Functional MicroRNA Target Evaluation Based on Disease Similarity with Sentence-Bert

Baiming Chen*[1,2] and Hsi-Yuan Huang†[1]

[1]School of Medicine, The Chinese University of Hong Kong, Shenzhen, Longgang District, Shenzhen 518172, China
[2]Shenzhen X-Institute, Nanshan District, Shenzhen, Guangdong Province 518055, China

June 23, 2024

Abstract

MicroRNAs (miRNAs) regulate gene expression by binding to target mRNAs, inhibiting translation, or promoting mRNA degradation. miRNAs are of great importance in the development of diseases. Currently, a variety of miRNA target prediction tools are available, involving analyzing sequence complementarity, thermodynamic stability, and evolutionary conservation to identify potential miRNA binding sites on the 3' untranslated region (3'UTR) of target mRNAs. We propose a further screening method for sequence-based predicted MTIs (MicroRNA Target Interactions) by considering the disease semantic similarity between miRNA and gene to establish a database of disease-specific MTIs. We fine-tuned a Sentence-BERT model to improve the performance of calculating disease semantic similarity. The method achieves a recall rate of 91% on the validation dataset, which comprises intersecting MTIs from miRTarbase and miRWalk that represent experimentally verified predicted MTIs. Additionally, the method demonstrates excellent generalizability across different datasets. The proposed method was utilized to calculate the semantic similarity of diseases in 3,900,041 experimentally validated and predicted MTIs, involving 7,727 genes and 1,263 miRNAs. The observed distribution of similarity aligns perfectly with existing biological evidence. The study holds the potential to offer valuable insights into comprehending gene regulatory networks and advancing progress in disease diagnosis, treatment, and drug development.

1 Introduction

In 1993, miRNA (Lin-4) was first discovered in humans and found to be critical for the timing of embryonic development in C. elegans [1]. Since then, a large number of miRNAs have been discovered with the development of high-throughput

*Corresponding author: baimingchen@link.cuhk.edu.cn
†First corresponding author: huanghsiyuan@cuhk.edu.cn

bioinformatics technologies [2]. miRNAs are classified as small non-coding RNAs (ncRNAs), play a crucial role in the regulation of various physiological processes, such as cell cycle [3], cell growth, development, differentiation, and apoptosis [4], as well as pathological processes [5]. In addition, miRNAs may be promising biomarker candidates for early detection or prognosis of various diseases [6]. Therefore, the identification of novel MTIs holds biological significance.

There are many existing MTIs prediction databases. miRDB utilizes a support vector machine (SVM) to identify miRNA-mRNA interactions in different species [7]. mirSVR trained a regression model to rank miRNA target sites based on down-regulation score, using data from miRanda [8]. DeepTarget uses Recurrent Neural Networks (RNN) for miRNA-mRNA interaction prediction [9]. TargetScan predicts miRNA-mRNA interaction based on seed match and conservativeness analysis in different species [10]. The RNA22 is designed for the identification of miRNA binding sites and their corresponding miRNA and mRNA complexes [11]. PITA predicts miRNA targets by considering both the binding free energy and the accessibility of the target site within the mRNA secondary structure [12]. StarBase explores and analyzes the RNA–RNA and protein–RNA interaction networks [13]. However, the availability of disease-specific MTIs prediction databases is limited.

The NLP method can be employed to address this issue. Sentence-BERT (SBERT) fine-tunes BERT using siamese and triple network structures and is an extension of the BERT (Bidirectional Encoder Representations from Transformers) model. The SBERT model demonstrates excellent performance in semantic similarity tasks [14]. Moreover, the SBERT has a faster speed compared to BERT [15] and RoBERTa [16] while maintaining accuracy.

The present study proposes a methodology for calculating the similarities of diseases between miRNAs and genes, aiming to further screen sequence-based predicted MTIs to establish a database of disease-specific MTIs. The similarity between the two diseases can be derived using the SBERT model. The subsequent step involves the utilization of clustering to group miRNA-associated diseases and gene-associated diseases based on disease similarities, followed by the calculation of either the intra-cluster mean or overall mean to obtain a comprehensive measure of similarity.

2 Materials and Methods

2.1 Overview

We present a comprehensive flow chart (Figure 1) that outlines our study's major steps and processes. The steps encompass fine-tuning the Sentence-bert model, utilizing clustering for computing disease similarity between miRNAs and genes, and classifying MTIs based on computed scores.

