# FKMU: K-Means Under-Sampling for Data Imbalance in Predicting TF-Target Genes Interactions

Thanh Tuoi Le[1, 2], Xuan Tho Dang[3]*
Faculty of Information Technology, Hanoi National University of Education, Hanoi City, Vietnam
Faculty of Information Technology, Vinh University of Technology Education, Vinh City, Vietnam
Academy of Policy and Development, Hanoi City, Vietnam

***Abstract*—Identifying interactions between transcription factors (TFs) and target genes is critical for understanding molecular mechanisms in biology and disease. Traditional experimental approaches are often costly and not scalable. We introduce FKMU, a K-means-based under-sampling method designed to address data imbalance in predicting TF-target interactions. By selecting low-frequency TF samples within each cluster and optimizing the balance ratio to 1:1 between known and unknown samples, FKMU significantly improves prediction accuracy for unobserved interactions. Integrated with a deep learning model that uses random walk sampling and skip-gram embeddings, FKMU achieves an average AUC of 0.9388 ± 0.0045 through five-fold cross-validation, outperforming state-of-the-art methods. This approach facilitates accurate and large-scale predictions of TF-target interactions, providing a robust tool for molecular biology research.**

***Keywords—K-means clustering; imbalanced data; TF-target gene interactions; heterogeneous network; meta-path***

## I. Introduction

Transcription Factors (TFs) are essential regulatory proteins in the process of gene transcription, which is the mechanism of transferring genetic information from DNA to RNA [8]. TFs perform their role by binding to specific DNA sequences, often located in or near gene promoters. Upon binding to DNA, TFs can either activate or inhibit the function of RNA polymerase, the enzyme responsible for transcribing DNA into RNA. Through this mechanism, TFs regulate gene expression, playing a crucial role in the development and maintenance of cellular functions. TFs are found in almost all living organisms and are vital for gene expression regulation. However, when TFs lose their function, the balance in gene regulation is disrupted, leading to severe diseases. Accurately identifying the relationships between TFs and target genes is a crucial step in understanding the complex molecular mechanisms involved in biological and pathological processes. These insights will pave the way for extensive research in molecular biology and applied medicine, laying the foundation for more effective diagnostic and therapeutic methods in the future.

Previously, identifying interactions between TFs and target genes relied primarily on experimental methods, which were costly and time-intensive. The emergence of large-scale techniques such as ChIP-seq and RNA-seq has made it more feasible to predict TF target genes across the entire genome [12, 13]. ChIP-seq maps TF-DNA interactions, while RNA-seq provides RNA expression data, shedding light on genes influenced by TFs [20]. However, these methods reveal only a small fraction of the complex gene regulatory network.

Many interactions between TFs and target genes remain unclear in existing databases. Datasets on TF-target gene interactions collected from ChIP-seq techniques provide a limited view of the complex gene regulatory network. Specifically, most current computational methods only identify binding sites without addressing the nature of these interactions. Although some recent studies have made progress in predicting these interactions, building high-quality datasets with both positive and negative samples remains a significant challenge. Furthermore, current methods often fail to effectively address the data imbalance issue, particularly in selecting negative samples. This limitation can result in the failure to detect potential interactions between TFs and target genes, reducing the accuracy of prediction models. Additionally, failure to address the data imbalance issue can introduce bias during the training process [2-4], impairing the ability to detect important interactions in the gene regulatory network. Therefore, the development of new methods focused on data balancing is crucial to improve prediction performance and provide a solid foundation for molecular biology and applied medicine research.

This paper introduces a novel approach to address data imbalance for improving the prediction of TF-target gene interactions. Key contributions of this study include:

*a)* We present the FKMU method, an under-sampling technique based on the K-means clustering algorithm and the inverse information principle, designed to enhance efficiency and stability in predicting TF-target gene interactions.

*b)* A novel meta-path schema has been developed to extend the capability of capturing potential links within heterogeneous networks, significantly improving the predictive performance of the model.

*c)* The FKMU model incorporates substantial advancements, achieving superior performance in identifying unknown TF-target gene interactions compared to existing approaches.

*d)* The effectiveness of the proposed method is validated through rigorous experiments, demonstrating outstanding results in terms of accuracy and predictive efficiency over current methods.

*e)* Experimental results confirm that FKMU is an effective and accurate solution, achieving an average AUC value

[*] Corresponding Author

superior to many existing methods, demonstrating its potential for widespread application in molecular biology research.

The remainder of this paper is organized as follows. Section II reviews related works on predicting interactions between TFs and target genes based on TF binding sites, gene expression data, and heterogeneous networks. Section III introduces the FKMU method, which combines K-means clustering and negative sampling. Section IV presents the experimental results, including evaluation metrics, parameter optimization, and performance comparisons. Finally, Section V concludes with the contributions of the study and suggests potential directions for future development.

## II. Related Works

Predicting interactions between TFs and target genes is a critical topic in the field of computational biology. Traditional experimental methods are often time-consuming, costly, and challenging to apply at scale, while also carrying the risk of failure. Artificial intelligence offers a powerful tool to support these experimental approaches, helping to narrow down the search for potential interactions between TFs and target genes and optimizing them for subsequent experimental validation. As a result, research time and costs can be significantly reduced, facilitating the research and development process. Research related to our approach can be divided into the following three subsections.

*1) Methods based on predicting transcription factor binding sites (TFBS)*: These methods primarily focus on identifying interactions between TFs and target genes by detecting their binding sites. The process involves determining the binding positions of TFBS, which are often integrated with deep learning models such as Convolutional Neural Networks (CNNs), as demonstrated in the research by S. Salekin et al. [21] and Ž. Avsec et al. [29], or Recurrent Neural Networks (RNNs), as referenced in the studies by J. Lanchantin et al. [10] and Z. Shen et al. [32]. However, these methods have a significant limitation, leading to a high false positive rate because TFBS are often located within long non-coding sequences. Furthermore, they do not directly predict TF-target gene interactions but rather infer them based on the locations of TFBS.

*2) Direct prediction methods for TF-target gene interactions based on gene expression data:* These methods do not rely on TFBS but instead use gene expression data, such as gene expression images from in situ hybridization (ISH) or single-cell RNA sequencing (scRNA-seq) data, to directly predict the relationship between TFs and target genes. For example, using gene expression image analysis, Y. Yang et al. [28] developed GripDL, an effective tool for studying transcriptional regulatory networks in Drosophila. GripDL utilizes ISH images as input, combined with a deep residual model to leverage known TF-target gene interactions. Results showed that GripDL outperformed traditional methods in accuracy and the ability to detect novel gene interactions, offering valuable insights into eye development in Drosophila and paving the way for new research on gene regulatory networks. Beyond gene expression images, single-cell RNA sequencing (scRNA-seq) data provides an additional perspective for understanding complex mechanisms by which TFs regulate target genes. Su et al. [15] developed NetAct, a computational platform for constructing transcription factor regulatory networks using transcriptomic data and gene databases. This tool has been effectively applied to model regulatory networks in epithelial-mesenchymal transition and macrophage polarization, highlighting its significant potential in analyzing complex gene networks. Y. Fan et al. [26] introduced the 3D Co-Expression Matrix Analysis (3DCEMA) method, employing 3D convolutional neural networks to predict regulatory relationships between genes. This approach helps minimize the effects of noise and data loss, significantly enhancing the accuracy of gene regulatory network inference compared to existing algorithms.

