Refining Protein-Level MicroRNA Target Interactions in Disease from Prediction Databases Using Sentence-BERT

Baiming Chen*[1,2]

[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

September 18, 2024

Abstract

miRNAs (MicroRNAs) regulate gene expression by binding to 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, which analyze sequence complementarity, thermodynamic stability, and evolutionary conservation to predict miRNA-target interactions (MTIs) within the 3' untranslated region (3'UTR). We propose a concept for further screening sequence-based predicted MTIs by considering the disease similarity between miRNA and gene to establish a prediction database of disease-specific MTIs. We fine-tuned a Sentence-BERT model to calculate disease semantic similarity. The method achieved an F1 score of 88% in accurately distinguishing protein-level experimentally validated MTIs and predicted MTIs. Moreover, the method exhibits exceptional generalizability across different databases. The proposed method was utilized to calculate the similarity of disease in 1,220,904 MTIs from miRTarbase, miRDB, and miRWalk, involving 6,085 genes and 1,261 pre-miRNAs. The study holds the potential to offer valuable insights into comprehending miRNA-gene regulatory networks and advancing progress in disease diagnosis, treatment, and drug development.

Keywords: MicroRNA (miRNA), Sentence-BERT, miRNA-Target Interactions (MTIs), Disease Semantic Similarity, Gene Regulation, Database

1 Introduction

In 1993, miRNA (Lin-4) was first discovered 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 bioinformatics technologies [2]. miRNAs are classified as small non-coding RNAs (ncRNAs),

*First and corresponding author: baimingchen@link.cuhk.edu.cn

and 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 MTIs in diseases holds biological significance.

There are many existing MTI prediction databases. miRWalk identifies consecutive complementary subsequences between miRNA and gene to construct a database for MTIs prediction in different species [7]. miRDB considered miRNA overexpression and CLIP-seq data, utilizing a support vector machine (SVM) for MTI prediction in different species [8]. mirSVR trained a regression model to rank miRNA target sites based on down-regulation score, using data from miRanda [9]. DeepTarget uses Recurrent Neural Networks (RNN) for learning sequence-sequence interaction patterns to predict MTIs [10]. TargetScan predicts MTIs based on seed match and conservativeness analysis in different species [11]. The RNA22 is designed for the identification of miRNA binding sites and their corresponding miRNA and mRNA complexes [12]. PITA predicts miRNA targets by considering both the binding free energy and the accessibility of the target site within the mRNA secondary structure [13]. StarBase explores miRNA-mRNA and miRNA-lncRNA interaction networks based on CLIP-Seq data [14]. However, there are few disease-specific MTI databases.

The NLP method can be applied to address this issue. Sentence-BERT (SBERT) improves the pre-trained BERT model by incorporating Siamese and triplet network structures [15]. BERT, a bidirectional Transformer model, captures the semantic meaning of each word based on its preceding and following context. Traditional BERT requires sentence pairs to be concatenated and processed together, making tasks like sentence similarity calculations computationally expensive. SBERT, by contrast, generates independent sentence embeddings that can be precomputed, allowing for sentence comparisons through simple similarity calculations such as cosine similarity. This approach significantly enhances inference speed. SBERT outperforms both BERT [16] and RoBERTa [17] in terms of speed while maintaining comparable accuracy.

The present study proposes a method for calculating the similarities of diseases between miRNAs and genes, aiming to further refine sequence-based MTI prediction to establish a database of disease-specific MTIs. We assume that protein-level experimentally verified MTIs (Functional MTIs) exhibit different disease similarity patterns from predicted MTIs. The experiment result confirms this hypothesis. The similarity between two diseases can be assessed using the SBERT model. We generate a similarity matrix by computing pairwise disease similarities between miRNAs and genes. The matrix is analyzed to determine the statistical distribution of its values, which is then used as a feature in the machine learning method for classifying MTIs.

2 Materials and Methods

2.1 Overview

We present a flow chart (Figure 1) that outlines the major steps of our study. These steps include fine-tuning the SBERT model, calculating the similarity matrix, and classifying MTIs based on the statistical distribution of values within the similarity matrix.

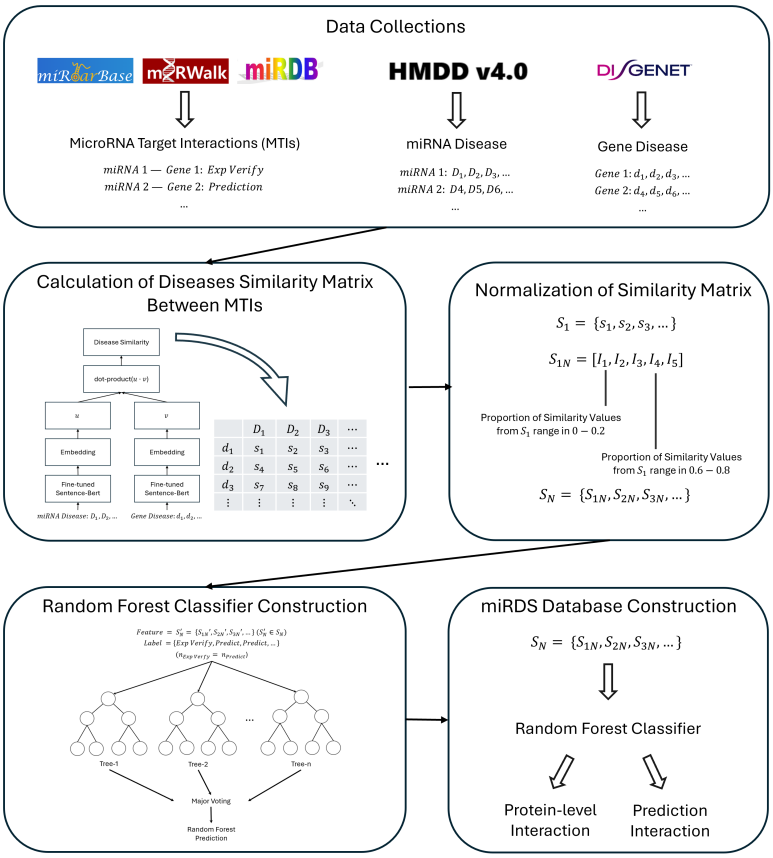


Figure 1: Flow Chart of Method

2.2 Data Collection

To fine-tune the SBERT model, we downloaded the MeSH (Medical Subject Headings) descriptors (https://meshb.nlm.nih.gov/) from the National Library of Medicine. The MeSH descriptors are standardized terms utilized for indexing and retrieving biomedical and health-related information, forming a hierarchical vocabulary tree (Figure 3). To test the SBERT model performance, we obtained the BIOSSES dataset (Figure 2), which consists of 100 pairs of biomedical statements with manually annotated text similarity scores provided by experts [18]. It can be used to evaluate the biomedicine text similarity calculation model. To obtain protein-level experimentally verified MTIs and predicted MTIs, we utilized the miRTarBase [19], miRWalk [7] and miRDB [8]. miRTarBase provides experimentally validated MTIs, which are primarily categorized into two types: Functional MTIs and Functional MTIs (Weak). Functional MTIs are validated through Reporter assays, Western blot, and other sensitive and specific techniques. Functional MTIs (Weak) are identified using Microarray, qPCR, or sequencing. miRWalk and miRDB provide sequenced-based predicted MTIs. For acquiring information on miRNA-associated and gene-associated diseases, we employed HMDD [20], DisGeNEt [21] and KEGG [22]. HMDD provides data on diseases associated with pre-miRNAs. Therefore, for disease associations involving mature miRNAs such as hsa-miR-122b-3p and hsa-miR-122b-5p, we will refer to their corresponding pre-miRNA, hsa-mir-122b, for subsequent calculations. DisGeNet and KEGG provide diseases associated with genes.

Sentence 1	Sentence 2	Similarity (0-4)
It has recently been shown that Craf is essential for Kras G12D-induced NSCLC.	It has recently become evident that Craf is essential for the onset of Kras-driven non-small cell lung cancer.	4
The Bcl-2 inhibitor ABT-737 induces regression of solid tumors and its derivatives are in the early clinical phase as cancer therapeutics; however, it targets Bcl-2, Bcl-XL, and Bcl-w, but not Mcl-1, which induces resistance against apoptotic cell death triggered by ABT-737.	Recently, it has been reported that ABT-737 is not cytotoxic to all tumors cells, and that chemoresistance to ABT-737 is dependent on appreciable levels of Mcl-1 expression, the one Bcl-2 family member it does not effectively inhibit.	3
...		

Figure 2: Example of BIOSSES Dataset

2.3 Model Construction

2.3.1 Selection of Pretrained Model

The ideal model we seek must fulfill two primary criteria: exceptional performance in semantic similarity calculation tasks and rapid execution speed to facilitate large-scale data processing (Table 1). Moreover, scalability in handling lengthy biomedical semantic strings is favorable. Based on the SBERT website (https://www.sbert.net/), we have selected the 'multi-qa-MiniLM-L6-cos-v1' [23] model for further investigation. This model excels in execution speed, and robust semantic similarity calculations, and supports extended input sequences, with a maximum sequence length of 512 compared to 256 for the 'all-MiniLM-L6-v2' model.

Rank[a]	Model Name	Emb. Perf.[b]	Sem. Search Perf.[c]	Enc. Spd.[d]
1	all-mpnet-base-v2	69.57	57.02	2800
3	all-distilroberta-v1	68.73	50.94	4000
4	all-MiniLM-L12-v2	68.70	50.82	7500
6	all-MiniLM-L6-v2	68.06	49.54	14200
7	multi-qa-MiniLM-L6-cos-v1	64.33	51.83	14200

[a] The ranking of the models is based on their average performance in sentence embedding and semantic search performance.
[b] The average performance of the model in embedding sentences in different domains. Higher values indicate better performance.
[c] The average performance of the model in semantic search tasks. Higher values indicate better performance.
[d] The encoding speed of the models measured in encoding sentences per second on a V100 GPU.

