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

October 17, 2024

Abstract

MicroRNAs (miRNAs) regulate gene expression by binding to mRNAs, inhibiting translation, or promoting mRNA degradation. miRNAs are of great importance in the development of various diseases. Currently, a variety of miRNA target prediction tools are available, that 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 of human sequence-based predicted MTIs by considering the disease similarity between miRNAs and genes to establish a prediction database of disease-specific MTIs. We fine-tuned the Sentence-BERT model to calculate the semantic similarity of disease. The method achieved an F1 score of 0.88 in accurately distinguishing human protein-level experimentally (Western Blot, Reporter Assay, etc.) validated MTIs and predicted MTIs. Moreover, this method exhibits exceptional generalizability across different databases. The proposed method was utilized to calculate the similarity of disease in 1,220,904 human 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

*baimingchen@link.cuhk.edu.cn

miRNAs have been discovered with the development of high-throughput bioinformatics technologies [2]. miRNAs are classified as small non-coding RNAs (ncRNAs), 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 miRNAs and genes to construct a database for MTIs prediction in different species [7]. miRDB considers miRNA overexpression and CLIP-seq data, utilizing a support vector machine (SVM) for MTI prediction in different species [8]. mirSVR trains a regression model to rank miRNA target sites based on the 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 miRNAs 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 experimental results confirm this hypothesis. The similarity between two diseases can be assessed using the SBERT model. We generated similarity matrices by computing pairwise disease similarities between miRNAs and genes. The matrices were analyzed to determine the statistical distribution of its values, which was 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) outling 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.

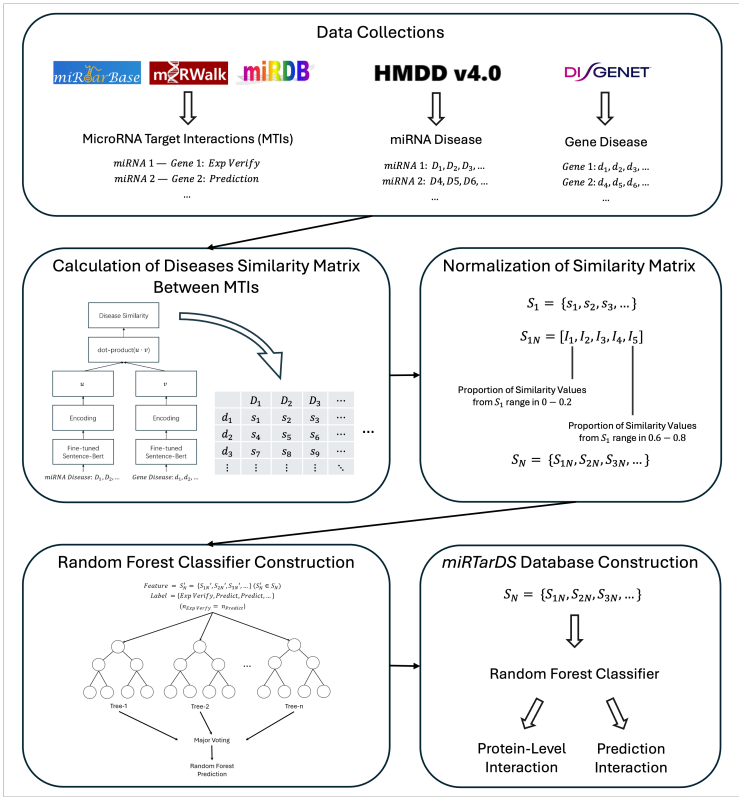

Figure 1: Flow Chart of Method

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. We utilized the miRTarBase [19], miRWalk [7] and miRDB [8] to obtain human protein-level experimentally verified MTIs and predicted MTIs. 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 by 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] databases. The 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. The DisGeNET and the KEGG databases 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 for 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:

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

A decay factor of 0.8 was selected in this study. 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.

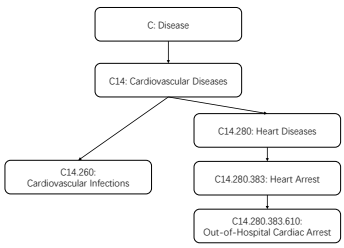

Figure 3: Disease Descriptor 'Cardiovascular Infections' and 'Out-of-Hospital Cardiac Arrest' Hierarchical Structure in MeSH Tree

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

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

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.

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

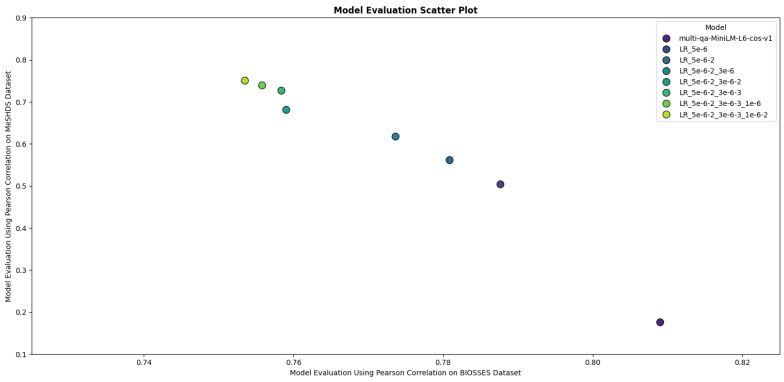

Figure 6: Model Performance Benchmark. The figure compares the performance of different fine-tuning strategies, as described below:

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.