However, the main drawback of these methods is the high cost of data collection, particularly for complex gene expression data like scRNA-seq, which limits their widespread practical application.

*3) Heterogeneous network-based methods:* With the rapid development of databases, a wealth of data on TF-target gene interactions has been collected from experiments and integrated into resources like the TRRUST database [7], providing extensive insights into human gene regulatory networks. Heterogeneous network-based methods offer a novel approach to directly predict TF-target gene interactions more effectively than TFBS or gene expression-based methods. These methods go beyond simply predicting binding sites by leveraging contextual biological factors and disease mechanisms that influence binding.

For instance, Y. A. Huang et al. [24] introduced a new deep learning model named HGETGI to predict TF-target gene interactions. HGETGI not only learns known interaction patterns between TFs and target genes but also integrates information on their roles in human pathological mechanisms. Using random walk sampling with meta-paths and skip-gram node embedding techniques, HGETGI achieved high prediction accuracy, with an average AUC of 0.8519 ± 0.0731 through five-fold cross-validation. Similarly, Z. H. Du et al. [30] proposed the GraphTGI model, which employs a graph-structured neural network to predict TF-target gene interactions, achieving an average AUC of 88.64% through five-fold cross-validation, proving its effectiveness in TF-target gene interaction prediction. GraphTGI is the first end-to-end model to incorporate the topological structure of the TF-target gene interaction network, alongside the chemical properties of genes in node features, creating automated embeddings that clarify relationships between TF-target gene pairs and support related tasks.

While these methods have made significant strides, challenges remain. Current approaches mainly focus on predicting TF-target gene interactions without optimizing for data balance, which hampers accurate predictions with uneven datasets.

This paper introduces a novel approach to address data imbalance in order to optimize the prediction of TF-target gene interactions. Numerous studies have proposed solutions for handling imbalanced data classification through various approaches, including data-level and algorithm-level strategies. In this study, we adopt a data-level approach, focusing on preprocessing to reduce imbalance before feeding data into the TF-target gene interaction prediction model to achieve better results. Several methods exist for adjusting data, such as oversampling or under-sampling. Moreover, combining these methods can further optimize classification and improve prediction performance [9].

Under-sampling is an effective method for handling imbalanced data by reducing the number of samples in the majority class to balance it with the minority class, thereby improving the predictive capability of the model. Among the under-sampling methods, Random Under-sampling (RUS) is a simple technique applied to balance datasets by randomly removing a number of samples from the majority class. However, such random data removal may lead to the loss of valuable samples and diminish the amount of useful information from the majority class, potentially negatively impacting performance in classification tasks. Therefore, C. M. Huang et al. [1] proposed an approach using K-means to select representative samples from the majority class, improving precision (PPV) by 20.2% while maintaining recall above 90% on Kawasaki Disease (KD) data. Q. Zhou et al. [19] suggested an adaptive K-means-based under-sampling method, where they calculate the distance between data points within each cluster and the cluster centroid using Manhattan distance and Cosine similarity. This algorithm employs these two metrics to select representative samples from the majority class, resulting in a more balanced dataset. The results indicate that this method determines an appropriate dynamic value of *k* for different datasets and generates a balanced dataset, thereby enhancing the classification performance of machine learning algorithms. T. Doan et al. [22] proposed GBDTLRL2D, a method for predicting lncRNA-disease relationships that combines Gradient Boosting Decision Trees (GBDT) and Logistic Regression, utilizing MetaGraph2Vec and K-means to preserve semantic features, achieving an average AUC of 0.98 in 10-fold cross-validation.

In contrast to under-sampling methods, over-sampling methods focus on increasing the number of samples in the minority class. The simplest over-sampling method is Random Oversampling, which involves randomly duplicating samples from the minority class to increase the number of samples in this class, thereby creating a balance with the majority class. However, this approach can easily lead to overfitting, reducing the generalization ability of the model. To mitigate this risk, N. V. Chawla et al. [17] proposed SMOTE, which generates synthetic samples through interpolation from the minority class data. However, SMOTE may not accurately reflect the complex characteristics of the minority class, especially in intricate models. D. X. Tho et al. [6] improved this approach with KNN-SMOTE, achieving superior performance in F-score, G-mean, and AUC on imbalanced datasets from UCI. H. Li et al. [9] proposed KM-GAN, which combines K-means and GAN to generate new samples from imbalanced industrial fault data. KM-GAN clusters the minority class samples and then utilizes GAN to create additional data, enhancing diagnostic efficacy through a combined DNN and DBN model, thus addressing the bias of traditional methods towards the majority class.

Current methods have effectively contributed to identifying interactions between TFs and target genes; however, they still face several challenges, such as a high false positive rate and significant data collection costs. Additionally, heterogeneous network-based methods have not been optimized for imbalanced data. Although many methods for handling imbalance, as introduced above, have achieved good predictive performance, further improvements are still necessary. To address these limitations and enhance the quality of the majority class, we have developed a new method called FKMU to improve the performance of predictive models. We anticipate that this method will enhance accuracy and applicability in empirical research.

## III. Methodology

In this section, we will outline the main tasks of our method aimed at predicting the relationships between TFs and target genes. As part of this narrative, we describe a heterogeneous network formed from biological databases related to TFs, target genes, and diseases, as shown in Step 1 of Fig. 1. We will perform data balancing by combining the K-means clustering algorithm with negative sampling, as detailed in Step 2 of Fig. 1. We will create new meta-paths, as illustrated in Step 3 of Fig. 1. Random walks will be conducted on the graph according to the meta-paths to generate training data for the embedding model, followed by the application of a deep learning model to learn the features of the nodes in the heterogeneous network, as shown in Steps 4 and 5 of Fig. 1. Finally, we will proceed to predict the interactions between TFs and target genes, as illustrated in Step 6 of Fig. 1.