Table 1: Performance Comparison of Pretrained SBERT Models

2.3.2 Fine-Tuning Pretrained Models to Enhance Disease Similarity Measurement

For the model to better handle the disease semantic similarity computation task while maintaining its original performance, we utilized MeSH descriptors from category C (diseases) to create a dataset containing pairs of diseases and their disease similarity. The distribution of disease similarities in the dataset we created closely mirrors the disease semantic similarity distribution produced by the 'multi-qa-MiniLM-L6-cos-v1' model without fine-tuning. Both types of distributions follow a log-normal distribution pattern. The 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 3, Figure 4). The weighted disease similarity was further adjusted to better align with the distribution of the output generated by the 'multi-qa-MiniLM-L6-cos-v1' model, resulting in a dataset, named 'MeSHDS'. We utilized this dataset to fine-tune the SBERT model.

2.3.3 Similarity Analysis Between miRNA and Gene

To analyze the similarity between diseases associated with miRNAs and genes, we compute the similarity matrix using two disease string lists: miRNA-associated disease string list and gene-associated disease string list, then analyze the value distribution within the similarity matrix.

Semantic String Embedding and Similarity Calculation First, each string in two lists is embedded using the Sentence-BERT model. The similarity between two strings is obtained by calculating the dot product of their embedding vectors:

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

Parameter Definitions: $\vec{v_i}$ represents the semantic vector of elements i in the miRNA-associated disease list embedded with sentence-bert. $\vec{v_j}$ represents the semantic vector of elements j in the gene-associated disease list embedded with sentence-bert.

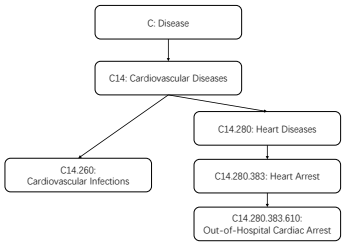


Figure 3: Disease Descriptor 'Cardiovascular Infections' and 'Out-of-Hospital Cardiac Arrest' Hierarchical Structure in MeSH Tree

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

Figure 4: Adjusted Similarity Calculation between 'Cardiovascular Infections' and 'Out-of-Hospital Cardiac Arrest'

Similarity Matrix Calculation The similarity matrix between miRNA-associated disease and gene-associated disease can be assessed using the following approach:

$$\text{Similarity Matrix}(i, j) = \text{Similarity}(\vec{v_j}, \vec{v_i}) \tag{7}$$

Parameter Definitions: Similarity Matrix(i, j) represents the element at the i-th row and j-th column of the similarity matrix.

Analyze the Value Distribution in the Similarity Matrix The similarity matrices of varying dimensions are transformed into a normalized disease similarity list of length 5 using statistical method, wherein each position represents the distribution of similarities within a specific interval. For example, the first position corresponds to values ranging from 0 to 0.2. This normalized list serves as a classification feature. Meanwhile, by extracting values above 0.8 in the similarity matrix, detailed information on miRNA-associated and gene-associated diseases can be obtained.

2.3.4 Identifying Optimal Fine-tuned Models and Classification Methods for Superior Classification Performance

Functional MTIs and MTIs from miRWalk and miRDB were classified based on the normalized disease similarity list, with the classifier's performance evaluated using the F1 score. The F1 score, calculated as the harmonic mean of precision and recall, was affected by different fine-tuned models and the choice of classification methods in this study. To ensure optimal performance, we conducted extensive tuning to identify the most suitable model and classification approach. The test set includes 20 genes that are associated with cancer and have sufficient disease data in DisGeNET, namely TP53, PTEN, KRAS, BCL2, FOXO3, KIF2C, PLK1, CENPA, STAT3, MYC, TNF, IL6, TGFB1, IL1B, IL10, IL17A, MMP9, CRP, TLR4, NLRP3 [24–43]. We hypothesize that protein-level experimentally verified MTIs (Functional MTIs) exhibit different disease similarity patterns from predicted MTIs, thus enabling their classification using machine learning techniques. Therefore, we employed five machine learning algorithms, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting Machine, to classify an equal number of Functional MTIs and MTIs from miRWalk & miRDB.

3 Results

3.1 The Performance of Fine-Tuned SBERT Model for Disease Similarity Calculation

The "MeSHDS" dataset comprises 2,249,975 disease pairs along with their corresponding similarities (Figure 5), a substantial volume of data that ensures the high quality of fine-tuning. To test the performance of the fine-tuned SBERT model, we calculated Pearson's correlation coefficient between the model output scores and scores from BIOSSES and MeSHDS (Figure 6). The outcome demonstrates its capacity to compute biomedical semantic similarity while learning to compute disease similarity through fine-tuning.

Disease 1	Disease 2	Similarity (0-1)
Acute Abdomen	Morton Neuroma	0.446130547
Acute Abdomen	Cellulite	0.310198272
...		

Figure 5: Example of MeSHDS Dataset

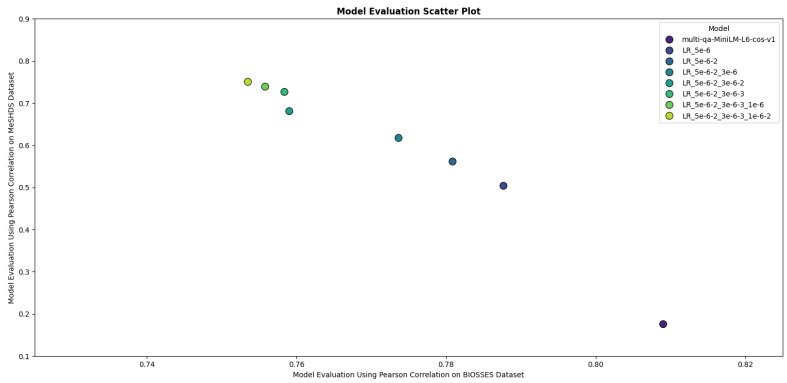


Figure 6: Model Performance Benchmark. The figure compares the performance of different fine-tuning strategies, as described below:

1. **LR_5e_6**: Fine-tuned for 1 epoch using a learning rate of 5e-6.
2. **LR_5e_6-2**: Fine-tuned for 2 epochs using a learning rate of 5e-6.
3. **LR_5e_6-2_3e_6**: Fine-tuned for 2 epochs with a learning rate of 5e-6, followed by 1 epoch with 3e-6.
4. **LR_5e_6-2_3e_6-2**: Fine-tuned for 2 epochs with a learning rate of 5e-6, followed by 2 epochs with 3e-6.
5. **LR_5e_6-2_3e_6-3**: Fine-tuned for 2 epochs with a learning rate of 5e-6, followed by 3 epochs with 3e-6.
6. **LR_5e_6-2_3e_6-3_1e_6**: Fine-tuned for 2 epochs with a learning rate of 5e-6, 3 epochs with 3e-6, and 1 epoch with 1e-6.
7. **LR_5e_6-2_3e_6-3_1e_6-2**: Fine-tuned for 2 epochs with a learning rate of 5e-6, 3 epochs with 3e-6, and 2 epochs with 1e-6.

3.2 Fine-Tuning Model and Classification Method Selection

The F1 score was employed as the primary selection criterion to identify the optimal SBERT model and machine learning method for classifying protein-level experimentally validated MTIs and predicted MTIs. The pre-trained model fine-tuned using MeSHDS for two rounds, with a learning rate of '5e-6', followed by an additional two rounds of fine-tuning at a learning rate of '3e-6' demonstrated remarkable performance improvement, achieving an 8% increase compared to the original model. The utilization of random forest and gradient descent classification methods in this model leads to optimal overall performance, as evidenced by an F1 Score exceeding 0.8 (Figure 7). After consideration, we have decided to select the 'LR_5e-6_2_3e-6' model and employ the random forest for further investigation.

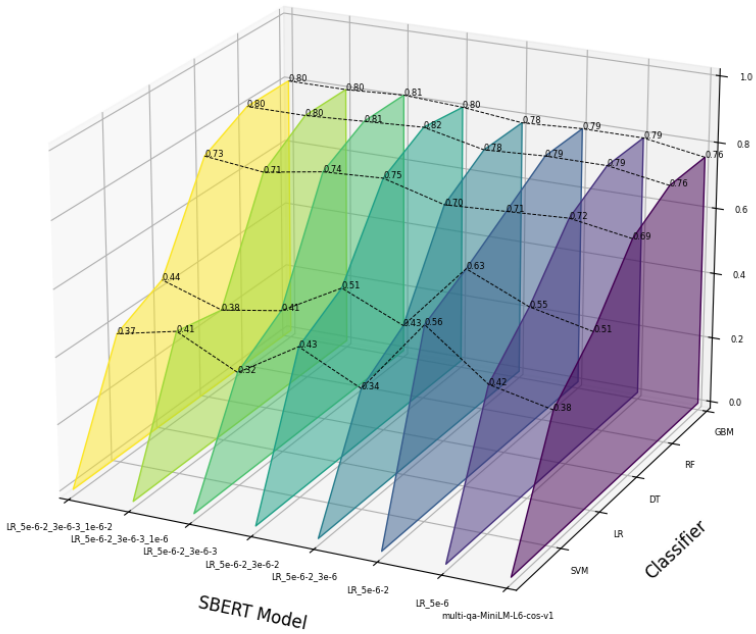


Figure 7: F1 Scores Across Different SBERT Models and Classification Methods.

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

The disease similarity matrix of 1,220,904 MTIs was computed, including 1,261 miRNAs from HMDD and 6,085 genes from DisGeNET. The following word clouds highlight the most frequently occurring genes (Figure 8) and pre-miRNAs (Figure 9).