2.2 Data Collection

To fine-tune the SBERT model, we obtained the MeSH descriptors (nlm.nih.gov) from the National Library of Medicine and specifically selected category C, which pertains to diseases. The MeSH (Medical Subject Headings) descriptors are standardized terms utilized for indexing and retrieving biomedical and health-related information, forming a hierarchical vocabulary. To test the SBERT model performance, we obtained the BIOSSES dataset, which consists of 100 pairs of biomed-

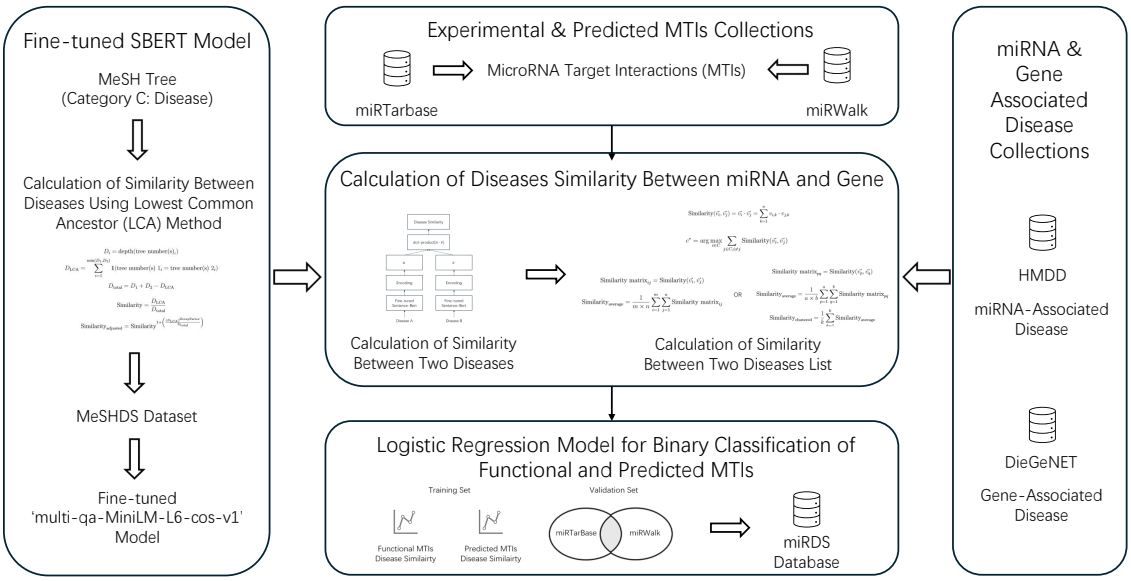


Figure 1: Flow Chart

ical statements with manually annotated text similarity scores provided by experts [17]. It can be used to evaluate text similarity measurement models in the field of biomedicine. To obtain experimental and predicted MTIs, we utilized the miRTarbase [18] and miRWalk [19]. miRTarbase provides experimentally validated MTIs. miRWalk provides sequenced-based predicted MTIs. For acquiring information on miRNA-associated and gene-associated diseases, we employed HMDD [20], DisGeNet [21] and KEGG [22]. HMDD offers miRNA-associated disease data. DisGeNet and KEGG offer gene-associated disease data, for each gene, we selected its top 40 most relevant diseases. We processed the miRTarbase and miRWalk datasets by merging the -3p and -5p of the same miRNA.

2.3 Model Construction

2.3.1 Fine-tuned SBERT Model for Disease Semantic String Analysis

For the model to have the ability to handle the biomedical string semantic similarity computation task while maintaining its original performance, we utilized the MeSH descriptors category C (diseases) to generate a disease semantic similarity dataset to fine-tune the model. To calculate the similarity between different disease semantic strings, we proposed a method that yields a distribution of values that is similar to the distribution of semantic similarity values output by the 'sentiment-transformers/multi-qa-MiniLM-L6-cos-v1' [23] model without fine-tuning. Both types of distributions follow a log-normal distribution pattern. The specific methods are as follows:

- Determine the depth of the disease descriptors in the MeSH Tree:

$$D_i = \text{depth}(\text{tree number(s)}_i) \tag{1}$$

- Determine the depth of the lowest common ancestor (LCA) between two disease descriptors:

$$D_{\text{LCA}} = \sum_{i=1}^{\min(D_1,D_2)} \mathbf{1}(\text{tree number(s) } 1_i = \text{tree number(s) } 2_i) \tag{2}$$

Parameter Definitions: tree number(s)1_i = tree number(s)2_i indicates the level at which the two diseases share a common node in their hierarchical structure. i ranges from 1 to the depth at which the two codes' lowest common ancestor (LCA) of the two tree numbers is located.

