentiation via cyclin-CDK complex control. CYCB2 has also been implicated in regulating endosperm development and seed size [109].

The sub-module 12 in this analysis is associated with the Abscisic acid-activated signaling pathway and ABI5 serves as an intra-modular hub which was predicted during the analysis. It plays a crucial role in multiple functions such as grain maturation, vigor, and dormancy, predominantly through regulating ABA-mediated transcription [76]. The sub-module 14 of the rice seed development sub-network is annotated to the brassinosteroid (BR)-mediated signaling pathway. Several studies have elaborated on the mediation of BRs in rice grain development. One study particularly emphasized that enhancing BR biosynthesis can boost crop productivity [110]. Additionally, it was observed that rice plants deficient in or insensitive to BRs produced smaller, shorter seeds [111]. Sterol delta-7-reductase (OS02T0465400-01) was identified as an intra-modular hub of this sub-module. This enzyme is involved in the biosynthesis of brassinosteroids [112]. However, its direct involvement with seed development is yet to be understood and serves as a candidate for future studies.

### Analysis of inter-modular hubs

Six inter-modular hubs (Table 8) were discovered during the analysis and they are described in detail in Additional file 1: Table S7. Among the discovered, ATP synthase subunit beta, mitochondrial and ATP synthase subunit gamma, mitochondrial, both predicted proteins for rice seed development, belong to the ATP synthase family, hinting at their involvement in ATP biogenesis. ATP synthase subunit gamma, mitochondrial acts as a connector hub, linking translation, endosperm development, hydrogen ion transmembrane transport, and transitional ion homeostasis, while ATP synthase subunit beta, mitochondrial connects pathways such as hydrogen ion transmembrane transport, leucine biosynthesis, and translation. Despite the energy-intensive nature of these pathways, no studies have yet confirmed their mediation through these inter-modular hubs. Therefore, these are ideal candidates for future investigations.

Furthermore, SDH1, central to the tricarboxylic acid (TCA) cycle and the electron transport chain [113], was identified as both an inter-modular and intra-modular hub during the analysis. It was found in the sub-module 4, which is annotated to leucine biosynthesis, and it interconnects various sub-modules, including hydrogen ion transmembrane transport, protein refolding, glycolytic process, and translation. Hydrogen ion (H+) transmembrane transport in rice plays a crucial role in regulating enzymes involved in the TCA cycle in the mitochondria [114]. The regulation of enzymes in the TCA cycle is complex, involving multiple levels of control, including transcriptional and translation regulation [115]. Protein biosynthesis and refolding in cells require ATP-dependent mechanisms, which can be energy-intensive [116]. Additionally, the redox status of cells can impact protein folding [104]. These findings emphasize the complex interplay between TCA cycle intermediates, protein translation, and H+ion transport in developing rice grain cells. Based on the findings, SDH1 plays a key role in regulating the crosstalk between various sub-modules in seed development. While no direct knockout experiments on SDH1 have been conducted in rice, studies on Arabidopsis have identified two SDH1 genes, SDH1-1 and SDH1-2, important in seed development. Knockout of SDH1-1 in Arabidopsis resulted in seed abortion, indicating its crucial role in early seed development [113]. However, further experimental validations are needed to confirm its use as a genetic resource for crop improvement in rice.

### Using an ensemble method and selection of the prediction thresholds

Ensemble methods combine the strengths of multiple models to improve the accuracy of predictions [49]. The rule of sum, a widely used technique to produce ensemble models, involves adding the predicted probabilities from each model to obtain a combined confidence score [33]. This study used the rule of sum to integrate four widely used network-based candidate protein prediction algorithms to predict novel seed development proteins. Using this ensemble model reduces the error and bias caused by individual algorithms and ensures the most accurate predictions. The benchmark results of the ensemble model compared to individual algorithms (Table 2 and Fig. 2) clearly prove the superiority of the ensemble model for the predictions based on the higher AUROC, AUPR, and $F_{max}$ scores obtained. To our knowledge, this is the first instance where an ensemble of network-based prediction algorithms was used to predict novel plant proteins.