The protein-level experimentally verified MTIs (Functional MTIs) from miRTarBase and predicted MTIs from miRWalk and miRDB, were binary classified using a Random Forest algorithm, with the similarity matrix employed as the feature. To address the imbalance between Functional MTIs and predicted MTIs, random sampling was employed for predicted MTIs. The classifier achieved an F1 score of 0.88 (Figure 12), with a ROC AUC of 0.94 (Figure 10) and a PRC AUC of 0.93 (Figure 11), demonstrating strong classification performance. Subsequently, the trained

SLC8A1
RPS6KA5
DYRK1A
TPM3
PPM1L
NTSK2
POLK
GREM1
PCYT1B
SDR42E1
IL31RA
TTBK1
AP1M1
ZNF81
NDUFA10
PCDH9
EPB41L3
HPGD
NE2
FGF1
BUB3
LPR
MTAP
KCNMA1
VSTM4
SELENOI
RBM28
KCNN3
HNF4A
MDM2
IL6ST
TK2
PPARA
SMAD2
FXN
MCPH1
FOOK3
KIF1B
NSD2
PSEN1
ABLIM1
THY1
NFIB
PDE1C
NUPR1
DNAL1
ONECUT2
GPSCLA
CCL28
CUX1
BRF2
CCDC134
RNF115
DYM
CAMK2A
ARMCX5
ADCYAP1R1
CALM1
RP1-257.1
RP3-434.2
RP11-1407.1
RP11-1399H1.3
RP11-1399H1.1
RP11-1399H1.2
RP11-1399H1.4
RP11-1399H1.5
RP11-1399H1.6
RP11-1399H1.7
RP11-1399H1.8
RP11-1399H1.9
RP11-1399H1.10
RP11-1399H1.11
RP11-1399H1.12
RP11-1399H1.13
RP11-1399H1.14
RP11-1399H1.15
RP11-1399H1.16
RP11-1399H1.17
RP11-1399H1.18
RP11-1399H1.19
RP11-1399H1.20
RP11-1399H1.21
RP11-1399H1.22
RP11-1399H1.23
RP11-1399H1.24
RP11-1399H1.25
RP11-1399H1.26
RP11-1399H1.27
RP11-1399H1.28
RP11-1399H1.29
RP11-1399H1.30
RP11-1399H1.31
RP11-1399H1.32
RP11-1399H1.33
RP11-1399H1.34
RP11-1399H1.35
RP11-1399H1.36
RP11-1399H1.37
RP11-1399H1.38
RP11-1399H1.39
RP11-1399H1.40
RP11-1399H1.41
RP11-1399H1.42
RP11-1399H1.43
RP11-1399H1.44
RP11-1399H1.45
RP11-1399H1.46
RP11-1399H1.47
RP11-1399H1.48
RP11-1399H1.49

across the two databases was calculated. The average Pearson's correlation coefficient reached 0.89 (Figure 14). This result serves as evidence that the method exhibits perfect generalizability across different databases.

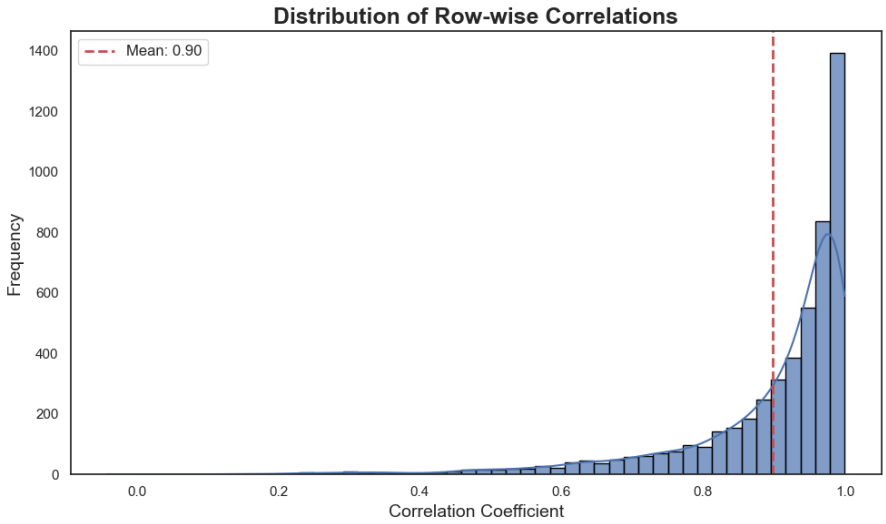


Figure 14: Row-wise Distribution of Normalized Disease Similarity List Correlation Coefficients Between DisGeNET and KEGG Databases

3.5 miRNA Target Disease Similarity (miRTarDS) Website Construction

To visualize the results of our calculations, we developed the miRTarDS website (Figure 15). The website allows users to explore the predicted types of each microRNA-target interaction (MTI) along with highly probable associated diseases. Users can search for specific MTIs (Figure 16) or diseases (Figure 17). The total number of searches and the total number of pages are displayed beneath the search bar, with 20 MTIs shown per page. The source data is provided at the end of the paper, and the website is scheduled to launch soon.

Figure 15: Home Page

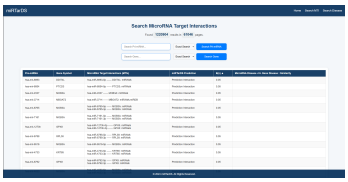


Figure 16: Search MTI

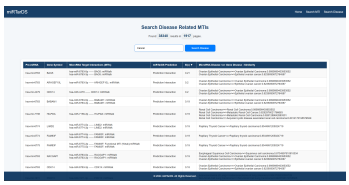


Figure 17: Search disease

Figure 18: miRTarDS Database

4 Discussion

The present study introduces a method to refine sequence-based predicted MTIs by considering disease similarity between miRNAs and genes. This approach ultimately leads to the development of a database that predicts potential protein-level MTIs and identifies diseases associated with these MTIs. Interestingly, we compared

the real data (Figure 19) with the classification data generated by the random forest model (Figure 20) and discovered a remarkable similarity between them, thereby validating the effectiveness of our classification methodology. Furthermore, it was observed that within the 0-0.2 range, Functional MTIs and Protein-level Interaction exhibited smaller variations compared to MTIs from miRWalk & miRDB and Prediction Interaction respectively. The results were reversed in the 0.2-0.4 range, demonstrating that MTIs experimentally verified at the protein level were associated with a higher similarity between miRNA-associated diseases and gene-associated diseases.

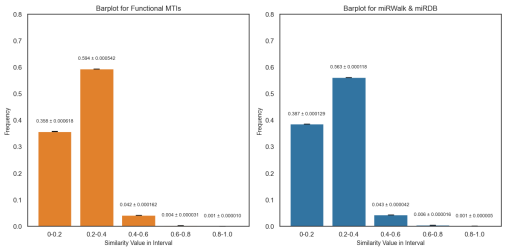


Figure 19: Average Value in Normalized Disease Similarity List of Functional MTIs and MTIs from miRWalk & miRDB (Real Data)

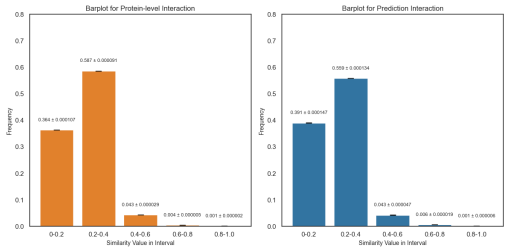


Figure 20: Average Value in Normalized Disease Similarity List of Protein-level Interaction and Prediction Interaction (Random Forest Prediction)

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 miRNAs and their associated diseases. The estimated total number of human genes is approximately 22,333 [44], while the estimated number of mature miRNAs in humans is around 2,300 [45]. Based on estimation, the method can encompass a substantial proportion of both miRNAs and genes. More importantly, compared to previous disease similarity algorithms, such as the MISIM and SemFunSim [46, 47], the SBERT-based disease similarity computation method does not require the input disease must be present on the MeSH Tree, which greatly improves the scope of the computation.

5 Perspective

The results of the experiment and the model's performance on BIOSSES and MeSHDS datasets demonstrate that the Natural Language Processing (NLP) method excels in calculating semantic string similarity in biomedical contexts after specific fine-tuning. The method demonstrates promising scalability and can also be applied to compute the disease similarity or functional similarity between other biomolecules, thereby facilitating the exploration of their interactions. In addition to predicting interactions based on physical and chemical properties, the method provides novel insights into disease similarity and functional similarity, thereby optimizing the analysis.

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 to enhance the identification of gene-associated diseases. By utilizing this approach, it may become feasible to construct a comprehensive MTI network in diseases, providing robust support for future diagnosis and treatment [48].

Declarations

Acknowledgments

Throughout this paper, the pronoun "We" is used instead of "I" to follow academic conventions. Welcome any feedback or suggestions, if any errors are found, please feel free to provide corrections. Thank you for your support. The computing resources of the Chinese University of Hong Kong, Shenzhen have not been utilized for this project, as I hereby affirm. At present, this paper is not being considered by any journal, and I would be delighted to receive an invitation. Without contacting me, please do not integrate the data from this paper into any databases.

Equipment

Intel(R) Xeon(R) Platinum 8358 CPUs @ 2.60GHz and 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 MeSHDS dataset and optimal fine-tuned model can be gained through 'Baiming123/MeSHDS' and 'Baiming123/Calcu_Disease_Similarity' in HuggingFace, respectively. miRTarDS source data can be gained through: 'AltasChen/miRTarDS' in GitHub.

Patent

Two patent applications related to the work have been filed. Patent Application No.202410559682.3 and Application No.202411135644.1. For commercial use, contact: baimingchen@link.cuhk.edu.cn.

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] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[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] Daniela N Mannel and Bernd Echtenacher. Tnf in the inflammatory response. *Chem Immunol*, 74(74):141–161, 2000.

[35] Toshio Tanaka, Masashi Narazaki, and Tadamitsu Kishimoto. Il-6 in inflammation, immunity, and disease. *Cold Spring Harbor perspectives in biology*, 6(10):a016295, 2014.

[36] John J Letterio and Anita B Roberts. Regulation of immune responses by tgf-β. *Annual review of immunology*, 16(1):137–161, 1998.