- Calculate the total depth:

$$D_{\text{total}} = D_1 + D_2 - D_{\text{LCA}} \tag{3}$$

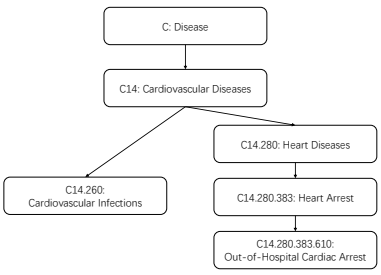
- Calculate the disease similarity:

$$\text{Similarity} = \frac{D_{\text{LCA}}}{D_{\text{total}}} \tag{4}$$

- Calculate the adjusted disease similarity:

$$\text{Similarity}_{\text{adjusted}} = \text{Similarity}^{1+\left(\frac{(D_{\text{LCA}})^{\text{DecayFactor}}}{D_{\text{total}}}\right)} \tag{5}$$

The decay factor of 0.8 was selected in this paper. According to the provided formula, the weighted disease similarity between the descriptor: Cardiovascular Infections, tree number(s): C14.260, and the descriptor: Out-of-Hosptial Cardiac Arrest, tree number(s): C14.280.383.610 is 0.14495593273553914 (Figure 2). The weighted disease similarity was adjusted to better align with the distribution of the output generated by the 'multi-qa-MiniLM-L6-cos-v1' model. The weighted disease similarity was reprocessed, resulting in a fine-tuned dataset, named 'MeSHDS'. We utilized this dataset to fine-tune the model.



$D_{\text{Cardiovascular Infections}} = 2,$
$D_{\text{Out-of-Hospital Cardiac Arrest}} = 4,$
$D_{\text{LCA}} = 1,$
$D_{\text{total}} = 5,$
$\text{Similarity} = 0.2,$
$\text{Similarity}_{\text{adjusted}} = 0.14495593273553914$

Figure 2: Example of adjusted similarity calculation between 'Cardiovascular Infections' and 'Out-of-Hospital Cardiac Arrest' with MeSH Tree

2.3.2 Disease Similarity Calculation Between miRNAs and Genes

To calculate the disease similarity between miRNA and gene. We propose an algorithm to compute the similarity of two lists of disease strings. The algorithm combines semantic string similarity calculation and clustering.

Semantic String Encoding and Similarity Calculation First, each string in the list is encoded using the Sentence-BERT model. The similarity between two semantic vectors is obtained by calculating the dot product:

$$\text{Similarity}(\vec{v_i}, \vec{v_j}) = \vec{v_i} \cdot \vec{v_j} \tag{6}$$

Parameter Definitions: $\text{Similarity}(\vec{v_i}, \vec{v_j})$ is the dot product between these vectors. $\vec{v_i}$ and $\vec{v_j}$ represent the semantic vectors of elements i and j after encoding with sentence-bert respectively.

Clustering To classify miRNA-associated and gene-associated diseases into different groups to determine if or how many diseases they associated in common. First, merge the miRNA-associated and gene-associated diseases. The first element of the merged list automatically becomes the cluster center. As iterating through the remaining elements, if an element's semantic similarity to the cluster center exceeds the threshold, the element is added to that cluster. Otherwise, it becomes a new cluster center. After clustering, the elements in each cluster are traversed to find the element with the highest similarity sum as the center of the cluster.

$$c^* = \arg\max_{i \in C} \sum_{j \in C, i \neq j} \text{Similarity}(\vec{v_i}, \vec{v_j}) \tag{7}$$

Parameter Definitions: c^* represents the cluster center.

Evaluation of Clustering Results The evaluation of clustering results is based on the number of successfully clustered clusters, where a successfully clustered cluster refers to the elements within the cluster that are not a subset of either the miRNA-associated disease or gene-associated disease.

Suppose the number of such clusters is smaller than a value n (in this study, we set n=5), the disease similarity between miRNA and gene is calculated:

$$\text{Similarity matrix}_{ij} = \text{Similarity}(\vec{v_i}, \vec{v_j}) \tag{8}$$

$$\text{Similarity}_{\text{average}} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \text{Similarity matrix}_{ij} \tag{9}$$

Parameter Definitions: m and n denote the number of the miRNA-associated disease and gene-associated disease, respectively.

If the number of successfully clustered clusters exceeds n, the similarity between the two lists is calculated:

$$\text{Similarity matrix}_{pq} = \text{Similarity}(\vec{v_p}, \vec{v_q}) \tag{10}$$

$$\text{Similarity}_{\text{average}} = \frac{1}{a \times b} \sum_{p=1}^{a} \sum_{q=1}^{b} \text{Similarity matrix}_{pq} \tag{11}$$

$$\text{Similarity}_{\text{clustered}} = \frac{1}{k} \sum_{k=1}^{k} \text{Similarity}_{\text{average}} \tag{12}$$

Parameter Definitions $\vec{v_p}$ and $\vec{v_q}$ represent the semantic vectors of the p-th and q-th elements of the successfully clustered cluster, respectively. a and b

denote the total number of miRNA-associated diseases and gene-associated diseases in clusters, respectively. k denotes the number of successfully clustered clusters.

Through these steps, we are able to evaluate the disease semantic similarity between the miRNA and gene.