The selection of the prediction score cutoff for novel candidate protein predictions is an important factor in PPI analysis [117]. It impacts the sensitivity and specificity of predicted interactions. Raising the cutoff enhances specificity but decreases sensitivity while lowering the cutoff enhances sensitivity but decreases specificity. Hence, the choice of cutoff must strike a balance between these two parameters. The most common approach for selecting the prediction score cutoff for protein function prediction studies is arbitrarily selecting a threshold [118]. Most studies have selected a number, such as the top 20 or 50 predictions, without providing a solid reason for the selection [118–120]. This approach lacks robustness because the effectiveness of applying a random threshold cannot be quantitatively assessed. Recently, some studies have employed a trial-and-error method, where they visually observe the network structure with varying thresholds and select the threshold with the best network visualization [10, 14]. Again, this method is not robust because of the inability to quantitatively assess the effectiveness of the selected threshold and the bias introduced through visual observation. Precision top-N curves, where precision is represented along varying thresholds, are used to select the optimum threshold corresponding to the highest precision in prediction studies [121, 122]. However, to our knowledge, this method has not been applied to protein function prediction. Therefore, this study determined the optimal cutoffs for the most reliable predictions using a modified precision top-N curve. Unlike the conventional approach of simply calculating the precision at the N number of top-ranked predictions [122], the modified curve considers the percentage of overlap between the DEPs and the top-ranked N predictions. Here, the DEPs are differentially expressed proteins associated with seed development collected from literature and act as an independent source of validation for the predicted proteins. Using the traditional top-N curve for detecting the cutoff may lead to a substantial inclusion of false positives in the results, which is considerably reduced when using the modified precision top-N curve. Furthermore, employing the modified precision top-N curve for this study offered a more robust and quantitative alternative to previously mentioned traditional methods, such as arbitrarily selecting thresholds or network visualization-based trial-and-error methods.

In PPI analysis, intra-modular hubs are important for identifying protein complexes as they often represent the core proteins within a module. The specific cutoff for picking hubs can have a significant impact on the results [123]. The methods used to identify intra-modular hubs, i.e., hubs that have a high number of interactions within a module, typically involve calculating the degree of each protein either within the whole network or within the module and selecting the proteins with the highest degree, often the top 10% of proteins [124]. However, this technique can introduce a bias towards detecting a majority of hubs from more highly connected modules. Hence the within-module degree Z-score (Z), which gauges the extent to which a node's degree centrality differs from the anticipated degree of nodes with the same module membership, was used for this study [60]. Previous works have used intra-modular Z-score cutoffs of 2.5 [125] and 1.5 [61] based on the properties of the network and research query. This study used 1.5 as the cutoff as 2.5 was unable to identify the protein with the highest within-module degree in the network.

For the prediction of inter-modular hubs, which connect different modules, the PC values were calculated. The choice of a PC cutoff for their detection in a PPI network may depend on the specific research question and the characteristics of the network being analyzed [126]. Inter-modular hubs were selected using a PC cutoff of 0.5, which has been employed in previous studies [65]. All of the resulting inter-moduler hubs were connected to at least three sub-modules.

For the STRING rice PPI network used for the analysis, the 0.7 combined score cutoff was applied because it was recommended by the STRING database to retain high-quality interactions [34, 35], which is also the most widely used cutoff for STRING PPI network analysis [127, 128]. However, the application of a cutoff inevitably removes a certain number of seed proteins from the network because those do not contain any interaction with a combined score above the selected cutoff. In this work, seven such seed proteins were removed from the original 102 (Additional file 1: Table S1) because they did not contain any interaction higher than the 0.7 combined score cutoff in the given sub-network. Lowering the cutoff may increase the number of seed proteins that can be retained, but the confidence of their interactions presents a serious issue for the downstream network analyses as the majority of those interactions are of low quality. Therefore, in this work, a few proteins had to be removed to maintain the quality of the predictions and analysis results, which is inevitable in STRING PPI network-based studies [31, 129].

### Limitations

This work revealed several hub proteins associated with seed development in rice. Generally, these hubs are considered more important for the regulation of the considered function and are often used as drug targets [27]. Some of these hub proteins, such as APx1 [38], APG [93] and DHAR1 [104], are seed proteins that are experimentally validated to be associated with seed development. This study also predicted new protein candidates, such as ABI5 and RPL3B, which were also discovered as hub proteins associated with seed development. These predicted proteins were validated as differentially expressed proteins associated with seed development from independent transcriptomic experiments [41, 42], but they lack experimental validation. Conducting experimental validations on the predicted proteins is beyond the scope of this work and is a major limitation of this study. It is important to conduct experimental validations of the newly predicted proteins before using them for biotechnological applications. However, this study provides a list of protein candidates for further experiments, such as quantitative Polymerase Chain Reaction (qPCR) [130] and gene knockout or knockdown studies [131], which could be used to experimentally validate the association with seed development. Widely used molecular techniques, such as CRISPR/cas9 [132, 133] and RNA interference [134], can be used to perform knockouts of the predicted hub proteins and observe their effect on seed development.

### Future work

In this work, an ensemble model composed of four network-based protein function prediction algorithms was used for predicting new candidate proteins associated with seed development. The performance evaluations showed a superior accuracy for the ensemble model over individual algorithms, indicating the suitability of algorithm integration for protein function prediction tasks. However, the accuracy of the current model has the potential to be further enhanced. Deep learning has emerged as a popular technique to improve the accuracy of traditional models [18, 22], and could be used for potential improvements of this model. For instance, a multimodal deep learning framework can be tested in the future for algorithm integration, which enables the integration of other types of data, such as protein structures [18, 49]. Such a framework could efficiently integrate different types of data compared to conventional methods, such as the rule of sum. Furthermore, the ensemble model used for this work explicitly used PPI network data retrieved from the STRING database. Other types of network data sourced from regulatory, metabolomic, and transcriptomic networks can be integrated with the PPI network data and tested in the future as possible enhancements. The successful use of the ensemble model in predicting seed development proteins in this work indicates the future potential of this model for predicting genes associated with other phenotypes, such as human diseases, plant diseases, and microbial growth, opening numerous opportunities for future applications.