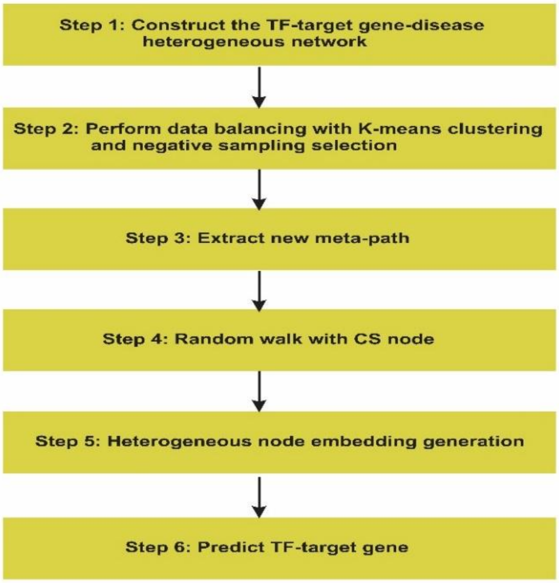

Step 1: Construct the TF-target gene-disease heterogeneous network
Step 2: Perform data balancing with K-means clustering and negative sampling selection
Step 3: Extract new meta-path
Step 4: Random walk with CS node
Step 5: Heterogeneous node embedding generation
Step 6: Predict TF-target gene

Fig. 1. General workflow containing six main steps.

### A. *Heterogeneous Network Construction*

Definition 1. A heterogeneous network [23] is defined as graph G = (V, E, T), where each node v and each edge e are associated with mapping functions $\phi(v): V \rightarrow T_V$ and $\varphi(e): E \rightarrow T_E$ , respectivelt. $T_V$ and $T_E$ represent the set of object types and relationship types, and satisfy the condition $|T_V| + |T_E| > 2$.

In this study, we construct a model to predict the associations between TFs and target genes. This heterogeneous network is defined as a graph G = (V, E, T), where each node represents TFs, target genes, or diseases, and each edge represents the relationships between these entities. TFs and target genes have been shown to have close associations with various diseases, and integrating information about these entities allows us to explore potential unknown associations between TFs and target genes.

### B. *Meta-Path in a Heterogeneous Network*

A meta-path, also known as a "hyperlink" is a model used to represent relationships between nodes in a heterogeneous network. It can be understood as a sequence of connections between nodes and their links, designed to express the relationship between two nodes under consideration within the network.

Definition 2. Meta-path [5]. A meta-path $\mathcal{P}$ is a path defined on the network schema $T_G = (\mathcal{A}, \mathcal{R})$ and is represented as $V_1 \overset{R_1}{\rightarrow} V_2 \overset{R_2}{\rightarrow} \dots \overset{R_l}{\rightarrow} V_{l+1}$, defining a composite relationship $R = R_1 \circ R_2 \circ \dots \circ R_l$ between the types $V_1$and $V_{l+1}$, where ∘ denotes the composition operator over relationships.

For example, the meta-path "CTCF (TF) - BRCA1 (Target Gene) - Breast Cancer (Disease) - TP53 (Target Gene) - MYC (TF) - Ovarian Cancer (Disease) - EGFR (Target Gene) - SP1 (TF)", as illustrated in Fig. 2, demonstrates how the transcription factors CTCF, MYC, and SP1 are linked to breast and ovarian cancers through intermediate target genes BRCA1, TP53, and EGFR.

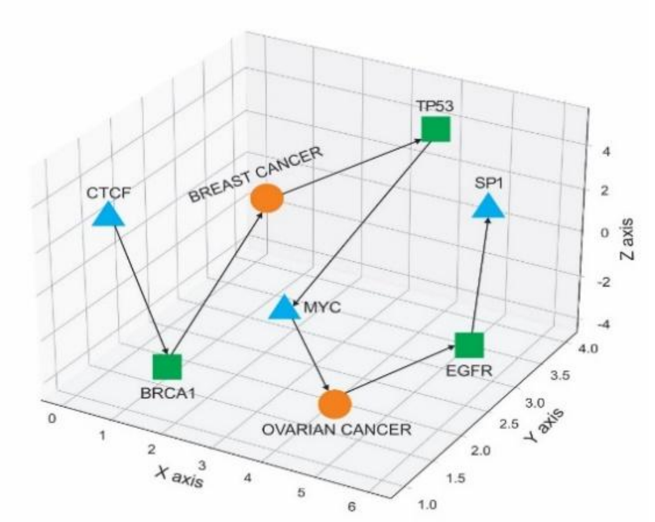


Fig. 2. Illustration of a meta-path in a heterogeneous TF-target gene-disease network.

1) *Meta-path random walks:* is a graph mining technique that generates paths based on the semantic and structural relationships between different types of nodes. This method helps transform the complex structure of the network into vectors, enabling effective extraction of information from the relationships.

For a heterogeneous network *G = (V, E, T)* and a meta-path schema $\mathcal{P}$, we can calculate the transition probability at step *k* as follow:

$$P(v^{k+1} | v_t^k, \mathcal{P}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^k)|} (v^{k+1}, v_t^k) \in E, \emptyset(V^{k+1}) = t+1 \\ 0 \quad (v^{k+1}, v_t^k) \in E, \emptyset(V^{k+1}) \neq t+1 \\ 0 \quad (v^{k+1}, v_t^k) \in E \end{cases} \tag{1}$$

where $v_t^k \in V_t$ and $N_{t+1}(v_t^k)$ denotes the type $V_{t+1}$ of the neighborhood of node $v_t^k$.

### C. *Dataset*

In this study, we use a dataset consisting of three types of nodes: TFs, target genes, and diseases, along with three types of relationships between these nodes [24]. Specifically, the three types of relationships include: the association between TFs and target genes, the association between TFs and diseases, and the association between target genes and diseases (Fig. 3).

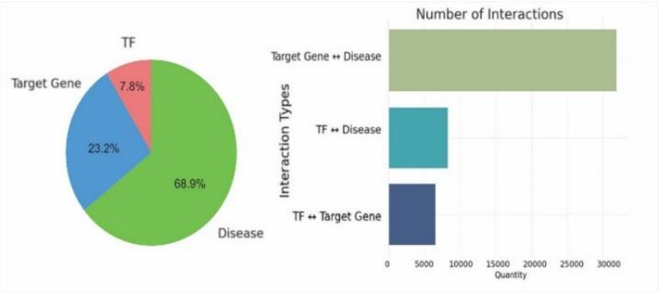


Fig. 3. Statistics of the heterogeneous TF-target gene-disease network information.

Data on interactions between human TFs and target genes were collected from the TRRUST database. This is a transcriptional regulatory network database that utilizes text mining techniques to gather and manually verify detailed information on interactions between TFs and human target genes, ensuring data accuracy. During processing, duplicate pairs were removed, resulting in a final dataset of 6,542 interactions between 696 TFs and 2,064 target genes. Additionally, these transcription factors and target genes were linked to diseases through the DisGeNET database, a resource focused on the genetic basis of human diseases. As a result, 8,199 links between TFs and diseases, along with 31,895 links between target genes and diseases, covering 6,121 different disease types, were collected.