2.3.3 Selection of SBERT Model and Cluster Parameter for the Best Disease Similarity Calculation Performance

The accuracy of classification is influenced by various models, clustering parameters, and the selection of classification methods. Therefore, we conducted parameter testing to identify the optimal model. The test set includes 10 genes that are strongly associated with cancer and have sufficient disease data in DisGeNET, namely TP53, PTEN, KRAS, BCL2, FOXO3, KIF2C, PLK1, CENPA, STAT3 and MYC [24–33]. We hypothesize that a higher degree of similarity indicates a greater likelihood of functional MTIs between miRNA and gene in specific disease. Hence, we employed two linear classification methods, SVM and Logistic Regression, to classify the Functional MTI class and miRWalk class in the test set. We assessed the accuracy of models fine-tuned with different combinations of learning rates and epochs to determine the optimal selection. Additionally, we explored various clustering thresholds and linear classification methods for each fine-tuned model to identify the threshold and classification method that yields the highest performance.

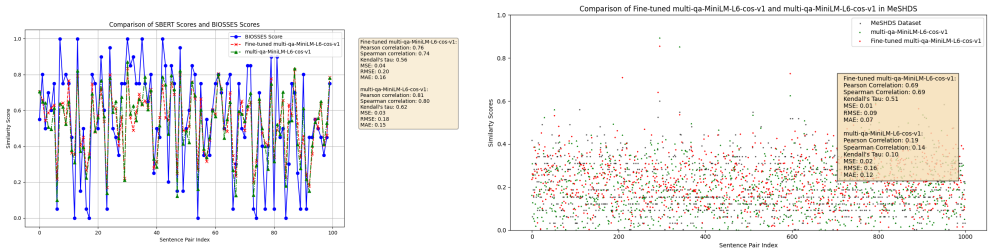
3 Results

3.1 The Performance of Fine-Tuned SBERT Model for Biomedical Semantic String Analysis

The "MeSHDS" dataset comprises 2,249,975 disease pairs along with their corresponding similarities, a substantial volume of data that ensures the high quality of the fine-tuning model. To test the performance of the fine-tuned SBERT model, we calculated Pearson's correlation coefficient and Spearman's correlation coefficient between the model output scores and the manually evaluated scores from BIOSSES. The Pearson correlation coefficient reached 0.76 and the Spearman correlation coefficient reached 0.74 (Figure 3a). We randomly selected 1000 sets of data from the MeSH dataset and calculated Pearson's correlation coefficient and Spearman's correlation coefficient between the model output scores and scores from MeSHDS (Figure 3b). The Pearson Correlation Coefficient and Spearman Correlation Coefficient reached 0.69. The outcome demonstrates its capacity to preserve the initial pattern while effectively acquiring new knowledge through fine-tuning.

3.2 Model and Cluster Parameter Selection

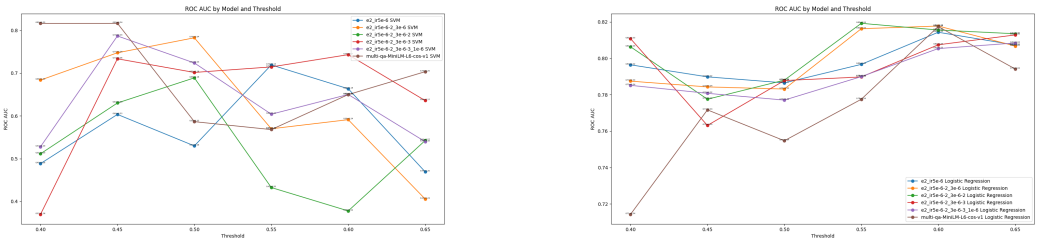
To achieve the optimal model and parameters, we assessed the classification outcomes through different models and parameter settings. Based on the results of parameter tuning, we discovered that the model achieved optimal performance (ROC AUC=0.82) by undergoing 2 epochs of training on the MeSHDS dataset with a learning rate of "5e-6", followed by an additional 2 epochs with a learning rate of "3e-6" (Figure 4). Subsequently, we selected this particular model and set the cluster threshold to 0.55 before applying logistic regression for further investigation.



(a) Model Performance in BIOSSES Dataset

(b) Model Performance in MeSHDS Dataset

Figure 3: Comparison of Model Performance



(a) Support Vector Machine Performance

(b) Logistic Regression Performance

Figure 4: ROC Curves of Different Models and Parameters

3.3 Calculation of Disease Similarity Between miRNAs and Genes throughout Collected MTIs

To obtain a substantial number of disease similarities between MTIs to build a logistic regression classifier. We analyzed 377,312 and 156,555,574 sets of MTIs from miRTarbase and miRWalk, respectively, and analyzed 1,900 miRNAs annotated with diseases from HMDD and 7,727 genes annotated with diseases from DisGeNET with at least 40 associated diseases, resulting in a total of 3,900,041 sets of MTIs. The statistical analysis reveals the distribution of scores as follows (Figure 5, Table 1).