## Conclusions

In this study, an ensemble of network-based algorithms was employed to predict new candidate proteins associated with rice seed development and study their systems biology landscape. This ensemble model outperformed individual algorithms when predicting seed development proteins. The use of the ensemble approach was novel in predicting candidate proteins and the relevance of the predicted candidates was demonstrated by the enrichment of the key GO-BP terms common with original seed proteins. Also, all the predicted candidates were found to be differentially expressed during seed development, providing extra validation. Furthermore, the submodule analysis revealed specific pathways linked to seed growth regulation, endosperm development, and general processes such as translation, protein refolding, cell differentiation, and plant hormone signaling.

This study revealed hub proteins that are central to the stability of the extracted PPI sub-network. Among the 17 intra-modular hubs, DHAR1, Inorganic disphosphatase, CYCB2-2, HMA4, and Copper chaperone were the top five according to their Z-scores. Additionally, six inter-modular hubs that regulate cross-talk between different sub-modules were identified, including two predicted proteins and four seed proteins. Of these, the SDH1 protein achieved the highest score and was also revealed to have a dual role as an intra-modular hub, underscoring its importance as a central regulator. Newly predicted proteins such as RPL3B, Sterol delta-7-reductase, ATP synthase subunit gamma, mitochondrial, HMA4, and ABI5 were also identified as important hub proteins which are ideal candidates for further experimental studies.

Collectively, these novel protein candidates offer promising avenues for gaining deeper insights into the regulatory mechanisms governing rice grain development. This work provides the first comprehensive view of the protein interaction landscape associated with rice seed development. It is crucial to experimentally validate the identified hub proteins and predicted proteins in future studies as they could be important targets for genetic research aiming to improve rice grain quality and yield in rice. This study demonstrates how systems biology analysis techniques can be used for studying crucial developmental biology processes and serves as a blueprint for future studies.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-025-06595-7.

Additional file 1: The supplementary tables (Tables S1–S7) in this file provide comprehensive datasets that support the main findings of this study. Table S1 lists the seed proteins, their associated biological processes, and their retention status in the final network after applying the 0.7 combined score cutoff. Table S2 presents DEPs related to rice seed development, collected from the literature. Table S3 shows enriched GO-BP terms for the seed proteins using DAVID. Table S4 provides GO enrichment results for each network sub-module using DAVID. Table S5 summarizes the most relevant GO-BP term and the protein composition of each sub-module. Table S6 highlights intra-modular hub proteins and those with literature support for their roles in seed development. Table S7 lists inter-module hub proteins, describing their functions and roles in connecting different sub-modules.

### Authors' contributions

All authors were involved in planning and designing the experiments. PCF conceived the study and MRP wrote the Python scripts for the analysis, except for the RWR script, which was written by JW. MRP and SP performed the experiments under the supervision of PCF. PCF, MRP, SP, and JW contributed to the preparation of the final manuscript. All authors analyzed the results and read and approved the final version of the manuscript.

### Funding

Not Applicable.

### Data availability

All data generated or analyzed in support of the findings of this study are included in the manuscript and its supplementary information file. The Python codes used for this study are available at https://github.com/rashpr88/RiceSeedDevelopment.

## Declarations

### Ethics approval and consent to participate

Not Applicable.

### Consent for publication

All authors have read and approved the final manuscript for publication.

### Competing interests

The authors declare no competing interests.