### D. *Data Balancing Solution with K-Means Clustering and Negative Sample Selection*

In this study, the dataset includes 696 TFs and 2,064 target genes, with a total of 1,436,544 TF-target gene pairs. As shown in Fig. 4, only 0.46% of the TF-target gene pairs have been identified as interactions, while the vast majority, accounting for

99.54%, are unknown interactions. This substantial imbalance highlights a severe data imbalance, posing a significant challenge for the prediction model, which can lead to bias and reduced accuracy. Therefore, we have applied data sampling methods to balance the dataset, thereby enhancing the prediction model's accuracy.

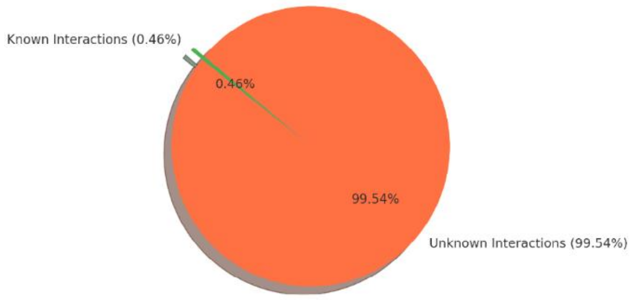


Fig. 4. Distribution of interactions between TF and target genes in the dataset.

In this study, we introduce the FKMU method, a K-means clustering-based Under-sampling technique for selecting negative samples. This method selects samples with the lowest occurrence frequency of TFs in each cluster, based on the inverse information principle as described in Fig. 5. The FKMU procedure is as follows:

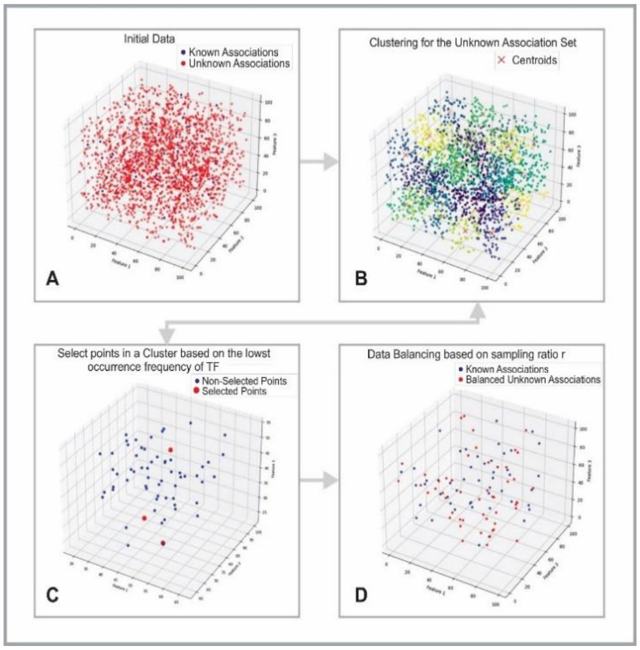


Fig. 5. The process of balancing the dataset using FKMU. A) initial known associations and unknown associations; B) Clustering for the set of unknown associations based on the feature matrix; C) Select points in a cluster based on the lowst occurrence frequency of TF; D) Data balancing based on sampling ratio *r*.

*1) Identify the known and unknown association sets from the TF-target gene adjacency matrix A*: The known association set *K* consists of TF and target gene pairs with a value of 1 in matrix *A*, representing the associations that have been confirmed. The unknown association set *U* consists of TF and target gene pairs with a value of 0 in matrix *A*, representing the associations that have yet to be evaluated.

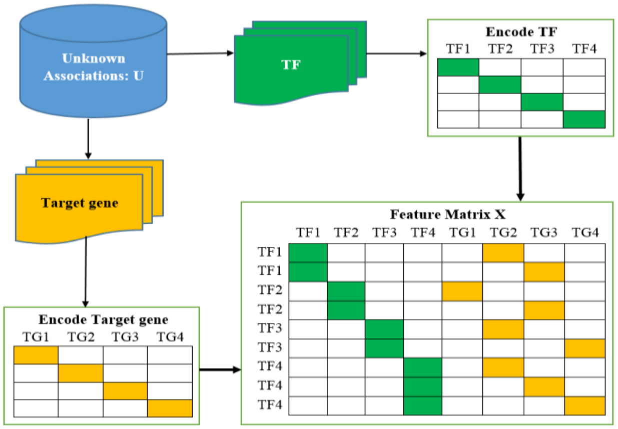
*2) One-hot encoding for the unknown association set*: Each pair (*i*, *j*) from the set *U* will be encoded into a feature vector using one-hot encoding for the following factors:

*a) TF encoding:* A value of 1 at position *i* corresponds to the TF.

*b) Target gene encoding:* A value of 1 at position *j* corresponds to the target gene.

*c) Create feature matrix X:* The matrix *X* is defined with dimensions |*U*| × (*m* + *n*), where each row is the one-hot vector of a pair (*i*, *j*) in *U*.

This process is illustrated in Fig. 6, which demonstrates the one-hot encoding structure for unknown associations.



Encode TF

| TF1 | TF2 | TF3 | TF4 |
|---|---|---|---|
| 🟩 | | | |
| | 🟩 | | |
| | | 🟩 | |
| | | | 🟩 |

Encode Target gene

| TG1 | TG2 | TG3 | TG4 |
|---|---|---|---|
| 🟨 | | | |
| | 🟨 | | |
| | | 🟨 | |
| | | | 🟨 |

Feature Matrix X

| | TF1 | TF2 | TF3 | TF4 | TG1 | TG2 | TG3 | TG4 |
|---|---|---|---|---|---|---|---|---|
| TF1 | 🟩 | | | | | 🟨 | | |
| TF1 | 🟩 | | | | | | 🟨 | |
| TF2 | | 🟩 | | | 🟨 | | | |
| TF2 | | 🟩 | | | | | 🟨 | |
| TF3 | | | 🟩 | | | 🟨 | | |
| TF3 | | | 🟩 | | | | | 🟨 |
| TF4 | | | | 🟩 | | 🟨 | | |
| TF4 | | | | 🟩 | | | 🟨 | |
| TF4 | | | | 🟩 | | | | 🟨 |

Fig. 6. One – hot encording for unknown associations.

*3) Perform K-means clustering on the unknown association set*: Apply K-means clustering to divide the unknown associations in *U* into *k* clusters. The feature matrix from step 2 is used to identify the cluster structure in the data, resulting in groups of unknown associations that share similar characteristics. The feature matrix is a sparse matrix. Therefore, before proceeding with the clustering, this matrix is represented in CSR (Compressed Sparse Row) format. CSR is one of the popular formats for storing sparse matrices. It stores the matrix by retaining non-zero values, which helps save memory for large matrices that contain many zero values.

*4) Calculate the number of associations to select from each cluster*: Based on the total number of known associations |*K*| and the sampling ratio *r*, determine the number of unknown associations $n_p$ to be selected from each cluster. This helps ensure a balance between known and unknown associations, according to the specified ratio.