[37] Yuki Nakamura, Riho Aihara, Hisataka Iwata, Takehito Kuwayama, and Koumei Shirasuna. Il1b triggers inflammatory cytokine production in bovine oviduct epithelial cells and induces neutrophil accumulation via ccl2. *American Journal of Reproductive Immunology*, 85(5):e13365, 2021.

[38] Margarida Saraiva and Anne O'garra. The regulation of il-10 production by immune cells. *Nature reviews immunology*, 10(3):170–181, 2010.

[39] Yoichiro Iwakura, Susumu Nakae, Shinobu Saijo, and Harumichi Ishigame. The roles of il-17a in inflammatory immune responses and host defense against pathogens. *Immunological reviews*, 226(1):57–79, 2008.

[40] Xiaodong Wang, Yan Yiu Yu, Shirley Lieu, Frank Yang, Jeffrey Lang, Chuanyong Lu, Zena Werb, Diane Hu, Theodore Miclau, Ralph Marcucio, et al. Mmp9 regulates the cellular response to inflammation after skeletal injury. *Bone*, 52(1):111–119, 2013.

[41] Marco Del Giudice and Steven W Gangestad. Rethinking il-6 and crp: Why they are more than inflammatory biomarkers, and why it matters. *Brain, behavior, and immunity*, 70:61–75, 2018.

[42] Suprabhat Mukherjee, Subhajit Karmakar, and Santi Prasad Sinha Babu. Tlr2 and tlr4 mediated host immune responses in major infectious diseases: a review. *Brazilian Journal of Infectious Diseases*, 20:193–204, 2016.

[43] Fayyaz S Sutterwala, Stefanie Haasken, and Suzanne L Cassel. Mechanism of nlrp3 inflammasome activation. *Annals of the New York Academy of Sciences*, 1319(1):82–95, 2014.

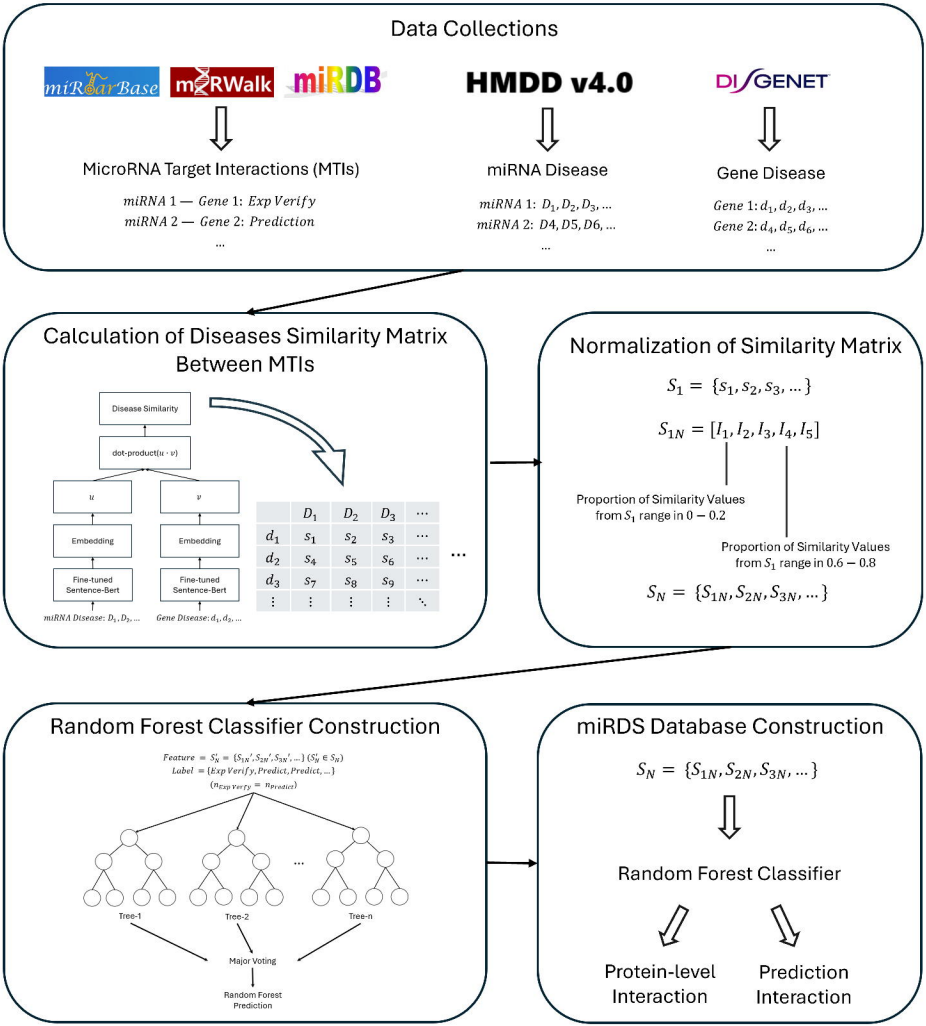
[44] Mihaela Pertea and Steven L Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome biology*, 11:1–7, 2010.

[45] 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.

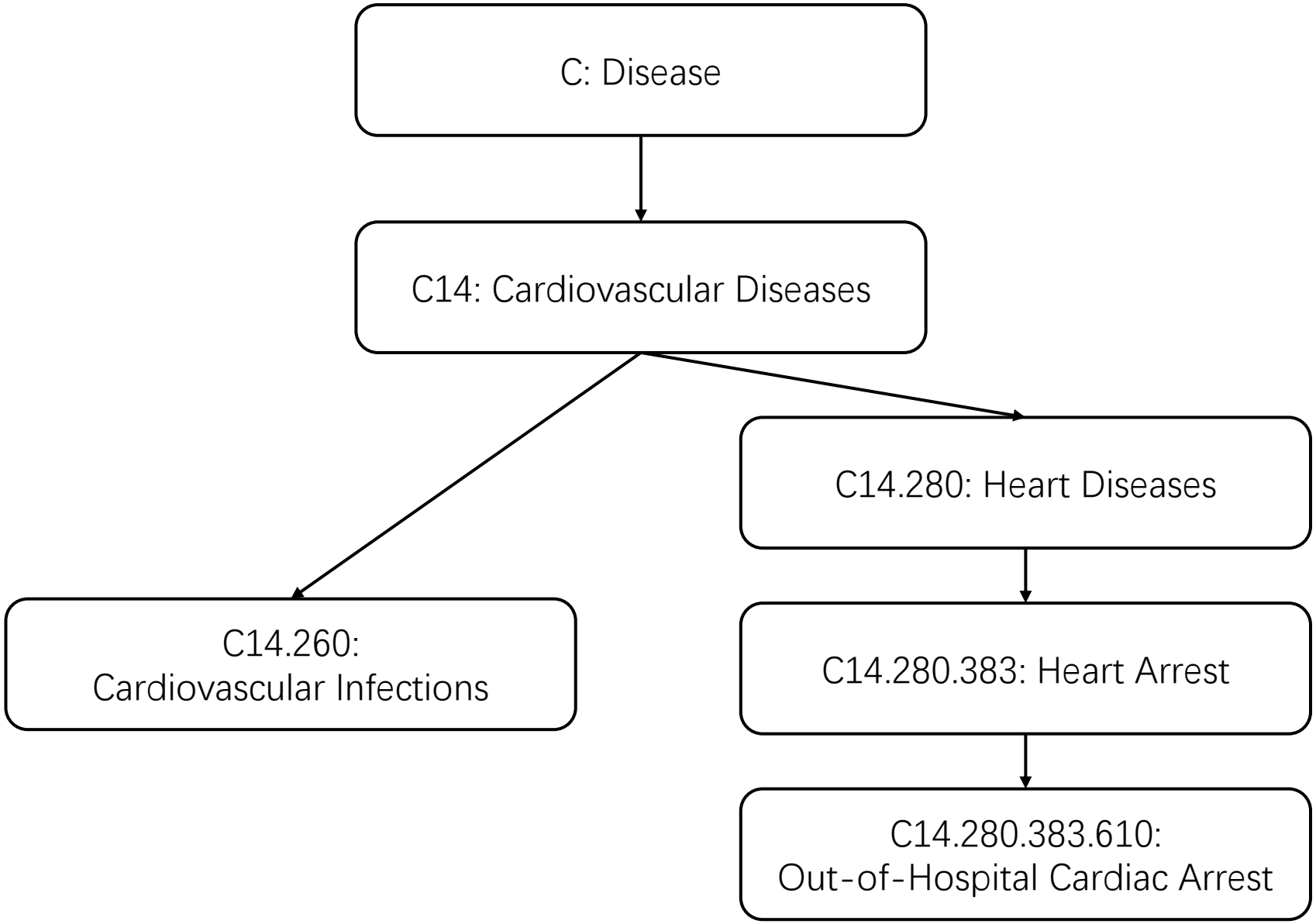
[46] Dong Wang, Juan Wang, Ming Lu, Fei Song, and Qinghua Cui. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics*, 26(13):1644–1650, 2010.

[47] 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.