Received: 19 November 2024 Accepted: 21 April 2025
Published online: 08 May 2025

### References

1. Huang R, Jiang L, Zheng J, Wang T, Wang H, Huang Y. Genetic bases of rice grain shape: So many genes, so little known. Trends Plant Sci. 2013;18:218–26.
2. Waterworth WM, Bray CM, West CE. Seeds and the art of genome maintenance. Front Plant Sci. Frontiers Media S.A.; 2019.
3. King T, Cole M, Farber JM, Eisenbrand G, Zabaras D, Fox EM. Food safety for food security: Relationship between global megatrends and developments in food safety. Trends Food Sci Technol. 2017;68:160–75.
4. Sadigov R. Rapid Growth of the World Population and Its Socioeconomic Results. Sci World J. 2022;2022(1):8110229.
5. Deng ZY, Gong CY, Wang T. Use of proteomics to understand seed development in rice. Proteomics. 2013;13(12–13):1784–800.
6. An L, Tao Y, Chen H, He M, Xiao F, Li G. Embryo-Endosperm Interaction and Its Agronomic Relevance to Rice Quality. Front Plant Sci. 2020;11:587641.
7. Kozaki A, Aoyanagi T. Molecular Aspects of Seed Development Controlled by Gibberellins and Abscisic Acids. Int J Mol Sci. 2022;23(3):1876.
8. Fitzgerald MA, McCouch SR, Hall RD. Not just a grain of rice: the quest for quality. Trends Plant Sci. 2009;14(3):133–9.
9. Tappiban P, Ying Y, Xu F, Bao J. Proteomics and post-translational modifications of starch biosynthesis-related proteins in developing seeds of rice. Int J Mol Sci. 2021;22(11):5901.
10. Wimalagunasekara SS, Weeraman JWJK, Tirimanne S, Fernando PC. Protein-protein interaction (PPI) network analysis reveals important hub proteins and sub-network modules for root development in rice (Oryza sativa). J Genet Eng Biotechnol. 2023;21(1):69.
11. Vella D, Marini S, Vitali F, Di Silvestre D, Mauri G, Bellazzi R. MTGO: PPI Network Analysis Via Topological and Functional Module Identification. Sci Rep. 2018;8(1):5499.
12. Fernando PC, Mabee PM, Zeng E. Integration of anatomy ontology data with protein-protein interaction networks improves the candidate gene prediction accuracy for anatomical entities. BMC Bioinf. 2020;21:1–26.
13. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3(1):88.
14. Fernando PC, Mabee PM, Zeng E. Protein–protein interaction network module changes associated with the vertebrate fin-to-limb transition. Sci Rep. 2023;13(1):22594.
15. Roy A, Kucukural A, Zhang Y. I-TASSER: A unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5(4):725–38.
16. Zheng LL, Li YX, Ding J, Guo XK, Feng KY, Wang YJ. A comparison of computational methods for identifying virulence factors. PLoS One. 2012;7(8):e42517.
17. Lin B, Luo X, Liu Y, Jin X. A comprehensive review and comparison of existing computational methods for protein function prediction. Brief. Bioinform: Oxford University Press; 2024.
18. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T. Structure-based protein function prediction using graph convolutional networks. Nat Commun. 2021;12(1):3168.
19. Song F V., Su J, Huang S, Zhang N, Li K, Ni M. DeepSS2GO: protein function prediction from secondary structure. Brief Bioinform. 2024;25(3):bbae196.
20. Avery C, Patterson J, Grear T, Frater T, Jacobs DJ. Protein Function Analysis through Machine Learning. Biomol. 2022;12(9):1246.
21. Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-dabbagh BSN. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J Big Data. 2023;10(1):46.
22. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP. A survey on deep learning: Algorithms, techniques, and applications. ACM Comput Surv. 2018;51(5):1–36.
23. Fortunato S. Community detection in graphs. Phys Rep. 2010;486(3–5):75–174.
24. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theor Exp. 2008;2008(10):P10008.
25. Kiran M, Nagarajaram HA. Interaction and localization diversities of global and local hubs in human protein-protein interaction networks. Mol Biosyst. 2016;12:2875–82.
26. He X, Zhang J. Why do hubs tend to be essential in protein networks? PLoS Genet. 2006;2:0826–34.