*5) Select the least frequent samples in each cluster*: Within each cluster, calculate the frequency of occurrence of each TF and sort the associations in the cluster by the ascending frequency of the TF. Next, select $n_p$ associations with the lowest frequency of TF occurrence in each cluster to minimize the bias caused by the high frequency of certain dominant TFs. This approach helps create a set of associations that follows the principle of inverse frequency, prioritizing less common associations to ensure that the sample dataset contains diverse types of associations. The goal of this principle is to enrich the information in the sample set, enabling the model to learn from both rare samples and those that are less biased, thereby enhancing the model's generalization ability for rare cases in real-world data.

*6) Return the balanced association set*: The set of least frequent associations selected from all clusters forms set *B*, which is the unknown association set balanced according to the ratio *r*. Set *B* is then used as a more balanced dataset for the subsequent steps of the predictive model.

**Algorithm 1:** Frequency-Based K-Means Under-sampling Algorithm

```
Input:
    • Association matrix $A \in \mathbb{R}^{m\times n}$, where m is the number of TFs and n is the number of target genes.
    • $k$: The number of clusters for performing K-means clustering.
    • $r$: The sampling ratio from unknown associations.
Output:
    • Balanced set of unknown associations B, sampled according to ratio r.
# Calculate known and unknown set from A:
K = $\{(i,j) \mid A_{ij} = 1\}$ (known association)
U = $\{(i,j) \mid A_{ij} = 0\}$ (unknown associations)
# One-hot encoding for TF and Target gene:
Create a feature matrix $X \in \mathbb{R}^{|U|\times(m+n)}$ # where each row corresponds to an unknown association pair (i, j)
for each pair $(i,j) \in U$ do
    $X_u = [0, \dots, 1_i, \dots, 0, 0, \dots, 1_j, \dots, 0]$ # Here, the 1 at position i represents the TF and the 1 at position j represents the Target gene, while the remaining positions are 0.
end for
Initialize k random cluster centers {$u_1, u_2, \dots, u_k$} from U
# Assign each data point x ∈ U to the nearest cluster center:
for $x \epsilon U$ do
    $c_i = \arg\min_j \|x - u_j\|^2$
end for
# Update the center of each cluster with $C_j$ being the set of data points belonging to cluster j:
for i = 1 to k do
    $u_j = \frac{1}{|C_j|}\sum_{x\in C_j} x$
end for
# Calculate the number of points to select from each cluster:
$n_{known} = |K|$
$n_p = \left\lceil \frac{n_{known}\times r}{k} \right\rceil$
Initialize B = []
# Select points from each cluster:
for i = 1 to k do
    # Calculate the frequency of each TF in the cluster:
    $freq(i) = \sum_{(i,j)\in C_j} 1$
    # Sort points in cluster $C_j$ by the increasing frequency of their TFs:
    $C_j' = sorted(C_j,\ key = \lambda x{:}freq(x_0))$ # where $x_0$ is the TF index in the pair (TF, Target gene)
    # Select $n_p$ points with the lowest TF frequency from each cluster:
    $B = B \cup C_j'\,[1{:}\,n_p]$
end for
# Balance the set of unknown associations:
$B = B\ [1{:}\ [n_{known} \times r]]$
Return B
```

### *E. Embedding Heterogeneous Network Nodes using Skip-Gram*

Specifically, in a heterogeneous network *G = (V, E, T)* with the number of node types $|T_V| > 1$, he objective is to maximize the co-occurrence probability *p* of the nodes within the same context window *k*, as follows [31]:

$$\underset{\theta}{\operatorname{argmax}} \sum_{v\epsilon V}\sum_{t\epsilon T_V}\sum_{c_t\epsilon N_{t(v)}} logp(c_t|v;\theta) \qquad (2)$$

where $N_{t(v)}$ is the set of neighboring nodes with node $v$ in the heterogeneous context with different node types, and $p(c_t|v;\theta)$ is defined as a softmax function [27] as follows:

$$p(c_t|v;\theta) = \frac{\exp(X_{c_t}X_v)}{\sum_{u\epsilon V}\exp(X_uX_v)} \qquad (3)$$

where $v$ and $c_t$ are the center node and the nodes in the scanning window, respectively, and, $X_v$ is the embedding vector of node $v$.

The number of nodes is often very large, so negative sampling techniques are commonly applied to approximate the estimation of probabilities. This method maximizes the probability such that the target node does not appear simultaneously with a randomly selected negative node. The ultimate maximization goal is expressed as follows:

$$O(X) = log\sigma(F(X_{c_t}||X_v) + log\sigma(-F(X_u||X_v)) \qquad (4)$$

where *F* represents the fully connected layer, || denotes the concatenation of the embedding vectors of the nodes, and σ(x) is calculated as follows:

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad (5)$$

## IV. Results

### *A. Evaluation Criteria*

To evaluate the performance of the proposed method, we applied *k*-fold cross-validation (*k* = 5). Specifically, the data was randomly divided into *k* approximately equal parts. In each iteration, one part was used as the test set, while the model was trained on the remaining *k*-1 parts. This process was repeated *k* times, ensuring that each part of the data was used as the test set once.

To demonstrate the effectiveness of the proposed method during $k$-fold cross-validation, we used the Area Under the ROC Curve (AUC) [24], which is calculated as follows:

TABLE I. COMPARISON OF PERFORMANCE BASED ON AUC DURING 5-FOLD CV WITH DIFFERENT PARAMETER SETS

| Number of walkers | 100 | 250 | 350 | 450 | 550 |
|---|---|---|---|---|---|
| Path length | 50 | 80 | 130 | 130 | 150 |
| Dimension of embedding | 128 | 200 | 300 | 450 | 550 |
| Average AUC | 0.9199 ± 0.0064 | **0.9388 ± 0.0045** | 0.9345 ± 0.0064 | 0.9100 ± 0.0038 | 0.8828 ± 0.0140 |

$$AUC = \frac{\sum_{e \epsilon e^+} Rank_e - \frac{|e^+| \times (|e^+ + 1|)}{2}}{|e^+| \times |e^-|} \tag{6}$$

where $e^+$ and $e^-$ represent the positive and negative samples, respectively, in the test set, and $Rank_e$ denotes the rank of edge $e$ based on the predicted score.

We conducted experiments with different values for three parameters: number of walkers, path length, and dimension of embedding, while comparing the corresponding prediction results when varying each parameter. The average AUC value for each experiment is presented in Table I. The best prediction results were achieved when the number of walkers was 250, the path length was 80, and the dimension of embedding was 200. The corresponding ROC curve is illustrated in Fig. 7, showing that our proposed method achieved an average AUC value of 0.9388 ± 0.0045 through 5-fold cross-validation.