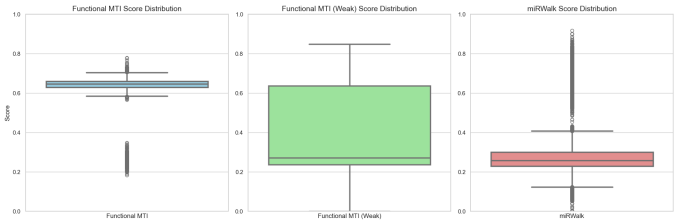


Figure 5: Scores in miRDS Distribution

Table 1: Score Summary

Total Count	3900041
Mean	0.327728
std	0.167878
min	0.004171
25%	0.230130
50%	0.258612
75%	0.303239
max	0.917340

The 'Functional MTIs' from miRTarbase and 'miRWalk' were binary classified based on the score using logistic regression. Since the 'Functional MTI' and the 'miRWalk' classes are unbalanced, random sampling was employed for the 'miRWalk' data. The model exhibited exceptional classification performance, achieving an ROC AUC of 0.97 during a 20-fold cross-validation (Figure 6), thereby establishing its superiority as a classifier. The decision boundary of the output was determined to be 0.5251100843352173, indicating that MTIs with scores above this threshold have

a high probability of being Functional MTIs in specific diseases. The number of scores that exceed the boundary is 755,705 out of a total of 3,900,041.

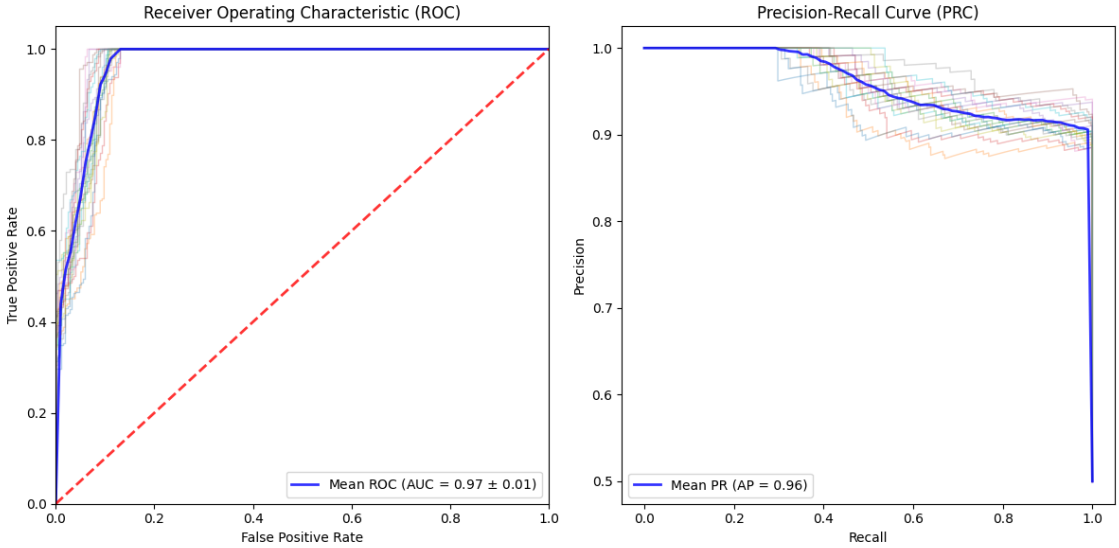


Figure 6: ROC and Precision-Recall Curves of Logistic Regression Classifier

The true accuracy of the model was also assessed simultaneously by selecting MTIs from the intersection of miRTarBase and miRWalk data, representing predicted MTIs that have been experimentally verified. The analysis included a total of 226 MTIs. After excluding rows with insufficient miRNA-associated disease (less than 20), a remaining count of 155 Functional MTIs was obtained, and 142 sets achieved scores above the threshold value of 0.5251100843352173 (Figure 7), demonstrating a recall rate of 91% (Table 2). This indicates that the model has exhibited commendable performance.