27. Hasan MI, Rahman MH, Islam MB, Islam MZ, Hossain MA, Moni MA. Systems Biology and Bioinformatics approach to Identify blood based signatures molecules and drug targets of patient with COVID-19. Inform Med Unlocked. 2021;28:100840.
28. Wu B, Xi S. Bioinformatics analysis of differentially expressed genes and pathways in the development of cervical cancer. BMC Cancer. 2021;21(1):733.
29. Zhuang DY, Jiang LI, He QQ, Zhou P, Yue T. Identification of hub subnetwork based on topological features of genes in breast cancer. Int J Mol Med. 2015;35(3):664–74.
30. Xu Z, Zhou Y, Cao Y, Dinh TLA, Wan J, Zhao M. Identification of candidate biomarkers and analysis of prognostic values in ovarian cancer by integrated bioinformatics analysis. Med Oncol. 2016;33:1–8.
31. Hozhabri H, Dehkohneh RSG, Razavi SM, Razavi SM, Salarian F, Rasouli A. Comparative analysis of protein-protein interaction networks in metastatic breast cancer. PLoS One. 2022;17(1):e0260584.
32. Backiyarani S, Sasikala R, Sharmiladevi S, Uma S. Decoding the molecular mechanism of parthenocarpy in Musa spp. through protein–protein interaction network. Sci Rep. 2021;11(1):14592.
33. Dua S, Chowriappa P. Data Mining for Bioinformatics. 1st ed. CRC Press; 2012.
34. Folador EL, Hassan SS, Lemke N, Barh D, Silva A, Ferreira RS. An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage. Integr Biol. 2014;6(11):1080–7.
35. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005;33(suppl_1):D433–7.
36. Chen L, Gao W, Chen S, Wang L, Zou J, Liu Y. High-resolution QTL mapping for grain appearance traits and co-localization of chalkiness-associated differentially expressed candidate genes in rice. Rice. 2016;9:1–17.
37. Mahto A, Mathew I, Agarwal P. Decoding the transcriptome of rice seed during development. In: Jimenez-Lopez JC, editor. Advances in Seed Biology. Spain: InTech; 2017. p. 25.
38. Kim YJ, Kim S-I, Kesavan M, Kwak JS, Song JT, Seo HS. Ascorbate Peroxidase OsAPx1 is Involved in Seed Development in Rice. Plant Breed Biotech. 2015;3:11–20.
39. Yang Y, Dai L, Xia H, Zhu K, Liu H, Chen K. Protein profile of rice (Oryza sativa) seeds. Genet Mol Biol. 2013;36:87–92.
40. You C, Chen L, He H, Wu L, Wang S, Ding Y. iTRAQ-based proteome profile analysis of superior and inferior Spikelets at early grain filling stage in japonica Rice. BMC Plant Biol. 2017;17:1–20.
41. Lee J, Koh HJ. A label-free quantitative shotgun proteomics analysis of rice grain development. Proteome Sci. 2011;9:1–10.
42. Xue LJ, Zhang JJ, Xue HW. Genome-wide analysis of the complex transcriptional networks of rice developing seeds. PLoS One. 2012;7(2):e31081.
43. Kotu V, Deshpande B. Predictive analytics and data mining : concepts and practice with RapidMiner. Burlington: Morgan Kaufmann; 2014.
44. Lee I, Nam H. Identification of drug-target interaction by a random walk with restart method on an interactome network. BMC Bioinf. 2018;19:9–18.
45. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol. 2000;18(12):1257–61.
46. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast. 2001;18(6):523–31.
47. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005;21(suppl_1):i302–10.
48. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: A universal amplifier of genetic associations. Nat Rev Genet. 2017;18:551–62.
49. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Comput Methods Programs Biomed. 2018;153:1–9.
50. Li JR, Chen L, Wang SP, Zhang YH, Kong XY, Huang T. A computational method using the random walk with restart algorithm for identifying novel epigenetic factors. Mol Genet Genomics. 2018;293(1):293–301.
51. Zhu L, Su F, Xu YC, Zou Q. Network-based method for mining novel HPV infection related genes using random walk with restart algorithm. Biochim Biophys Acta Mol Basis Dis. 2018;1864(6):2376–83.
52. Carlin DE, Demchak B, Pratt D, Sage E, Ideker T. Network propagation in the cytoscape cyberinfrastructure. PLoS Comput Biol. 2017;13(10):e1005598.
53. Jiang Z, Liu H, Fu B, Wu Z, Zhang T. Recommendation in heterogeneous information networks based on generalized random walk model and Bayesian Personalized Ranking. In: Yi C, Chengxiang Z, editors. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. New York: Association for Computing Machinery; 2018. p. 288–96.
54. Zhang Z, Zhang J. A big world inside small-world networks. PLoS One. 2009;4(5):e5686.
55. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44:837–45.
56. Li X, Lv J, Yi Z. Outlier Detection Using Structural Scores in a High-Dimensional Space. IEEE Trans Cybern. 2020;50(5):2302–10.
57. Fan G, Wei J. Identification of potential novel biomarkers and therapeutic targets involved in human atrial fibrillation based on bioinformatics analysis. Kardiol Pol. 2020;78:694–702.
58. Alcalá-Corona SA, Sandoval-Motta S, Espinal-Enríquez J, Hernández-Lemus E. Modularity in Biological Networks. Front Genet. 2021;12:701331.
59. Tang D, Zhao X, Zhang L, Wang Z, Wang C. Identification of hub genes to regulate breast cancer metastasis to brain by bioinformatics analyses. J Cell Biochem. 2019;120(6):9522–31.
60. McGarry K, Daniel U. Computational techniques for identifying networks of interrelated diseases. In: Neagu D, editor. 2014 14th UK Workshop on Computational Intelligence (UKCI). Bradford: Institute of Electrical and Electronics Engineers (IEEE); 2014. p. 1–8.
61. Zilidou VI, Frantzidis CA, Romanopoulou ED, Paraskevopoulos E, Douka S, Bamidis PD. Functional Re-organization of Cortical Networks of Senior Citizens After a 24-Week Traditional Dance Program. Front Aging Neurosci. 2018;10:422.
62. Ghalmane Z, El Hassouni M, Cherifi C, Cherifi H. Centrality in modular networks. EPJ Data Sci. 2019;8(1):15.
63. Vértes PE, Rittman T, Whitaker KJ, Romero-Garcia R, Váša F, Kitzbichler MG. Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. Philos Trans R Soc B. 2016;371(1705):20150362.
64. Power JD, Schlaggar BL, Lessov-Schlaggar CN, Petersen SE. Evidence for hubs in human functional brain networks. Neuron. 2013;79(4):798–813.
65. Liu Y, Hong X, Bengson JJ, Kelley TA, Ding M, Mangun GR. Deciding Where to Attend: Large-scale Network Mechanisms Underlying Attention and Intention Revealed by Graph-theoretic Analysis. Neuroimage. 2017;157:45–60.
66. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):D439–44.
67. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
68. Sabelli PA, Larkins BA, editors. Advances in Seed Biology. Advances in Seed Biology. Lausanne: Frontiers Media SA; 2015.
69. Wang W, Li Y, Dang P, Zhao S, Lai D, Zhou L. Rice secondary metabolites: Structures, roles, biosynthesis, and metabolic regulation. Molecules. 2018;23(12):3098.
70. Adom KK, Liu RH. Antioxidant activity of grains. J Agric Food Chem. 2002;50(21):6182–7.
71. Shu K, Zhou W, Chen F, Luo X, Yang W. Abscisic acid and gibberellins antagonistically mediate plant development and abiotic stress responses. Front Plant Sci. 2018;9:416.
72. Kusano H, Arisu Y, Nakajima J, Yaeshima M, She KC, Shimada H. Implications of the gene for F1-ATPase β subunit (AtpB) for the grain quality of rice matured in a high-temperature environment. Plant Biotechnol. 2016;33:169–75.