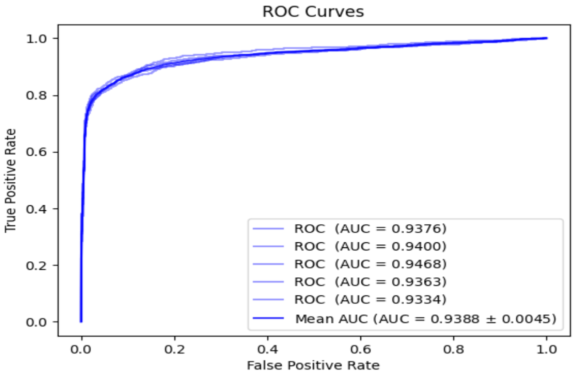


Fig. 7. ROC curve through 5-fold cross-validation.

## B. Determine the Optimal Number of Clusters

Determining the optimal number of clusters $k$ is a crucial factor in the effectiveness of the K-means algorithm. Choosing $k$ too small can lead to data points being grouped together, overlooking significant differences between clusters. Conversely, if $k$ is too large, the data may be unnecessarily divided into clusters, reducing generalization. For our experiment, we applied the Elbow method with values ranging from 10 to 200. Specifically, for each $k$ value, the K-means algorithm was executed, and the WCSS (Within-Cluster Sum of Squares) was calculated.

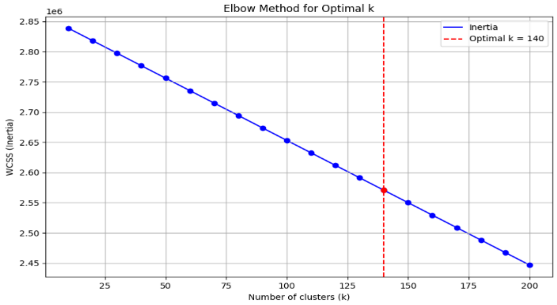


Fig. 8. The elbow method for selecting the optimal number of clusters *k*.

The results in Fig. 8 show that as $k$ increases from 10 to around 140, the WCSS decreases significantly, indicating that increasing the number of clusters improved data clustering. However, after $k$ = 140, the WCSS begins to decrease more slowly, suggesting that adding more clusters no longer provides substantial benefits for data separation. Therefore, we chose $k$ = 140 as the optimal number of clusters, as it achieves a balance between reducing WCSS and maintaining model generalization.

## C. Negative Sampling Rate

Choosing a balance ratio $r$ between Unknown Associations and Known Associations is aimed at optimizing the model's performance in distinguishing between these two groups. A reasonable balance ratio enables the model to learn the characteristics of both groups without bias, thereby achieving the highest AUC value and ensuring accuracy when applied to new data.

TABLE II. THE BALANCE RATIO BETWEEN UNKNOWN ASSOCIATION AND KNOWN ASSOCIATION

| Balance Ratio r<br>(Unknown : Known Association) | Average AUC Value |
|---|---|
| 1:1 | 0.9388 ± 0.0045 |
| 2:1 | 0.9264 ± 0.0028 |
| 3:1 | 0.9229 ± 0.0047 |
| 4:1 | 0.9275 ± 0.0055 |
| 5:1 | 0.9273 ± 0.0041 |
| 6:1 | 0.9212 ± 0.0025 |
| 7:1 | 0.9231 ± 0.0030 |
| 8:1 | 0.9229 ± 0.0041 |
| 9:1 | 0.9210 ± 0.0050 |
| 10:1 | 0.9175 ± 0.0017 |

When conducting experiments with different values of the balance ratio *r*, we obtained the results shown in Table II. The results indicate that when the balance ratio is 1:1, the AUC value reaches its highest point at 0.9388 ± 0.0045, suggesting that this is the optimal balance ratio for effectively distinguishing between Unknown Associations and Known Associations. As the balance ratio increases from 2:1 to 10:1, the AUC value gradually decreases, particularly at a ratio of 10:1, where the AUC value significantly drops to 0.9175 ± 0.0017. This indicates that the model's effectiveness diminishes when Unknown Associations constitute too large a proportion compared to Known Associations.

### D. The Impact of Selecting Meta-Paths

The random walk strategy based on meta-paths ensures that the model accurately integrates the semantic relationships between different types of nodes. Utilizing different meta-path schemas to generate sequences of nodes can capture the diverse semantic and structural relationships among these node types.

In this experiment, we designed a new meta-path schema “TF-Target gene-Disease/CS-Target gene-TF-Disease/CS-Target gene-TF” while also using the original meta-path “TF-Target gene-Disease/CS-Target gene-TF” [24] to conduct the random walk process and evaluate the prediction effectiveness of each schema. The results in Table III show that the new schema achieves a slightly higher average AUC value (0.9388 ± 0.0045) compared to the original schema (0.9366 ± 0.0044), suggesting that the new schema can improve predictive performance by capturing additional potential links within the heterogeneous network. Here, CS (Cold Start node) is a node added to the paths to address the cold start problem in the model. This issue arises when certain nodes (especially TFs or target genes) have no links to any target genes in the training data, making it difficult to learn embedding vectors for these nodes. By adding the CS node and setting its embedding to a vector where all elements have a value of 1, the model can learn information from paths containing the CS node, helping to mitigate the lack of link data for these nodes and enhance the overall performance of the model across the heterogeneous network.

TABLE III. COMPARING AUC PERFORMANCE ACROSS DIFFERENT META-PATHS

| Meta-paths | Average AUC Value |
|---|---|
| TF-target gene-disease/CS-target gene-TF | 0.9366 ± 0.0044 |
| TF-target gene-disease/CS-target gene-TF-disease/CS-target gene-TF | 0.9388 ± 0.0045 |

### E. Predicted Scores for TF-Target Gene Pairs

After training the model, we obtain low-dimensional embedding vectors for TFs and target genes. From this, we create an embedding matrix *M* for TFs and an embedding matrix *G* for target genes. The predicted scores for the interactions between TFs and target genes are determined as follows:

$$P = M.G^T \tag{7}$$

where the value in the *i-th* row and *j-th* column represents the interaction score between the *i-th* TF and the *j-th* target gene.

### F. Analysis and Comparison with Recent Studies

To evaluate the superior performance of the proposed model, we compared its predictive capabilities with recent studies, including Metapath2vec [25], HGETGI [24], and GraphTGI [30]. The comparison results shown in Table IV indicate that our method achieves the highest average AUC value of 0.9388, outperforming the other three methods. This confirms that our model is highly effective in predicting unobserved target genes for specific TFs.

Although models such as Metapath2vec, HGETGI, and GraphTGI effectively utilize heterogeneous graphs, particularly for predicting TF-target gene interactions, GraphTGI demonstrates impressive performance with an AUC of 88.64% in five-fold cross-validation, while HGETGI excels in leveraging semantic information through graph embeddings. However, all three methods lack robust mechanisms for handling imbalanced data and selecting negative samples, which can limit their ability to optimize performance on complex and highly imbalanced datasets. FKMU addresses these challenges through a K-means-based under-sampling strategy, ensuring a balanced dataset and enhancing the robustness of the model. Furthermore, the introduction of a novel meta-path allows FKMU to capture semantic relationships within the graph more effectively and optimize the detection of potential interactions. These advancements establish FKMU as an effective and superior method for predicting TF-target gene interactions while offering broad applicability to more complex and diverse problems, especially for large-scale and heterogeneous datasets in the future.