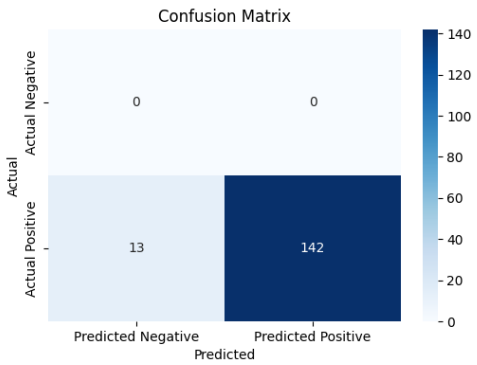


Figure 7: Confusion Matrix of Model Predictions on Intersection MTIs of miRTarbase and miRWalk

Table 2: Model Evaluation Metrics

Metric	Formula	Value
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	0.9161
Precision	$\frac{TP}{TP+FP}$	1.0000
Recall	$\frac{TP}{TP+FN}$	0.9161
F1 Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$	0.9562

TP: True Positives (Correctly Predicted Positives)

TN: True Negatives (Correctly Predicted Negatives)

FP: False Positives (Incorrectly Predicted Positives)

FN: False Negatives (Incorrectly Predicted Negatives)

3.4 Model Generalizability Across Different Databases

The method's generalizability was also assessed by testing its performance on different databases. We manually downloaded the gene-associated disease data from the KEGG database, and within the 10 previously mentioned genes associated with

cancer, only TP53, PTEN, KRAS, and MYC have enough associated diseases in KEGG. These genes were selected for generating 2994 MTIs disease similarity data with both the DisGeNET database and KEGG database respectively. The scores obtained from these two databases were calculated, resulting in a correlation coefficient of approximately 0.8 (0.796944711694144) (Figure 8). This result serves as evidence that the method exhibits excellent performance across different databases.

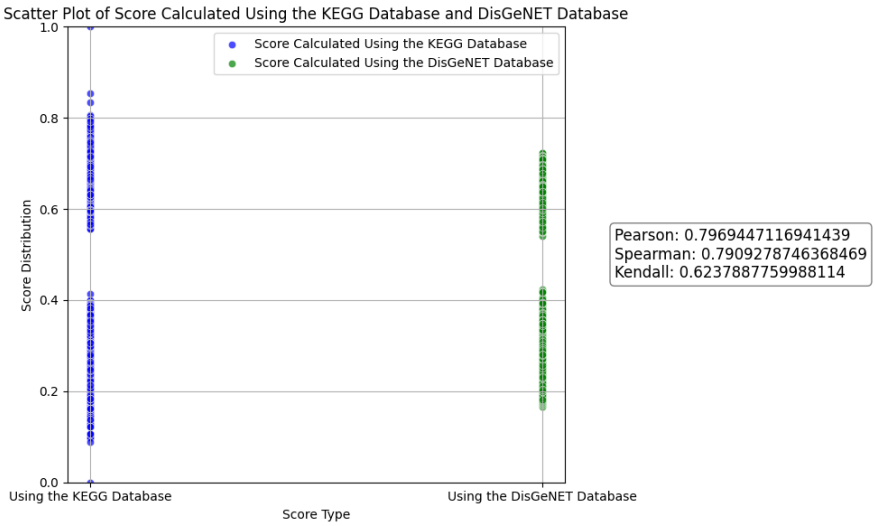
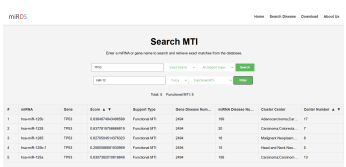


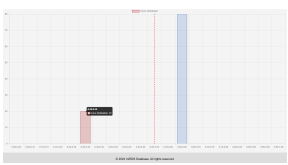
Figure 8: Scatter Plot of Score Calculated Using the KEGG Database and DisGeNET Database

3.5 miRNA Disease Similarity (miRDS) Website Construction

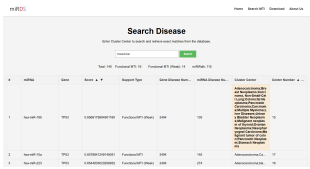
The decision to establish a database for storing the results of calculations was driven by the exceptional accuracy in classifying 'Functional MTIs' and sequence-based predicted MTIs (miRWalk). We have developed the miRDS website (Figure 9), which enables users to search for MTIs and obtain corresponding disease semantic similarity scores. Additionally, specific diseases can be searched to identify functional MTIs associated with them. The website also offers the flexibility to sort the output in ascending or descending order, along with a comprehensive histogram displaying the score distribution of the search results at the bottom of each page. The MTIs located to the right of the dotted line are considered potential Functional MTIs. We recommend selecting MTIs with a minimum of 20 miRNA-associated diseases to enhance the credibility of outcome predictions.



(a) Search MTIs



(b) Score Histogram of MTIs Search



(c) Search Functional MTIs in Diseases

Figure 9: miRDS Database Website

4 Discussion

The present study proposes a methodology to further refine the selection of MTIs from sequence-based prediction tools by incorporating disease similarity analysis between miRNAs and genes, ultimately leading to the establishment of a comprehensive database for disease-specific MTIs. The distribution of the three different types of scores shows that the scores for functional MTIs are overall greater than functional MTI(weak), and functional MTI(weak) is greater than miRWalk. These findings support our hypothesis that higher scores indicate a greater likelihood of functional MTIs. The reason behind the observation may attributed to the fact that a majority of experimentally validated MTIs possess specific biological implications, such as playing a significant role in disease pathogenesis. The method demonstrates a strong correlation between MTIs and diseases, as the computation of disease semantic similarity between miRNA and Gene can greatly predict the type of MTIs.