73. Borisjuk L, Rolletschek H, Radchuk R, Weschke W, Wobus U, Weber H. Seed development and differentiation: A role for metabolic regulation. Plant Biol. 2004;6(04):375–86.
74. Muench SP, Trinick J, Harrison MA. Structural divergence of the rotary ATPases. Q Rev Biophys. 2011;44:311–56.
75. Zhu M, Zang Y, Zhang X, Shang S, Xue S, Chen J. Insights into the regulation of energy metabolism during the seed-to-seedling transition in marine angiosperm Zostera marina L.: Integrated metabolomic and transcriptomic analysis. Front Plant Sci. 2023;14:1130292.
76. Ali F, Qanmber G, Li F, Wang Z. Updated role of ABA in seed maturation, dormancy, and germination. J Adv Res. 2022;35:199–214.
77. Gampala SSL, Finkelstein RR, Sun SSM, Rock CD. ABI5 interacts with abscisic acid signaling effectors in rice protoplasts. J Biol Chem. 2002;277(3):1689–94.
78. Zheng M, Wang Y, Liu X, Sun J, Wang Y, Xu Y. The RICE MINUTE-LIKE1 (RML1) gene, encoding a ribosomal large subunit protein L3B, regulates leaf morphology and plant architecture in rice. J Exp Bot. 2016;67:3457–69.
79. Huang XY, Deng F, Yamaji N, Pinson SRM, Fujii-Kashino M, Danku J. A heavy metal P-type ATPase OsHMA4 prevents copper accumulation in rice grain. Nat Commun. 2016;7(1):12138.
80. Walker EL, Waters BM. The role of transition metal homeostasis in plant seed development. Curr Opin Plant Biol. 2011;14(3):318–24.
81. Troncoso-Ponce MA, Kruger NJ, Ratcliffe G, Garcés R, Martínez-Force E. Characterization of glycolytic initial metabolites and enzyme activities in developing sunflower (Helianthus annuus L.) seeds. Phytochemistry. 2009;70(9):1117–22.
82. Dirk LMA, Downie AB. An examination of Job's rule: Protection and repair of the proteins of the translational apparatus in seeds. Seed Sci Res. 2018;28(3):168–81.
83. Luan X, Ke S, Liu S, Tang G, Huang D, Wei M. OsPEX1, a leucine-rich repeat extensin protein, functions in the regulation of caryopsis development and quality in rice. Crop J. 2022;10:704–15.
84. Wu X, Liu J, Li D, Liu CM. Rice caryopsis development II: Dynamic changes in the endosperm. J Integr Plant Biol. 2016;58:786–98.
85. Patrick JW, Offler CE. Compartmentation of transport and transfer events in developing seeds. J Exp Bot. 2001;52(356):551–64.
86. Hacham Y, Shitrit O, Nisimi O, Friebach M, Amir R. Elucidating the importance of the catabolic enzyme, methionine-gamma-lyase, in stresses during Arabidopsis seed development and germination. Front Plant Sci. 2023;14:1143021.
87. Ben-Nissan G, Sharon M. Regulating the 20S proteasome ubiquitin-independent degradation pathway. Biomolecules. 2014;4(3):862–84.
88. Wang WQ, Liu SJ, Song SQ, Møller IM. Proteomics of seed development, desiccation tolerance, germination and vigor. Plant Physiol Biochem. 2015;86:1–5.
89. Yin CC, Zhao H, Ma B, Chen SY, Zhang JS. Diverse roles of ethylene in regulating agronomic traits in rice. Front Plant Sci. 2017;8:1676.
90. Dante RA, Larkins BA, Sabelli PA. Cell cycle control and seed development. Front Plant Sci. 2014;5:493.
91. Zhang C, Bai M yi, Chong K. Brassinosteroid-mediated regulation of agronomic traits in rice. Plant Cell Rep. 2014;33:683–96.
92. Usman B, Nawaz G, Zhao N, Liu Y, Li R. Generation of high yielding and fragrant rice (Oryza sativa l.) lines by CRISPR/Cas9 targeted mutagenesis of three homoeologs of cytochrome p450 gene family and osbadh2 and transcriptome and proteome profiling of revealed changes triggered by mutations. Plants. 2020;9:1–28.
93. Heang D, Sassa H. An atypical bHLH protein encoded by POSITIVE REGULATOR OF GRAIN LENGTH 2 is involved in controlling grain length and weight of rice through interaction with a typical bHLH protein APG. Breed Sci. 2012;62:133–41.
94. Nallamilli BRR, Zhang J, Mujahid H, Malone BM, Bridges SM, Peng Z. Polycomb Group Gene OsFIE2 Regulates Rice (Oryza sativa) Seed Development and Grain Filling via a Mechanism Distinct from Arabidopsis. PLoS Genet. 2013;9(3):e1003322.
95. Wang L, Yuan J, Ma Y, Jiao W, Ye W, Yang DL. Rice Interploidy Crosses Disrupt Epigenetic Regulation, Gene Expression, and Seed Development. Mol Plant. 2018;11:300–14.
96. Yamaji N, Xia J, Mitani-Ueno N, Yokosho K, Ma JF. Preferential delivery of zinc to developing tissues in rice is mediated by P-type heavy metal ATPase OsHMA2. Plant Physiol. 2013;162:927–39.