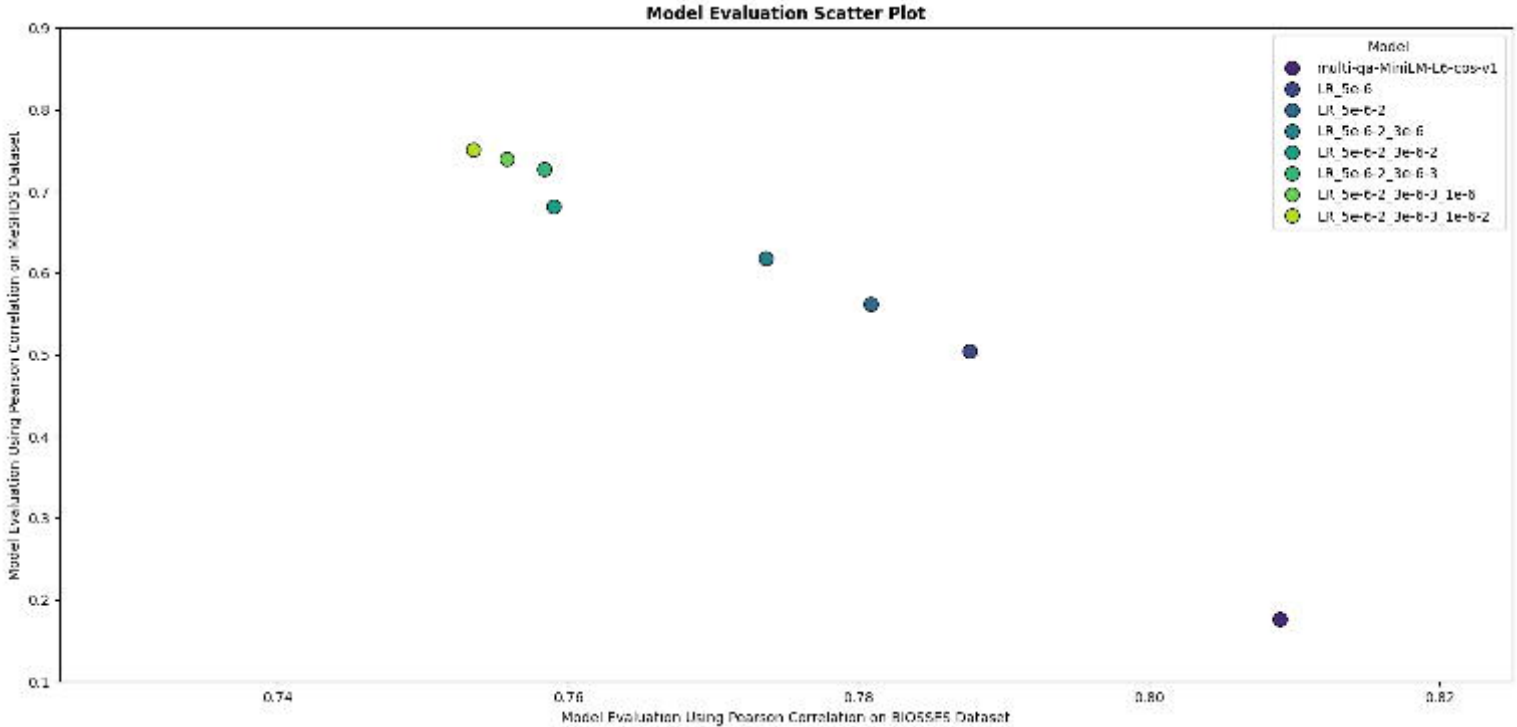
[48] Rajesha Rupaimoole and Frank J Slack. Microrna therapeutics: towards a new era for the management of cancer and other diseases. *Nature reviews Drug discovery*, 16(3):203–222, 2017.

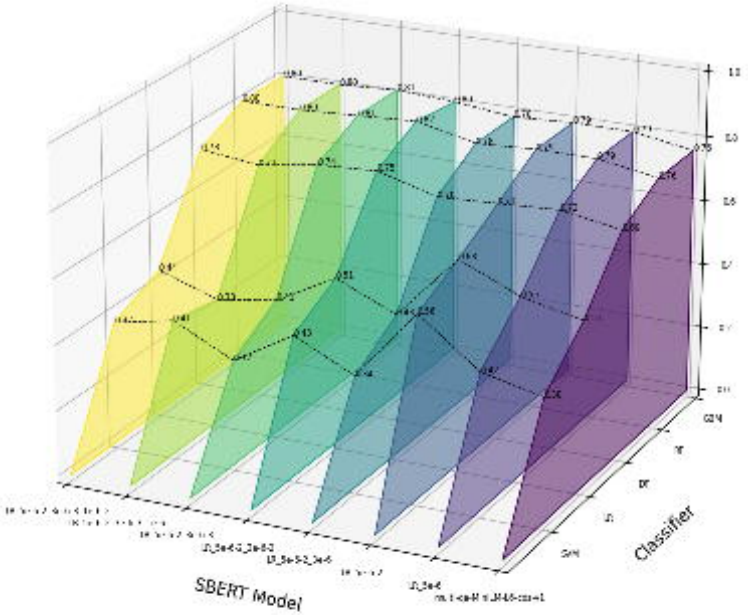

Data Collections
miRTarBase
miRWalk
miRDB
HMDD v4.0
DisGENET
MicroRNA Target Interactions (MTIs)
miRNA 1 — Gene 1: Exp Verify
miRNA 2 — Gene 2: Prediction
...
miRNA Disease
miRNA 1: D1, D2, D3, ...
miRNA 2: D4, D5, D6, ...
...
Gene Disease
Gene 1: d1, d2, d3, ...
Gene 2: d4, d5, d6, ...
...
Calculation of Diseases Similarity Matrix Between MTIs
Disease Similarity
dot-product(u · v)
u
v
Embedding
Embedding
Fine-tuned Sentence-Bert
Fine-tuned Sentence-Bert
miRNA Disease: D1, D2, ...
Gene Disease: d1, d2, ...
D1 D2 D3 ...
d1 s1 s2 s3 ...
d2 s4 s5 s6 ...
d3 s7 s8 s9 ...
Normalization of Similarity Matrix
S1 = {s1, s2, s3, ...}
S1N = [I1, I2, I3, I4, I5]
Proportion of Similarity Values from S1 range in 0 – 0.2
Proportion of Similarity Values from S1 range in 0.6 – 0.8
SN = {S1N, S2N, S3N, ...}
Random Forest Classifier Construction
Feature = S'N = {S1N', S2N', S3N', ...} (S'N ∈ SN)
Label = {Exp Verify, Predict, Predict, ...}
(nExp Verify = nPredict)
Tree-1
Tree-2
Tree-n
Major Voting
Random Forest Prediction
miRDS Database Construction
SN = {S1N, S2N, S3N, ...}
Random Forest Classifier
Protein-level Interaction
Prediction Interaction

Sentence 1	Sentence 2	Similarity (0-4)
It has recently been shown that Craf is essential for Kras G12D-induced NSCLC.	It has recently become evident that Craf is essential for the onset of Kras-driven non-small cell lung cancer.	4
The Bcl-2 inhibitor ABT-737 induces regression of solid tumors and its derivatives are in the early clinical phase as cancer therapeutics; however, it targets Bcl-2, Bcl-XL, and Bcl-w, but not Mcl-1, which induces resistance against apoptotic cell death triggered by ABT-737.	Recently, it has been reported that ABT-737 is not cytotoxic to all tumors cells, and that chemoresistance to ABT-737 is dependent on appreciable levels of Mcl-1 expression, the one Bcl-2 family member it does not effectively inhibit.	3
...		


C: Disease
C14: Cardiovascular Diseases
C14.280: Heart Diseases
C14.260:
Cardiovascular Infections
C14.280.383: Heart Arrest
C14.280.383.610:
Out-of-Hospital Cardiac Arrest

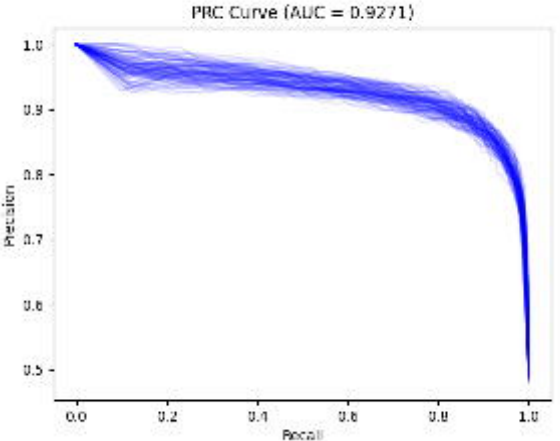
Disease 1	Disease 2	Similarity (0-1)
Acute Abdomen	Morton Neuroma	0.446130547
Acute Abdomen	Cellulite	0.310198272
...		

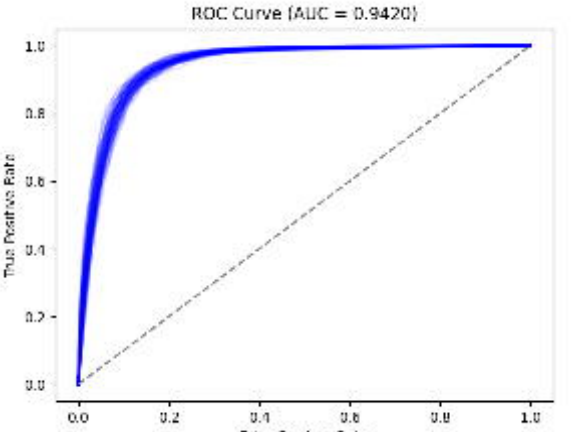

Model Evaluation Scatter Plot
Model
multi-qa-MiniLM-L6-cos-v1
LR_5e-6
LR_5e-6-2
LR_5e-6-2_3e-6
LR_5e-6-2_3e-6-2
LR_5e-6-2_3e-6-3
LR_5e-6-2_3e-6-3_1e-6
LR_5e-6-2_3e-6-3_1e-6-2
Model Evaluation Using Pearson Correlation on MedNLI's Dataset
Model Evaluation Using Pearson Correlation on BIOSSES Dataset
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.74
0.76
0.78
0.80
0.82