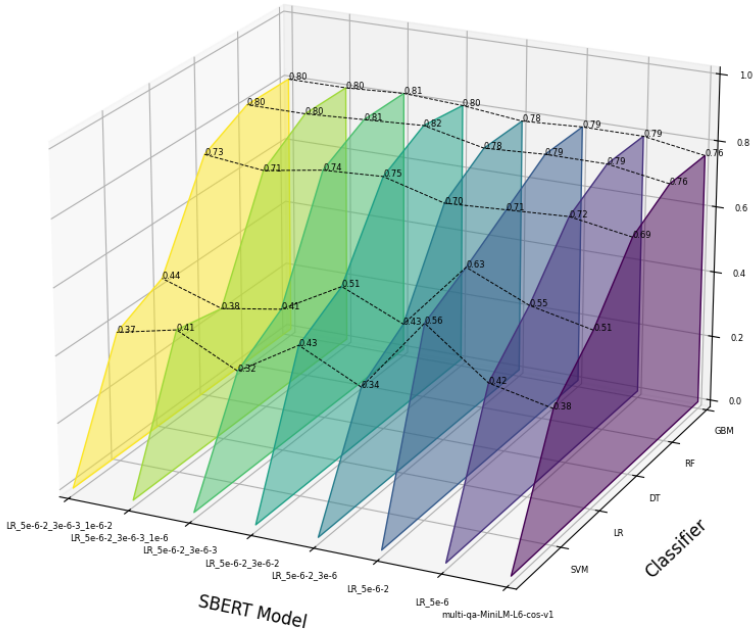

Figure 7: F1 Scores Across Different SBERT Models and Classification Methods.

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).

Figure 8: Gene Symbol Word Cloud

Figure 9: Pre-miRNA Word Cloud

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, Table 2), with an ROC AUC of 0.94 (Figure 10) and a PRC AUC of 0.93 (Figure 11), demonstrating strong classification performance. Subsequently, the trained Random Forest model was applied to all MTIs, resulting in 1,220,904 predictions. Of these predictions, 189,759 were classified as Protein-level Interaction while 1,031,145 were classified as Prediction Interaction (Figure 13). We accurately recognized 4,039 out of 4,113 Functional MTIs from miRTarbase using our method.

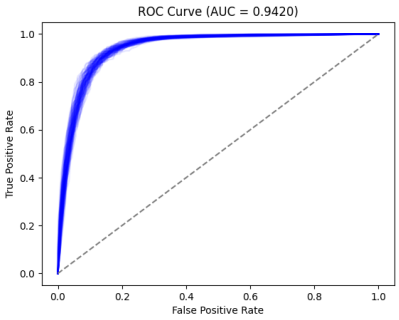

Figure 10: ROC Curve of Random Forest Classifier

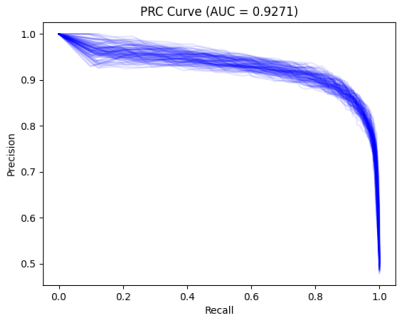

Figure 11: PRC Curve of Random Forest Classifier

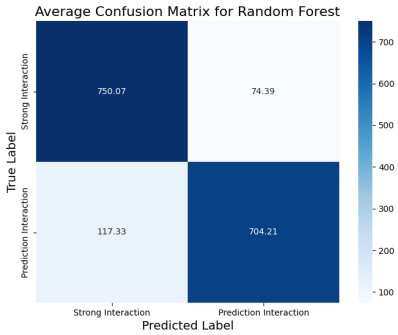

Figure 12: Confusion Matrix Analysis: 80/20 Train-Test Split with 100 Iterations

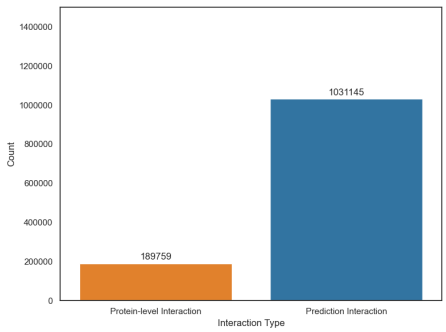

Figure 13: Distribution of Predicted Interaction Labels

Table 2: Random Forest Model Evaluation Metrics

Metric	Formula	Value
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	0.884
Precision	$\frac{TP}{TP+FP}$	0.910
Recall	$\frac{TP}{TP+FN}$	0.865
F1 Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$	0.887

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 generalizability of the method was assessed by testing its performance on different databases. We manually downloaded the gene-associated disease data from the KEGG database, and within the 20 previously mentioned genes, only TP53, PTEN, KRAS, MYC, TNF, IL6, IL10, and CRP have enough gene-associated diseases in KEGG. These genes were selected to generate 5,263 MTIs' disease similarity matrices with both the DisGeNET and the KEGG databases. The correlation coefficients between the normalized disease similarity lists of each MTI across the two databases were calculated. The average Pearson's correlation coefficient is approximately 0.90 (Figure 14). This result provides evidence that the proposed method exhibits perfect generalizability across different databases.