97. Zhang L, Ren Y, Lu B, Yang C, Feng Z, Liu Z. FLOURY ENDOSPERM7 encodes a regulator of starch synthesis and amyloplast development essential for peripheral endosperm development in rice. J Exp Bot. 2016;67:633–47.
98. Gayen D, Ghosh S, Paul S, Sarkar SN, Datta SK, Datta K. Metabolic regulation of carotenoid-enriched golden rice line. Front Plant Sci. 2016;7:1622.
99. He W, Wang L, Lin Q, Yu F. Rice seed storage proteins: Biosynthetic pathways and the effects of environmental factors. J Integr Plant Biol. 2021;63(12):1999–2019.
100. Timabud T, Yin X, Pongdontri P, Komatsu S. Gel-free/label-free proteomic analysis of developing rice grains under heat stress. J Proteomics. 2016;133:1–19.
101. Shi Y, Zhang Y, Sun Y, Xie Z, Luo Y, Long Q. Natural variations of OsAUX5, a target gene of OsWRKY78, control the neutral essential amino acid content in rice grains. Mol Plant. 2023;16:322–36.
102. Sikdar MSI, Kim JS. Isolation of a gene encoding 3-isopropylmalate dehydrogenase from rice. Russ J Plant Physiol. 2011;58:190–6.
103. Li Y, Fan C, Xing Y, Yun P, Luo L, Yan B. Chalk5 encodes a vacuolar H + -translocating pyrophosphatase influencing grain chalkiness in rice. Nat Genet. 2014;46:398–404.
104. Kim YS, Kim IS, Bae MJ, Choe YH, Kim YH, Park HM. Homologous expression of cytosolic dehydroascorbate reductase increases grain yield and biomass under paddy field conditions in transgenic rice (Oryza sativa L. japonica). Planta. 2013;237:1613–25.
105. Erales J, Coffino P. Ubiquitin-independent proteasomal degradation. Biochim Biophys Acta, Mol Cell Res. 2014;1843:216–21.
106. Kyosseva SV. Mitogen-Activated Protein Kinase Signaling. Int Rev Neurobiol. 2004;59:201–20.
107. Minkenberg B, Xie K, Yang Y. Discovery of rice essential genes by characterizing a CRISPR-edited mutation of closely related rice MAP kinase genes. Plant J. 2017;89:636–48.
108. Zhao H, Yin CC, Ma B, Chen SY, Zhang JS. Ethylene signaling in rice and Arabidopsis: New regulators and mechanisms. J Integr Plant Biol. 2021;63(1):102–25.
109. Yang BJ, Wendrich JR, De Rybel B, Weijers D, Xue HW. Rice microtubule-associated protein IQ67-DOMAIN14 regulates grain shape by modulating microtubule cytoskeleton dynamics. Plant Biotechnol J. 2020;18:1141–52.
110. Divi UK, Krishna P. Brassinosteroid: a biotechnological target for enhancing crop yield and stress tolerance. New Biotechnol. 2009;26(3–4):131–6.
111. Hong Z, Ueguchi-Tanaka M, Fujioka S, Takatsuto S, Yoshida S, Hasegawa Y. The rice brassinosteroid-deficient dwarf2 mutant, defective in the rice homolog of arabidopsis DIMINUTO/DWARF1, is rescued by the endogenously accumulated alternative bioactive brassinosteroid, dolichosterone. Plant Cell. 2005;17:2243–54.
112. Ito Y, Thirumurugan T, Serizawa A, Hiratsu K, Ohme-Takagi M, Kurata N. Aberrant vegetative and reproductive development by overexpression and lethality by silencing of OsHAP3E in rice. Plant Sci. 2011;181:105–10.
113. Huang S, Millar AH. Succinate dehydrogenase: the complex roles of a simple enzyme. Curr Opin Plant Biol. 2013;16:344–9.
114. Liao JL, Zhou HW, Peng Q, Zhong PA, Zhang HY, He C. Transcriptome changes in rice (Oryza sativa L.) in response to high night temperature stress at the early milky stage. BMC Genomics. 2015;16:1–4.
115. Zhang Y, Fernie AR. On the role of the tricarboxylic acid cycle in plant productivity. J Integr Plant Biol. 2018;60(12):1199–216.
116. Cao H, Duncan O, Millar AH. The molecular basis of cereal grain proteostasis. Essays Biochem. 2022;66(2):243–53.
117. Gorji-bahri G, Moghimi HR, Hashemi A. RAB5A is associated with genes involved in exosome secretion: Integration of bioinformatics analysis and experimental validation. J Cell Biochem. 2021;122(3–4):425–41.
118. Zhu L, Zhang H, Cao D, Xu Y, Li L, Ning Z. Drought Stress-Related Gene Identification in Rice by Random Walk with Restart on Multiplex Biological Networks. Agriculture. 2022;13(1):53.
119. Yang P, Li X, Wu M, Kwoh CK, Ng SK. Inferring Gene-Phenotype associations via global protein complex network propagation. PLoS One. 2011;6(7):e21502.
120. Zamanian-Azodi M, Rezaei-Tavirani M, Rahmati-Rad S, Hasanzadeh H, Tavirani MR, Seyyedi SS. Protein-protein interaction network could