TABLE IV. COMPARING THE PERFORMANCE OF RESEARCH METHODS

| Methods | Average AUC Value |
|---|---|
| Metapath2vec [25] | 0.8239 ± 0.0057 |
| HGETGI [24] | 0.8519 ± 0.0731 |
| GraphTGI [30] | 0.8864 ± 0.0057 |
| FKMU | 0.9388 ± 0.0045 |

### G. Case Study

To evaluate the predictive performance of the model in identifying potential target genes associated with TFs, we conducted experiments on the transcription factors CTCF and TP53. Specifically, we removed the links between the specific TFs used in the experiments and their target genes. We then reconstructed the heterogeneous network. Finally, we trained the model and tested it for each specific TF case to assess the model's performance.

The transcription factor CTCF (CCCTC-binding factor) is an important protein involved in regulating the structure and function of the genome. CTCF binds to DNA sequences to create insulator regions and chromatin loops, helping to regulate gene activity. CTCF can activate or repress genes depending on its binding location. Mutations in the CTCF gene are associated with various diseases such as cancers (breast, colorectal, prostate), neurodevelopmental disorders (Bardet-Biedl syndrome, autism spectrum disorders), and rare genetic diseases, primarily due to disruptions in chromatin structure and gene regulation.

The transcription factor TP53 (tumor protein p53) is an important gene, often referred to as the “guardian of the genome” due to its key role in maintaining genetic stability and preventing tumor formation. This gene encodes the p53 protein, a tumor suppressor that plays a crucial role in controlling cell division, repairing damaged DNA, and activating apoptosis when cells sustain irreparable damage. TP53 mutations are a common cause in many types of cancer, including lung cancer, breast cancer, colorectal cancer, and skin cancer. Research on TP53 not only elucidates the mechanisms of cancer but also opens new avenues for treatments aimed at restoring p53 function to prevent the development of cancer cells.

TABLE V. TOP 20 TARGET GENES FOR CTCF

| Target gene | CTCF-Related Target |
|---|---|
| CDKN1A | Confirmed |
| MTHFR | Unconfirmed |
| VEGFA | Confirmed |
| TNF | Confirmed |
| SOD2 | Confirmed |
| IL6 | Confirmed |
| PTGS2 | Confirmed |
| BCL2 | Confirmed |
| CCND1 | Confirmed |
| CDKN2A | Unconfirmed |
| MMP9 | Confirmed |
| KRAS | PMID:32374727 |
| CDH1 | Confirmed |
| IFNG | Unconfirmed |
| TGFB1 | Confirmed |
| TERT | Confirmed |
| PTEN | Confirmed |
| NOS2 | Confirmed |
| ERBB2 | Confirmed |
| IL1B | Unconfirmed |

TABLE VI. TOP 20 TARGET GENES FOR TP53

| Target gene | TP53-Related Target |
|---|---|
| CDKN1A | Confirmed |
| VEGFA | Confirmed |
| MTHFR | Confirmed |
| IL6 | Confirmed |
| TNF | Confirmed |
| SOD2 | Confirmed |
| CCND1 | Confirmed |
| PTGS2 | Unconfirmed |
| BCL2 | Confirmed |
| IFNG | Unconfirmed |
| TGFB1 | Confirmed |
| IL1B | PMID:34986125 |
| CDH1 | Confirmed |
| MMP9 | Unconfirmed |
| KRAS | Confirmed |
| CDKN2A | Confirmed |
| ABCB1 | Confirmed |
| TERT | PMID:23284306 |
| ERBB2 | Confirmed |
| EGFR | Confirmed |

We ranked the predicted scores based on the weighted matrix to identify potential target genes. Then, we assessed the accuracy of these target genes by comparing them with the hTFtarget database [18]. Specifically, we focused on testing and validating the top 20 predicted target genes to ensure the reliability and accuracy of the predictive model. This process helps us confirm the model's capability in identifying potential target genes associated with the TFs.

The experimental results are presented in Table V for CTCF and in Table VI for TP53, respectively. According to these tables, 75% (15/20) of the predicted target genes have been validated against the hTFtarget dataset. Additionally, we conducted supplementary research and discovered genes such as KRAS, which, although not listed as interacting with CTCF in hTFtarget, have been reported to interact with CTCF in other studies, as indicated by the PMID [11] codes in Table V. Similarly, genes such as IL1B and TERT were also found to interact with TP53, as shown by the PMID [14, 16] codes in Table VI. These results demonstrate the effectiveness of the proposed method.

## V. CONCLUSION

Predicting interactions between transcription factors and target genes remains a significant challenge, particularly in the context of the complex relationships within the gene regulatory network that have not been fully explored. To address this issue, we propose the FKMU method, a novel approach aimed at handling data imbalance when predicting interactions between TFs and target genes. FKMU combines K-means clustering and inverse information principles to select underrepresented samples in the dataset, thereby balancing sample ratios and improving the accuracy of the model. This method applies the K-means algorithm to partition unknown samples into clusters, subsequently prioritizing TFs with low occurrence frequency in each cluster to enhance diversity and representation within the data. Experimental results on real datasets demonstrate that FKMU achieves superior performance in accurately predicting interactions between TFs and target genes compared to current methods, with a significantly higher average AUC value. We expect that the FKMU method will pave the way for new avenues in scientific research, improving the handling of imbalanced data and enhancing the accuracy of predictive models in the biomedical field.

## REFERENCES

[1] C. M. Hoang et al., “A K-means Clustering Based Under-Sampling Method for Imbalanceed Dataset Classification,” 2024 International Conference on Information Networking (ICOIN), pp. 708-713, 2024, doi: 10.1109/ICOIN59985.2024.10572133.

[2] D. N. Anh, B. D. Hung, P. Q. Huy, and D. X. Tho, “Feature analysis for imbalanced learning,” Journal of Advanced Computational Intelligent Informatics, vol. 24, no. 5, pp. 648–655, Sep. 2020.

[3] D. X. Tho, B. D. Hung et al., “Prediction of autism-related genes using a new clustering-based under-sampling method,” In: 2019 11th International Conference on Knowledge and System Engineering (KSE), pp. 1-6, IEEE, 2019.

[4] D. X. Tho, D. N. Anh, “Imbalance In the Learning Chest X-Ray images For COVID-19 detection,” In: Soft Computing: Biomedical and Related Applications, pp. 107-119. Springer Berlin Heidelberg, 2021.