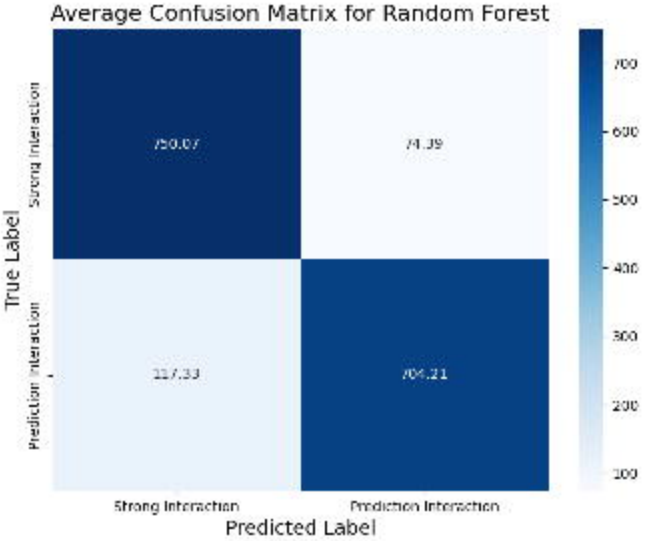

SBERT Model
Classifier

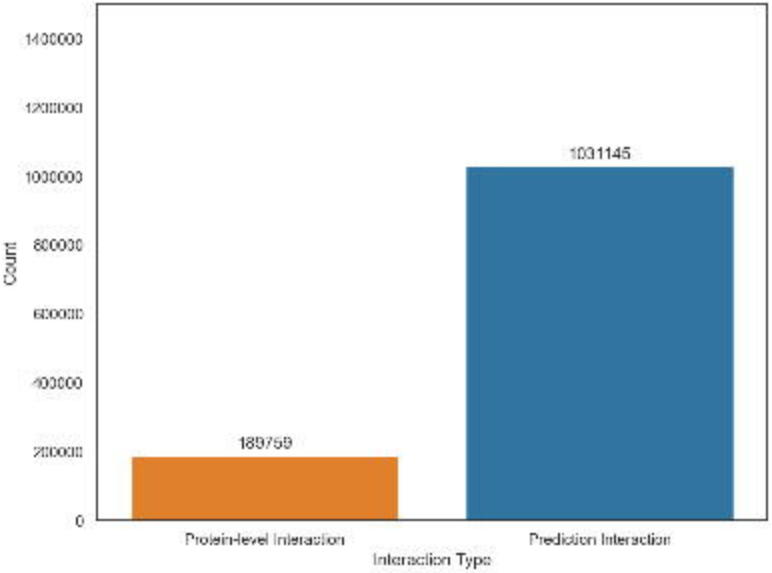
SLC8A1
MRK2
RPS6KA5
DYRK1A
TPM3
PPM1L
POLK
HPGD
NE2
RBM28
PCDH9
EPB41
NDUFA10
ZNF81
AP1M1
FXN
PCYT1B
SDR42E1
IL31RA
TTBK1
GREM1
MTAP
KCNMA1
VSTM4
AKT3
KCNN3
LPR
SMAD2
HNF4A
FGF1
BUB3
SELENOI
ONECUT2
DNAL1
NFIB
NUPR1
PDE1C
KIF1B
NSD2
SLC1A2
PSEN1
THY1
ABLIM1
MCPH1
HOOK3
NHEJ1
PIK3C3
IL6ST
NDM2
TK2
PPARA
CCDC134
RNF115
CAMK2A
FOXP1
GNAL
GNB5
DDX3
AP4S1
CALM1
LRPAP1
FAM83F
CAMK4
THBP
DNASH1
CTD
PPARGC1A
CLPB
CAPN1
CTSD
CSNK2
CDX1
CDX2
CDH1
CDH2
CDH3
CDH4
CDH5
CDH6
CDH7
CDH8
CDH9
CDH10
CDH11
CDH12
CDH13
CDH14
CDH15
CDH16
CDH17
CDH18
CDH19
CDH20
CDH21
CDH22
CDH23
CDH24
CDH25
CDH26
CDH27
CDH28
CDH29
CDH30
CDH31
CDH32
CDH33
CDH34
CDH35
CDH36
CDH37
CDH38
CDH39
CDH40
CDH41
CDH42
CDH43
CDH44
CDH45
CDH46
CDH47
CDH48
CDH49
CDH50
CDH51
CDH52
CDH53
CDH54
CDH55
CDH56
CDH57
CDH58
CDH59
CDH60
CDH61
CDH62
CDH63
CDH64
CDH65
CDH66
CDH67
CDH68
CDH69
CDH70
CDH71
CDH72
CDH73
CDH74
CDH75
CDH76
CDH77
CDH78
CDH79
CDH80
CDH81
CDH82
CDH83
CDH84
CDH85
CDH86
CDH87
CDH88
CDH89
CDH90
CDH91
CDH92
CDH93
CDH94
CDH95
CDH96
CDH97
CDH98
CDH99
CDH100
CDH101
CDH102
CDH103
CDH104
CDH105
CDH106
CDH107
CDH108
CDH109
CDH110
CDH111
CDH112
CDH113
CDH114
CDH115
CDH116
CDH117
CDH118
CDH119
CDH120
CDH121
CDH122
CDH123
CDH124
CDH125
CDH126
CDH127
CDH128
CDH129
CDH130
CDH131
CDH132
CDH133
CDH134
CDH135
CDH136
CDH137
CDH138
CDH139
CDH140
CDH141
CDH142
CDH143
CDH144
CDH145
CDH146
CDH147
CDH148
CDH149
CDH150
CDH151
CDH152
CDH153
CDH154
CDH155
CDH156
CDH157
CDH158
CDH159
CDH160
CDH161
CDH162
CDH163
CDH164
CDH165
CDH166
CDH167
CDH168
CDH169
CDH170
CDH171
CDH172
CDH173
CDH174
CDH175
CDH176
CDH177
CDH178
CDH179
CDH180
CDH181
CDH182
CDH183
CDH184
CDH185
CDH186
CDH187
CDH188
CDH189
CDH190
CDH191
CDH192
CDH193
CDH194
CDH195
CDH196
CDH197
CDH198
CDH199
CDH200
CDH201
CDH202
CDH203
CDH204
CDH205
CDH206
CDH207
CDH208
CDH209
CDH210
CDH211
CDH212
CDH213
CDH214
CDH215
CDH216
CDH217
CDH218
CDH219
CDH220
CDH221
CDH222
CDH223
CDH224
CDH225
CDH226
CDH227
CDH228
CDH229
CDH230
CDH231
CDH232
CDH233
CDH234
CDH235
CDH236
CDH237
CDH238
CDH239
CDH240
CDH241
CDH242
CDH243
CDH244
CDH245
CDH246
CDH247
CDH248
CDH249
CDH250
CDH251
CDH252
CDH253
CDH254
CDH255
CDH256
CDH257
CDH258
CDH259
CDH260
CDH261
CDH262
CDH263
CDH264
CDH265
CDH266
CDH267
CDH268
CDH269
CDH270
CDH271
CDH272
CDH273
CDH274
CDH275
CDH276
CDH277
CDH278
CDH279
CDH280
CDH281
CDH282
CDH283
CDH284
CDH285
CDH286
CDH287
CDH288
CDH289
CDH290
CDH291
CDH292
CDH293
CDH294
CDH295
CDH296
CDH297
CDH298
CDH299
CDH300
CDH301
CDH302
CDH303
CDH304
CDH305
CDH306
CDH307
CDH308
CDH309
CDH310
CDH311
CDH312
CDH313
CDH314
CDH315
CDH316
CDH317
CDH318
CDH319
CDH320
CDH321
CDH322
CDH323
CDH324
CDH325
CDH326
CDH327
CDH328
CDH329
CDH330
CDH331
CDH332
CDH333
CDH334
CDH335
CDH336
CDH337
CDH338
CDH339
CDH340
CDH341
CDH342
CDH343
CDH344
CDH345
CDH346
CDH347
CDH348
CDH349
CDH350
CDH351
CDH352
CDH353
CDH354
CDH355
CDH356
CDH357
CDH358
CDH359
CDH360
CDH361
CDH362
CDH363
CDH364
CDH365
CDH366
CDH367
CDH368
CDH369
CDH370
CDH371
CDH372
CDH373
CDH374
CDH375
CDH376
CDH377
CDH378
CDH379
CDH380
CDH381
CDH382
CDH383
CDH384
CDH385
CDH386
CDH387
CDH388
CDH389
CDH390
CDH391
CDH392
CDH393
CDH394
CDH395
CDH396
CDH397
CDH398
CDH399
CDH400
CDH401
CDH402
CDH403
CDH404
CDH405
CDH406
CDH407
CDH408
CDH409
CDH410
CDH411
CDH412
CDH413
CDH414
CDH415
CDH416
CDH417
CDH418
CDH419
CDH420
CDH421
CDH422
CDH423
CDH424
CDH425
CDH426
CDH427
CDH428
CDH429
CDH430
CDH431
CDH432
CDH433
CDH434
CDH435
CDH436
CDH437
CDH438
CDH439
CDH440
CDH441
CDH442
CDH443
CDH444
CDH445
CDH446
CDH447
CDH448
CDH449
CDH450
CDH451
CDH452
CDH453
CDH454
CDH455
CDH456
CDH457
CDH458
CDH459
CDH460
CDH461
CDH462
CDH463
CDH464
CDH465
CDH466
CDH467
CDH468
CDH469
CDH470
CDH471
CDH472
CDH473
CDH474
CDH475
CDH476
CDH477
CDH478
CDH479
CDH480
CDH481
CDH482
CDH483
CDH484
CDH485
CDH486
CDH487
CDH488
CDH489
CDH490
CDH491
CDH492
CDH493
CDH494
CDH495
CDH496
CDH497
CDH498
CDH499
CDH500
CDH501
CDH502
CDH503
CDH504
CDH505
CDH506
CDH507
CDH508
CDH509
CDH510
CDH511
CDH512
CDH513
CDH514
CDH515
CDH516
CDH517
CDH518
CDH519
CDH520
CDH521
CDH522
CDH523
CDH524
CDH525
CDH526
CDH527
CDH528
CDH529
CDH530
CDH531
CDH532
CDH533
CDH534
CDH535
CDH536
CDH537
CDH538
CDH539
CDH540
CDH541
CDH542
CDH543
CDH544
CDH545
CDH546
CDH547
CDH548
CDH549
CDH550
CDH551
CDH552
CDH553
CDH554
CDH555
CDH556
CDH557
CDH558
CDH559
CDH560
CDH561
CDH562
CDH563
CDH564
CDH565
CDH566
CDH567
CDH568
CDH569
CDH570
CDH571
CDH572
CDH573
CDH574
CDH575
CDH576
CDH577
CDH578
CDH579
CDH580
CDH581
CDH582
CDH583
CDH584
CDH585
CDH586
CDH587
CDH588
CDH589
CDH590
CDH591
CDH592
CDH593
CDH594
CDH595
CDH596
CDH597
CDH598
CDH599
CDH600
CDH601
CDH602
CDH603
CDH604
CDH605
CDH606
CDH607
CDH608
CDH609
CDH610
CDH611
CDH612
CDH613
CDH614
CDH615
CDH616
CDH617
CDH618
CDH619
CDH620
CDH621
CDH622
CDH623
CDH624
CDH625
CDH626
CDH627
CDH628
CDH629
CDH630
CDH631
CDH632
CDH633
CDH634
CDH635
CDH636
CDH637
CDH638
CDH639
CDH640
CDH641
CDH642
CDH643
CDH644
CDH645
CDH646
CDH647
CDH648
CDH649
CDH650
CDH651
CDH652
CDH653
CDH654
CDH655
CDH656
CDH657
CDH658
CDH659
CDH660
CDH661
CDH662
CDH663
CDH664
CDH665
CDH666
CDH667
CDH668
CDH669
CDH670
CDH671
CDH672
CDH673
CDH674
CDH675
CDH676
CDH677
CDH678
CDH679
CDH680
CDH681
CDH682
CDH683
CDH684
CDH685
CDH686
CDH687
CDH688
CDH689
CDH690
CDH691
CDH692
CDH693
CDH694
CDH695
CDH696
CDH697
CDH698
CDH699
CDH700
CDH701
CDH702
CDH703
CDH704
CDH705
CDH706
CDH707
CDH708
CDH709
CDH710
CDH711
CDH712
CDH713
CDH714
CDH715
CDH716
CDH717
CDH718
CDH719
CDH720
CDH721
CDH722
CDH723
CDH724
CDH725
CDH726
CDH727
CDH728
CDH729
CDH730
CDH731
CDH732
CDH733
CDH734
CDH735
CDH736
CDH737
CDH738
CDH739
CDH740
CDH741
CDH742
CDH743
CDH744
CDH745
CDH746
CDH747
CDH748
CDH749
CDH750
CDH751
CDH752
CDH753
CDH754
CDH755
CDH756
CDH757
CDH758
CDH759
CDH760
CDH761
CDH762
CDH763
CDH764
CDH765
CDH766
CDH767
CDH768
CDH769
CDH770
CDH771
CDH772
CDH773
CDH774
CDH775
CDH776
CDH777
CDH778
CDH779
CDH780
CDH781
CDH782
CDH783
CDH784
CDH785
CDH786
CDH787
CDH788
CDH789
CDH790
CDH791
CDH792
CDH793
CDH794
CDH795
CDH796
CDH797
CDH798
CDH799
CDH800
CDH801
CDH802
CDH803
CDH804
CDH805
CDH806
CDH807
CDH808
CDH809
CDH810
CDH811
CDH812
CDH813
CDH814
CDH815
CDH816
CDH817
CDH818
CDH819
CDH820
CDH821
CDH822
CDH823
CDH824
CDH825
CDH826
CDH827
CDH828
CDH829
CDH830
CDH831
CDH832
CDH833
CDH834
CDH835
CDH836
CDH837
CDH838
CDH839
CDH840
CDH841
CDH842
CDH843
CDH844
CDH845
CDH846
CDH847
CDH848
CDH849
CDH850
CDH851
CDH852
CDH853
CDH854
CDH855
CDH856
CDH857
CDH858
CDH859
CDH860
CDH861
CDH862
CDH863
CDH864
CDH865
CDH866
CDH867
CDH868
CDH869
CDH870
CDH871
CDH872
CDH873
CDH874
CDH875
CDH876
CDH877
CDH878
CDH879
CDH880
CDH881
CDH882
CDH883
CDH884
CDH885
CDH886
CDH887
CDH888
CDH889
CDH890
CDH891
CDH892
CDH893
CDH894
CDH895
CDH896
CDH897
CDH898
CDH899
CDH900
CDH901
CDH902
CDH903
CDH904
CDH905
CDH906
CDH907
CDH908
CDH909
CDH910
CDH911
CDH912
CDH913
CDH914
CDH915
CDH916
CDH917
CDH918
CDH919
CDH920
CDH921
CDH922
CDH923
CDH924
CDH925
CDH926
CDH927
CDH928
CDH929
CDH930
CDH931
CDH932
CDH933
CDH934
CDH935
CDH936
CDH937
CDH938
CDH939
CDH940
CDH941
CDH942
CDH943
CDH944
CDH945
CDH946
CDH947
CDH948
CDH949
CDH950
CDH951
CDH952
CDH953
CDH954
CDH955
CDH956
CDH957
CDH958
CDH959
CDH960
CDH961
CDH962
CDH963
CDH964
CDH965
CDH966
CDH967
CDH968
CDH969
CDH970
CDH971
CDH972
CDH973
CDH974
CDH975
CDH976
CDH977
CDH978
CDH979
CDH980
CDH981
CDH982
CDH983
CDH984
CDH985
CDH986
CDH987
CDH988
CDH989
CDH990
CDH991
CDH992
CDH993
CDH994
CDH995
CDH996
CDH997
CDH998
CDH999
CDH1000