reveal the relationship between the breast and colon cancer. GHFBB. 2015;8(3):215.

121. Rahman MM, Vadrev SM, Magana-Mora A, Levman J, Soufan O. A novel graph mining approach to predict and evaluate food-drug interactions. Sci Rep. 2022;12(1):1061.
122. Liong VE, Lu J, Wang G, Moulin P, Zhou J. Deep Hashing for Compact Binary Codes Learning. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston; 2015. p. 2475–83.
123. Xiong Y, You W, Wang R, Peng L, Fu Z. Prediction and validation of hub genes associated with colorectal cancer by integrating PPI network and gene expression data. Biomed Res Int. 2017;2017(1):2421459.
124. Wang W, Shen J, Qi C, Pu J, Chen H, Zuo Z. The key candidate genes in tubulointerstitial injury of chronic kidney diseases patients as determined by bioinformatic analysis. Cell Biochem Funct. 2020;38:761–72.
125. Guimerà R, Mossa S, Turtschi A, Amaral LAN, Wachter KW. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. PNAS. 2005;102:7794–9.
126. Joyce KE, Laurienti PJ, Burdette JH, Hayasaka S. A new measure of centrality for brain networks. PLoS One. 2010;5(8):e12200.
127. Deng JL, Xu YH, Wang G. Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. Front Genet. 2019;10:695.
128. Flórez AF, Park D, Bhak J, Kim BC, Kuchinsky A, Morris JH. Protein network prediction and topological analysis in Leishmania major as a tool for drug target selection. BMC Bioinf. 2010;11:1–9.
129. Huang S, Zhong J, Qi Q, Liu G, Gong M. CircRNA expression profile and potential role of hsa_circ_0040039 in intervertebral disc degeneration. Medicine. 2022;101(32):e30035.
130. Kelder TP, Penning ME, Uh HW, Cohen D, Bloemenkamp KWM, Bruijn JA. Quantitative polymerase chain reaction-based analysis of podocyturia is a feasible diagnostic tool in preeclampsia. Hypertension. 2012;60(6):1538–44.
131. Jiao SY, Yang YH, Chen SR. Molecular genetics of infertility: Loss-of-function mutations in humans and corresponding knockout/mutated mice. Hum Reprod Update. 2021;27(1):154–89.
132. Ghavami S, Pandi A. CRISPR interference and its applications. Prog Mol Biol Transl Sci. 2021;180:123–40.
133. Geng S, Sohail H, Cao H, Sun J, Chen Z, Zhou L. An efficient root transformation system for CRISPR/Cas9-based analyses of shoot-root communication in cucurbit crops. Hortic Res. 2022;9:uhab082.
134. Watts GF, Schwabe C, Scott R, Gladding PA, Sullivan D, Baker J. RNA interference targeting ANGPTL3 for triglyceride and cholesterol lowering: phase 1 basket trial cohorts. Nat Med. 2023;29(9):2216–23.

## Publisher's Note