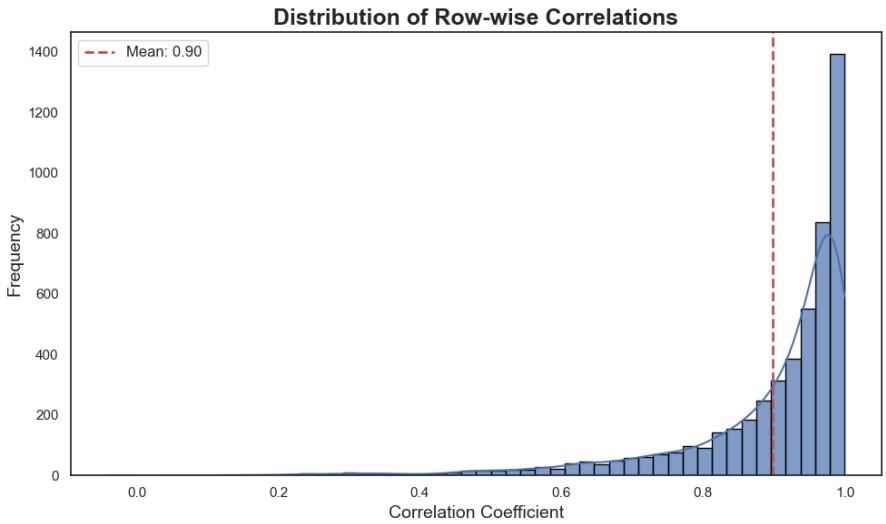

Figure 14: Row-wise Distribution of Normalized Disease Similarity List Correlation Coefficients Between DisGeNET and KEGG Databases

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 miRNA-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. Source data are provided at the end of the paper, and the website is scheduled to be launched soon.

Figure 15: Home Page

Figure 16: Search MTI

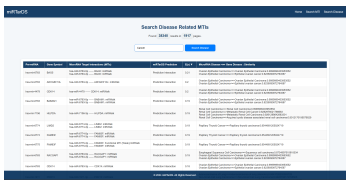

Figure 17: Search disease

Figure 18: miRTarDS Database

4 Discussion

The present study introduces a method to refine sequence-based predicted MTIs by considering the 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 prediction data generated by the random forest classifier (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 lower frequency 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.

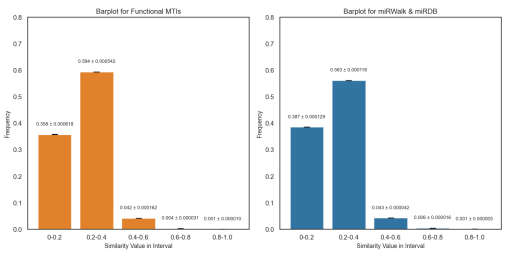

Figure 19: Average Value in Normalized Disease Similarity List of Functional MTIs and MTIs from miRWalk & miRDB (Real Data)

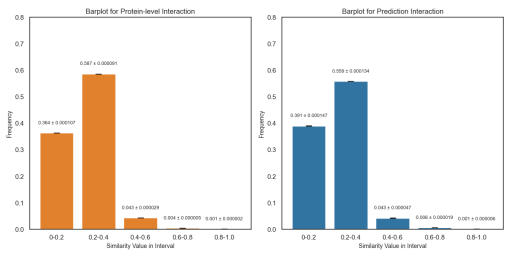

Figure 20: Average Value in Normalized Disease Similarity List of Protein-level Interaction and Prediction Interaction (Random Forest Prediction)

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 this estimation, the method can encompass a substantial proportion of both miRNAs and genes. More importantly, compared to previous disease similarity algorithms, such as 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.

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 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 semantic similarity and functional semantic similarity, thereby optimizing the analysis.

Declarations

Acknowledgments

Throughout this paper, the pronoun "We" is used instead of "I" to follow academic conventions. The computing resources of the Chinese University of Hong Kong, Shenzhen have not been utilized for this project, as I hereby affirm. Without contacting me, please do not integrate the data from this study 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 obtained from `Baiming123/MeSHDS` and `Baiming123/Calcu_Disease_Similarity` on HuggingFace, respectively. miRTarDS source data are available at `AltasChen/miRTarDS` on GitHub.

Patent

Two Chinese patent applications related to this work have been filed. Patent Application No.202410559682.3 and Application No.202411135644.1. Both patents

passed the preliminary examination.

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.