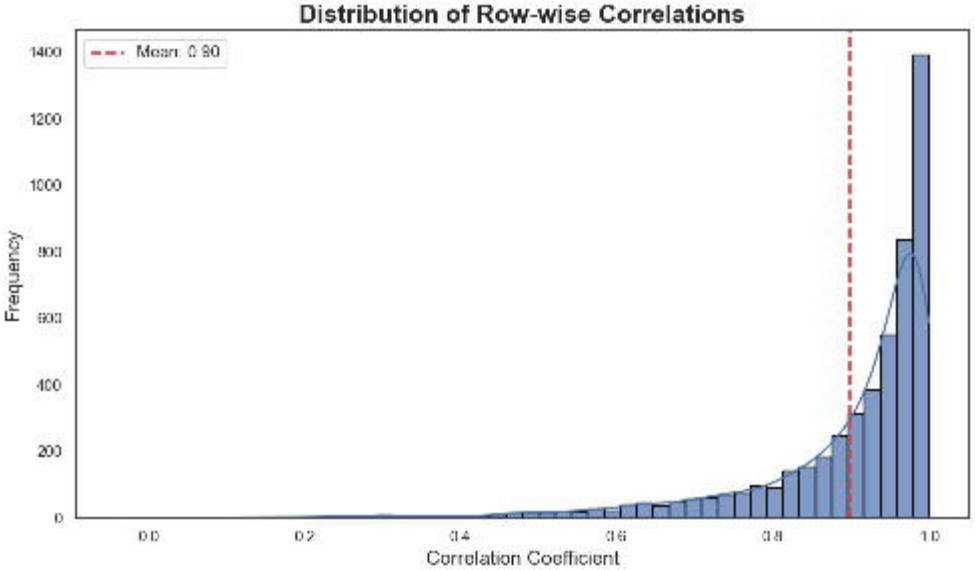
hsa-mir-6740
hsa-mir-6833
hsa-mir-6747
hsa-mir-7106
hsa-mir-4433a
hsa-mir-7110
hsa-mir-6780b
hsa-mir-4646
hsa-mir-6875
hsa-mir-4667
hsa-mir-4728
hsa-mir-149
hsa-mir-6887
hsa-mir-5010


PRC Curve (AUC = 0.9271)
1.0
0.9
0.8
0.7
0.6
0.5
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall


ROC Curve (AUC = 0.9420)
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate


Average Confusion Matrix for Random Forest
True Label
Strong Interaction
Prediction Interaction
750.07
74.39
117.33
704.21
Strong Interaction
Prediction Interaction
Predicted Label
700
600
500
400
300
200
100


1400000
1200000
1000000
800000
600000
400000
200000
0
Count
1031145
189759
Protein-level Interaction
Prediction Interaction
Interaction Type


Distribution of Row-wise Correlations
Mean: 0.90
Frequency
Correlation Coefficient
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0

Welcome to the miRTarDS Database

Explore the disease connections between miRNA and gene to enhance your research.

Search for MTIs

Search for Disease

© 2024 miRTarDS. All Rights Reserved.

Search MicroRNA Target Interactions

Found **1220904** results in **61046** pages.

Search Pri-miRNA... Exact Search Search Pri-miRNA

Search Gene... Exact Search Search Gene

Pre-miRNA	Gene Symbol	MicroRNA Target Interactions (MTIs)	miRTarDS Prediction	E(x) ▲	MicroRNA Disease <=> Gene Disease : Similarity
hsa-mir-6893	DDIT4L	hsa-miR-6893-5p —— DDIT4L: miRWalk	Prediction Interaction	0.05	
hsa-mir-6854	PTCD3	hsa-miR-6854-5p —— PTCD3: miRWalk	Prediction Interaction	0.05	
hsa-mir-4537	MOB3A	hsa-miR-4537 —— MOB3A: miRWalk	Prediction Interaction	0.05	
hsa-mir-3714	MBOAT2	hsa-miR-3714 —— MBOAT2: miRWalk;miRDB	Prediction Interaction	0.05	
hsa-mir-6765	MOB3A	hsa-miR-6765-5p —— MOB3A: miRWalk hsa-miR-6765-3p —— MOB3A: miRWalk	Prediction Interaction	0.05	
hsa-mir-7161	MOB3A	hsa-miR-7161-5p —— MOB3A: miRWalk hsa-miR-7161-3p —— MOB3A: miRWalk	Prediction Interaction	0.05	
hsa-mir-1273h	GPX8	hsa-miR-1273h-5p —— GPX8: miRWalk hsa-miR-1273h-3p —— GPX8: miRWalk	Prediction Interaction	0.05	
hsa-mir-6788	RPL38	hsa-miR-6788-5p —— RPL38: miRWalk hsa-miR-6788-3p —— RPL38: miRWalk	Prediction Interaction	0.05	
hsa-mir-6516	MOB3A	hsa-miR-6516-5p —— MOB3A: miRWalk	Prediction Interaction	0.05	
hsa-mir-4733	KRT85	hsa-miR-4733-5p —— KRT85: miRWalk hsa-miR-4733-3p —— KRT85: miRWalk	Prediction Interaction	0.05	
hsa-mir-6792	GPX8	hsa-miR-6792-5p —— GPX8: miRWalk hsa-miR-6792-3p —— GPX8: miRWalk	Prediction Interaction	0.05	
hsa-mir-4633	MOB3A	hsa-miR-4633-5p —— MOB3A: miRWalk hsa-miR-4633-3p —— MOB3A: miRWalk	Prediction Interaction	0.05	

© 2024 miRTarDS. All Rights Reserved.