Moreover, considering the abundance of miRNA-associated disease and gene-associated disease data, our proposed method holds great potential for widespread application. Notably, the DisGeNET database encompasses 21,671 genes linked to various disease types, while HMDD provides information on 1,900 microRNAs and their associated diseases. The estimated total number of human genes is approximately 22,333 [34], while the estimated number of mature miRNAs in humans is around 2,300 [35]. Based on estimation, the method has the capability to encompass a substantial proportion of both miRNAs and genes. More importantly, compared to previous disease similarity algorithms, such as the MISIM and SemFunSim [36, 37], the sentence-bert-based disease similarity computation method does not require the input that the disease must be present on the MeSH Tree, which greatly improves the scope of the computation.

However, there are still limitations. Despite the extensive coverage of miRNAs and genes in the database, the number of diseases associated with a single miRNA and gene remains limited. In the case of DisGeNet, when I screened genes with over 40 diseases, the gene count decreased from approximately 20,000 to around 8000. The lack of sufficient disease data can undermine the accuracy of classification and weaken the persuasiveness of the results. Additionally, there is a computational challenge in the method as only the top 40 gene-associated diseases are considered for analysis. As the number of gene-associated diseases increases, it significantly prolongs the running time of the model, posing difficulties in handling large-scale data calculation.

5 Perspective

The results of the experiment and the model's performance on the BIOSSES dataset demonstrate that the Natural Language Processing (NLP) method excels in the biomedical domain, providing compelling evidence for its high accuracy in calculating semantic similarity within biomedical vocabulary after specific fine-tuning. The method demonstrates promising scalability and can also be applied to compute the disease semantic similarity or functional semantic similarity between other biomolecules, thereby facilitating the exploration of their interactions. In addition to predicting interactions based on physical and chemical properties, the method can offer a novel prediction insight into disease semantic similarity or functional semantic similarity for optimizing prediction results.

Regarding application prospects, we propose a scenario where performing miRNA

profiling on diseased tissue samples from clinical patients allows obtaining a substantial amount of miRNA-associated disease data. A similar approach can also be employed to acquire gene expression data in order to enhance the identification of gene-associated diseases. By utilizing this approach, it may become feasible to construct an exceptionally accurate prediction network that is expected to provide robust support for future diagnosis and treatment [38].

Declarations

Equipment

The computations were performed using an NVIDIA A100 GPU with 80GB of memory.

Funding

This research was partially funded by the Shenzhen X-institute ESRT Scholarship.

Availability of data and materials

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. The MeSHDS dataset can be gained through 'Baiming123/MeSHDS' in HuggingFace. The miRDS database will be available soon.

Patent

A patent application related to the work presented in this paper has been filed Patent Application No.202410559682.3 Title: 基于文本的MicroRNA和基因关系挖掘方法、装置、设备及介质

References

[1] Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *cell*, 75(5):843–854, 1993.

[2] Marieke Van Kouwenhove, Martijn Kedde, and Reuven Agami. Microrna regulation by rna-binding proteins and its implications for cancer. *Nature Reviews Cancer*, 11(9):644–656, 2011.

[3] Michael Carleton, Michele A Cleary, and Peter S Linsley. Micrornas and cell cycle regulation. *Cell cycle*, 6(17):2127–2132, 2007.

[4] Angie M Cheng, Mike W Byrom, Jeffrey Shelton, and Lance P Ford. Antisense inhibition of human mirnas and indications for an involvement of mirna in cell growth and apoptosis. *Nucleic acids research*, 33(4):1290–1297, 2005.

[5] Meredith Hill and Nham Tran. mirna interplay: mechanisms and consequences in cancer. *Disease models & mechanisms*, 14(4):dmm047662, 2021.

[6] Colin C Pritchard, Heather H Cheng, and Muneesh Tewari. Microrna profiling: approaches and considerations. *Nature Reviews Genetics*, 13(5):358–369, 2012.

[7] Yuhao Chen and Xiaowei Wang. mirdb: an online database for prediction of functional microrna targets. *Nucleic acids research*, 48(D1):D127–D131, 2020.

[8] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11:1–14, 2010.

[9] Byunghan Lee, Junghwan Baek, Seunghyun Park, and Sungroh Yoon. deeptarget: end-to-end learning framework for microrna target prediction using deep recurrent neural networks. In *Proceedings of the 7th ACM international conference on bioinformatics, computational biology, and health informatics*, pages 434–442, 2016.