[5] D. X. Tho, L. M. Hung, and D. N. Anh, “Drug Repositioning for Drug Disease Association in Meta-paths,” In Deep Learning and Other Soft Computing Techniques: Biomedical and Related Applications, 2023, pp. 39-51, Cham: Springer Nature Switzerland.

[6] D. X. Tho, and T. T. Le, “KNN-SMOTE: An Innovative Resampling Technipue Enhancing the Efficary of Imbalanced Biomedial Classification,” Machine Learning and Other Soft Computing Techniques: Biomedical and Related Applications, pp. 111-121, 2024, doi: 10.1007/978-3-031-63929-6_11.

[7] H. Han, J. W. Cho, S. Lee, A. Yun, H. Kim, D. Bae at al., “TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions,” Nucleic Acids Research., vol. 46, pp. D380-D386, 2017, doi: 10.1093/nar/gkx1013.

[8] H. He, M. Yang, S. Li, G. Zhang et al., “Mechanisms and biotechnological applications of transcription factor,” Synthetic and Systems Biotechnology, Vol. 8, pp. 565-577, 2023.

[9] H. Li, R. Fan, Q. Shi, and Z. Du, “Class Imbalanced Fault Diagnosis via Combining K-Means Clustering Algorithm with Generative Adversarial Networks,” Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 25, No.3, pp. 346-355, 2021.

[10] J. Lachantin, R. Singh, B. Wang et al., “Deep motif dashboarch: Visualizing and understanding genomic sequences using deep neural network,” Pac Symp Biocomput., vol. 22, pp. 254–265, 2017, doi:10.1142/9789813207813_0025.

[11] J. Rinal, E. S. Sokol, R. J. Hartmaier, S. E. Trabucco et al., “The genomic landscape of metastatic breast cancer: Insights from 11,000 tumors,” PloS One, Vol. 15, No. 5, e0231999, 2020, PMID:32374727.

[12] J. T. Wade, “Mapping Transcription Regulatory networks with CHIP-seq and RNA-seq,” Adv Exp Med Biol, vol. 883, pp. 119–134, 2015, doi: 10.1007/978-3-319-23603-2_7.

[13] J. Wang, “TF-Target Finder: An R Web Application Bridging Multiple Predictive Models for Decoding Transcription Factor-Target Interactions,” Preprints.org, 2024, doi: 10.20994/preprints202404.1212.v1.

[14] K. Gao, Y. Zhu, H. Wang, X. Gong et al., “Network Pharmacology reveals the potential mechanism of Baiing Qinghou decoction in treating laryngeal squamous cell carcinoma,” Aging (Albany NY), vol. 13, no. 24, pp. 26003-26021, 2021, PMID:34986125.

[15] K. Su, A. Katebi, V. Kohar, B. Clausss, D. Gordin, Z. S. Qin et al., “NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity,” Genome Biology, vol. 23, no. 1, pp. 1-21, 2022.

[16] L. Xie, C. Gazin, S. M. Park, Li. J. Zhu et al., “A Synthetic Interaction Screen Identifies Factors Selectively Required for Proliferation and TERT Transcription in p53-Deficient Human Cancer Cells,” Plos Genetics, Vol. 8, No. 12, e1003151, PMID:23284306.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. of Artificial Intelligence Research, Vol. 16, No. 1, pp. 321-357, 2002.

[18] Q. Zhang et al., “hTFtarget: a comprehensive database for regulations of human transcription factors and their targets,” Genomics Proteomics Bioinf., vol. 18, pp. 120–128, 2020.

[19] Q. Zhou, B. Sun, “Adaptive K-means clustering based under-sampling methods to solve the class imbalance problem,” Data and Information Management, vol. 8, no. 3, pp. 1-12, 2024, doi:10.1016/j.dim.2023.100064.

[20] R. Mundade, H. G. Ozer, H. Wei, L. Prabhu, and T. Lu, “Role of CHIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond,” Cell Cycle., vol. 13, pp. 2847–2852, 2014, doi: 10.4161/15384101.2014.949201.

[21] S. Salekin, JM. Zhang, and Y. Huang, “Based-pair resolution detection of transcription factor binding site by deep decovolutional network,” Bioinformatics., vol. 34, no. 20, pp. 3446-3453, 2018.

[22] T. Doan, Z. Kuang, J. Wang, Z. Ma, “GBDTLRL2D Predicts LncRNA-Disease Associations Using MetaGraph2Vec and K-Means Based on Heterogeneous Network,” Frontiers in Cell and Developmental Biology, vol. 9:753027, 2021, doi: 10.3389/fcell.2021.753027.

[23] T. V. Thai, B. D. Hung, D. X. Tho et al., “A New Computational Method Based on Heterogeneous Network for Predicting MicroRNA-Disease Associations,” Soft Computing for Biomedical Applications and Related Topics, 2021, pp. 205–219.

[24] Y. A. Huang et al., “Heterogeneous graph embedding model for predicting interactions between TF and Target gene,” Bioinformatics, vol. 38, no. 9, pp. 2554-2560, 2022, doi: 10.1093/bioinformatics/btac148.

[25] Y. Dong et al., “metapath2vec: Scalable representation learning for heterogeneous network,” In Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining., 2017, pp. 135–144.

[26] Y. Fan, X. Ma, “Gene regulatory network inference using 3D convolutional Neural Network,” In Proceeding of the AAAI Conference on Artificial intelligence, 2021, vol. 35, pp. 99-106.

[27] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, “Pathsim: Meta path-based top-k similarity search in heterogeneous information network,” In VLDB'11., pp. 992-1003, 2011.

[28] Y. Yang, Q. Fang, H. B. Shen, “Predicting gene regulatory interractions based on spatial gene expression data and deep learning,” PloS Comput Biol, 2019, 15(9):e1007324.

[29] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, and S. Masri, “Based-resolution models of transcription-factor binding reveal soft motif syntax,” Nature Genetics., vol. 53, pp. 345–366, 2021, doi: 10.1038/s41588-021-00782.

[30] Z. H. Du, Y. H. Wu, Y. A. Huang, J. Chen, G. Q. Pan, L. Hu, Z. H. You, and J. Q. Li, “GrapTGI: An attention-based graph embedding model for predicting TF-Target gene interactions,” Briefings in Bioinformatics, vol. 23, no. 3, pp. 1-11, 2022, doi: 10.1093/bib/bbac148.

[31] Z. Liu, S. Zhang, J. Zhang, M. Jiang, M. Liu, “HeteEdgeWalk: A Heterogeneous Edge Memory Random Walk for Heterogeneous Information Network Embedding”, Entropy, 2023, 25(998) doi:10.3390/e25070998.

[32] Z. Shen, W. Bao, and D. S. Huang, “Recurrent neural network for predicting transcription factor binding sites,” Scientificreports., vol. 8, no. 1, pp. 1–10, 2018, doi: 10.1038/s41598-018-33321-1.