Search Disease Related MTIs

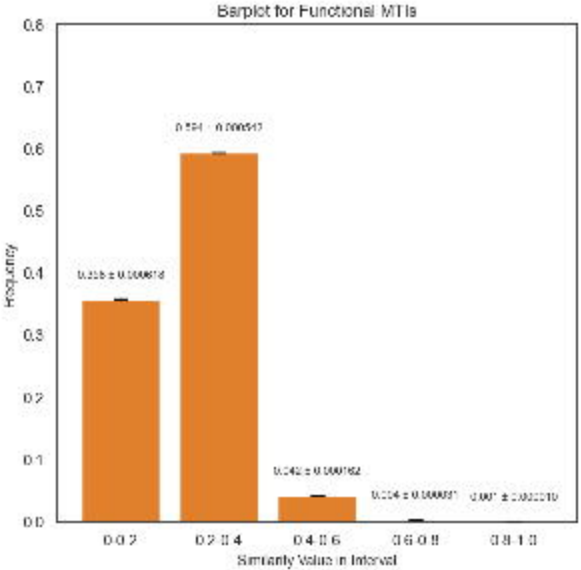
Found **38340** results in **1917** pages.

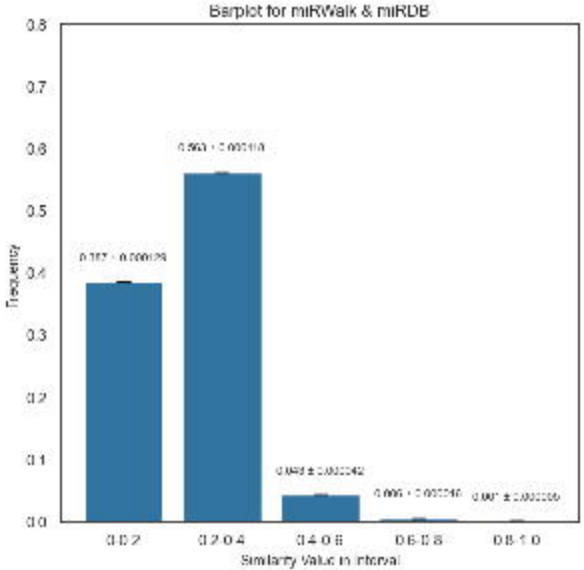
Cancer

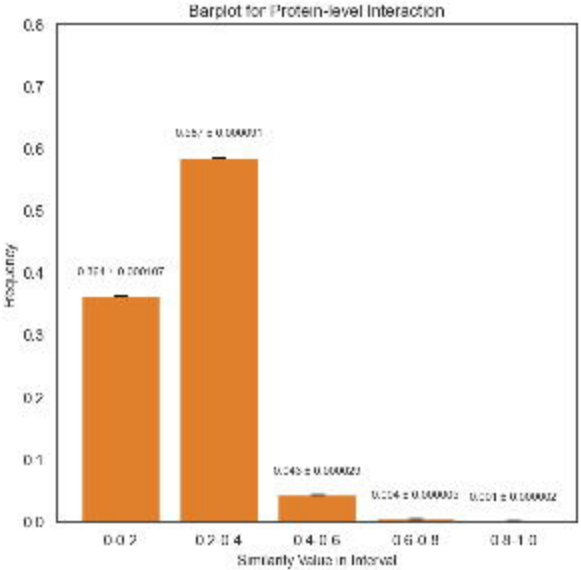
Search Disease

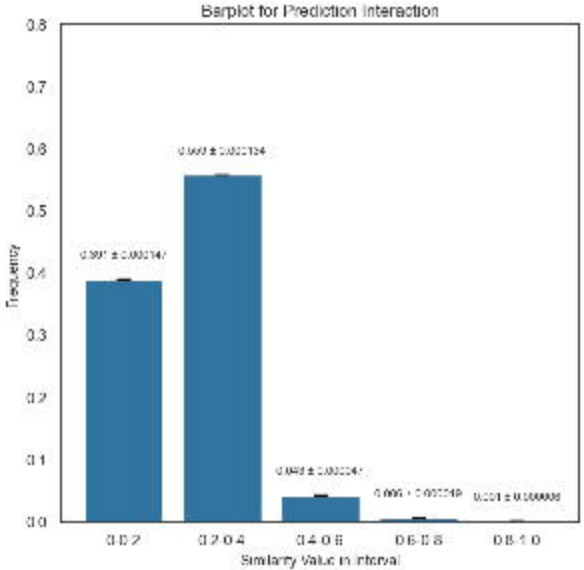
Pre-miRNA	Gene Symbol	MicroRNA Target Interactions (MTIs)	miRTarDS Prediction	E(x) ▼	MicroRNA Disease <=> Gene Disease : Similarity
hsa-mir-6763	BAG5	hsa-miR-6763-5p —— BAG5: miRWalk hsa-miR-6763-3p —— BAG5: miRWalk	Prediction Interaction	0.21	Ovarian Epithelial Carcinoma<=>Ovarian Epithelial Carcinoma:0.9999999403953552 Ovarian Epithelial Carcinoma<=>Epithelial ovarian cancer:0.8238090872764587
hsa-mir-6763	ARHGEF10L	hsa-miR-6763-3p —— ARHGEF10L: miRWalk	Prediction Interaction	0.2	Ovarian Epithelial Carcinoma<=>Ovarian Epithelial Carcinoma:0.9999999403953552 Ovarian Epithelial Carcinoma<=>Epithelial ovarian cancer:0.8238090872764587
hsa-mir-4479	CDK14	hsa-miR-4479 —— CDK14: miRWalk	Prediction Interaction	0.2	Ovarian Epithelial Carcinoma<=>Ovarian Epithelial Carcinoma:0.9999999403953552 Ovarian Epithelial Carcinoma<=>Epithelial ovarian cancer:0.8238090872764587
hsa-mir-6763	BABAM1	hsa-miR-6763-5p —— BABAM1: miRWalk hsa-miR-6763-3p —— BABAM1: miRWalk	Prediction Interaction	0.19	Ovarian Epithelial Carcinoma<=>Ovarian Epithelial Carcinoma:0.9999999403953552 Ovarian Epithelial Carcinoma<=>Epithelial ovarian cancer:0.8238090872764587
hsa-mir-7156	HILPDA	hsa-miR-7156-3p —— HILPDA: miRWalk	Prediction Interaction	0.19	Renal Cell Carcinoma<=>Renal Cell Carcinoma:0.9999999403953552 Renal Cell Carcinoma<=>Metastatic Renal Cell Cancer:0.8292976021766663 Renal Cell Carcinoma<=>Metastatic Renal Cell Carcinoma:0.9261288642883301 Renal Cell Carcinoma<=>Acquired cystic disease associated renal cell carcinoma:0.8012170195579529
hsa-mir-6774	LIMD2	hsa-miR-6774-5p —— LIMD2: miRWalk hsa-miR-6774-3p —— LIMD2: miRWalk	Prediction Interaction	0.19	Papillary Thyroid Cancer<=>Papillary thyroid carcinoma:0.8544691205024719
hsa-mir-6774	FAM83F	hsa-miR-6774-5p —— FAM83F: miRWalk hsa-miR-6774-3p —— FAM83F: miRWalk	Prediction Interaction	0.19	Papillary Thyroid Cancer<=>Papillary thyroid carcinoma:0.8544691205024719
hsa-mir-6775	FAM83F	hsa-miR-6775-3p —— FAM83F: Functional MTI (Weak);miRWalk hsa-miR-6775-5p —— FAM83F: miRWalk	Prediction Interaction	0.19	Papillary Thyroid Cancer<=>Papillary thyroid carcinoma:0.8544691205024719
hsa-mir-6763	RACGAP1	hsa-miR-6763-5p —— RACGAP1: miRWalk hsa-miR-6763-3p —— RACGAP1: miRWalk	Prediction Interaction	0.19	Esophageal Squamous Cell Carcinoma<=>Squamous cell carcinoma:0.8734997510910034 Ovarian Epithelial Carcinoma<=>Ovarian Epithelial Carcinoma:0.9999999403953552 Ovarian Epithelial Carcinoma<=>Epithelial ovarian cancer:0.8238090872764587
hsa-mir-6763	CDK14	hsa-miR-6763-3p —— CDK14: miRWalk	Prediction Interaction	0.18	Ovarian Epithelial Carcinoma<=>Ovarian Epithelial Carcinoma:0.9999999403953552 Ovarian Epithelial Carcinoma<=>Epithelial ovarian cancer:0.8238090872764587

© 2024 miRTarDS. All Rights Reserved.


Barplot for Functional MTIs
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0 0.2
0.2 0.4
0.4 0.6
0.6 0.8
0.8 1.0
Similarity Value in Interval


Barplot for miRWalk & miRDB
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0 0.2
0.2 0.4
0.4 0.6
0.6 0.8
0.8 1.0
Similarity Value in Interval


Barplot for Protein-level Interaction
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0 0.2
0.2 0.4
0.4 0.6
0.6 0.8
0.8 1.0
Similarity Value in Interval


Barplot for Prediction Interaction
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0 0.2
0.2 0.4
0.4 0.6
0.6 0.8
0.8 1.0
Similarity Value in Interval