[10] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005, 2015.

[11] Kevin C Miranda, Tien Huynh, Yvonne Tay, Yen-Sin Ang, Wai-Leong Tam, Andrew M Thomson, Bing Lim, and Isidore Rigoutsos. A pattern-based method for the identification of microrna binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–1217, 2006.

[12] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microrna target recognition. *Nature genetics*, 39(10):1278–1284, 2007.

[13] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, 42(D1):D92–D97, 2014.

[14] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[17] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 2017.

[18] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Jing Li, Kai-Yao Huang, Sirjana Shrestha, Hsiao-Chin Hong, Yun Tang, Yi-Gang Chen, Chen-Nan Jin, Yuan Yu, et al. mirtarbase 2020: updates to the experimentally validated microrna–target interaction database. *Nucleic acids research*, 48(D1):D148–D154, 2020.

[19] Carsten Sticht, Carolina De La Torre, Alisha Parveen, and Norbert Gretz. mirwalk: an online resource for prediction of microrna binding sites. *PloS one*, 13(10):e0206239, 2018.

[20] Zhou Huang, Jiangcheng Shi, Yuanxu Gao, Chunmei Cui, Shan Zhang, Jianwei Li, Yuan Zhou, and Qinghua Cui. Hmdd v3. 0: a database for experimentally supported human microrna–disease associations. *Nucleic acids research*, 47(D1):D1013–D1017, 2019.

[21] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943, 2016.

[22] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[23] sentence-transformers/multi-qa-minilm-l6-cos-v1 · hugging face.

[24] A Petitjean, MIW Achatz, AL Borresen-Dale, P Hainaut, and M Olivier. Tp53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*, 26(15):2157–2165, 2007.

[25] Nader Chalhoub and Suzanne J Baker. Pten and the pi3-kinase pathway in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 4:127–150, 2009.

[26] Hua Yu, Drew Pardoll, and Richard Jove. Stats in cancer inflammation and immunity: a leading role for stat3. *Nature reviews cancer*, 9(11):798–809, 2009.

[27] Niki Karachaliou, Clara Mayo, Carlota Costa, Ignacio Magrí, Ana Gimenez-Capitan, Miguel Angel Molina-Vila, and Rafael Rosell. Kras mutations in lung cancer. *Clinical lung cancer*, 14(3):205–214, 2013.

[28] Anna Frenzel, Francesca Grespi, Waldemar Chmelewskij, and Andreas Villunger. Bcl2 family proteins in carcinogenesis and the treatment of cancer. *Apoptosis*, 14:584–596, 2009.

[29] Shang Yao, Lavender Yuen-Nam Fan, and Eric Wing-Fai Lam. The foxo3-foxm1 axis: A key cancer drug target and a modulator of cancer drug resistance. In *Seminars in cancer biology*, volume 50, pages 77–89. Elsevier, 2018.

[30] Shanshan Liu, Ziwei Ye, Vivian Weiwen Xue, Qi Sun, Huan Li, and Desheng Lu. Kif2c is a prognostic biomarker associated with immune cell infiltration in breast cancer. *BMC cancer*, 23(1):307, 2023.

[31] Zhixian Liu, Qingrong Sun, and Xiaosheng Wang. Plk1, a potential target for cancer therapy. *Translational oncology*, 10(1):22–32, 2017.

[32] Jing Han, Rongkai Xie, Ying Yang, Diangang Chen, Li Liu, Jiayang Wu, and Sufen Li. Cenpa is one of the potential key genes associated with the proliferation and prognosis of ovarian cancer based on integrated bioinformatics analysis and regulated by mybl2. *Translational cancer research*, 10(9):4076, 2021.

[33] Chi V Dang. Myc on the path to cancer. *Cell*, 149(1):22–35, 2012.

[34] Mihaela Pertea and Steven L Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome biology*, 11:1–7, 2010.

[35] Julia Alles, Tobias Fehlmann, Ulrike Fischer, Christina Backes, Valentina Galata, Marie Minet, Martin Hart, Masood Abu-Halima, Friedrich A Grässer, Hans-Peter Lenhof, et al. An estimate of the total number of true human mirnas. *Nucleic acids research*, 47(7):3353–3364, 2019.

[36] Dong Wang, Juan Wang, Ming Lu, Fei Song, and Qinghua Cui. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics*, 26(13):1644–1650, 2010.

[37] Liang Cheng, Jie Li, Peng Ju, Jiajie Peng, and Yadong Wang. Semfunsim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PloS one*, 9(6):e99415, 2014.

[38] Rajesha Rupaimoole and Frank J Slack. Microrna therapeutics: towards a new era for the management of cancer and other diseases. *Nature reviews Drug discovery*, 16(3):203–222, 2017.