Author/s:
Uthayopas, K;de Sá, AGC;Alavi, A;Pires, DEV;Ascher, DB

Title:
TSMDA: Target and symptom-based computational model for miRNA-disease-association prediction

Date:
2021-12-03

Citation:
Uthayopas, K., de Sá, A. G. C., Alavi, A., Pires, D. E. V. & Ascher, D. B. (2021). TSMDA: Target and symptom-based computational model for miRNA-disease-association prediction. Molecular Therapy Nucleic Acids, 26, pp.536-546. https://doi.org/10.1016/j.omtn.2021.08.016.

Persistent Link:
https://hdl.handle.net/11343/290091

Molecular Therapy
Nucleic Acids
Original Article

# TSMDA: Target and symptom-based computational model for miRNA-disease-association prediction

Korawich Uthayopas,[1,2,3] Alex G.C. de Sá,[1,2,3,4] Azadeh Alavi,[1,2,3] Douglas E.V. Pires,[1,2,3,5] and David B. Ascher[1,2,3,4,6]

[1]Structural Biology and Bioinformatics, Department of Biochemistry, University of Melbourne, Parkville 3052, VIC, Australia; [2]Systems and Computational Biology, Bio21 Institute, University of Melbourne, Parkville 3052, VIC, Australia; [3]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne 3004, VIC, Australia; [4]Baker Department of Cardiometabolic Health, Melbourne Medical School, University of Melbourne, Parkville 3010, VIC, Australia; [5]School of Computing and Information Systems, University of Melbourne, Parkville 3052, VIC, Australia; [6]Department of Biochemistry, University of Cambridge, 80 Tennis Ct Rd, Cambridge CB2 1GA, UK

**The emergence of high-throughput sequencing techniques has revealed a primary role of microRNAs (miRNAs) in a wide range of diseases, including cancers and neurodegenerative disorders. Understanding novel relationships between miRNAs and diseases can potentially unveil complex pathogenesis mechanisms, leading to effective diagnosis and treatment. The investigation of novel miRNA-disease associations, however, is currently costly and time consuming. Over the years, several computational models have been proposed to prioritize potential miRNA-disease associations, but with limited usability or predictive capability. In order to fill this gap, we introduce TSMDA, a novel machine-learning method that leverages target and symptom information and negative sample selection to predict miRNA-disease association. TSMDA significantly outperforms similar methods, achieving an area under the receiver operating characteristic (ROC) curve (AUC) of 0.989 and 0.982 under 5-fold cross-validation and blind test, respectively. We also demonstrate the capability of the method to uncover potential miRNA-disease associations in breast, prostate, and lung cancers, as case studies. We believe TSMDA will be an invaluable tool for the community to explore and prioritize potentially new miRNA-disease associations for further experimental characterization. The method was made available as a freely accessible and user-friendly web interface at http://biosig.unimelb.edu.au/tsmda/.**

## INTRODUCTION

MicroRNAs (miRNAs) are small regulatory non-coding RNAs with a typical length of 21–25 nucleotides. Human mature miRNAs control the gene expression of target messenger RNAs (mRNAs) by partially complementary base pairing with the 3′ untranslated region.[1] This interaction generally results in post-transcriptional repression, occasionally leading to miRNA degradation.[2] Various physiological processes, such as cell proliferation and cell death, are regulated by a complex network of miRNAs.[2]

The advent of high-throughput sequencing techniques has been contributing to the growing evidence of associations between miRNAs and diseases. Deregulation of several miRNAs is correlated with the development of multiple diseases, such as cancers and brain and cardiovascular diseases.[3–5] For example, pancreatic carcinogenesis may occur from the upregulation of miR-21, miR-155, miR-181, miR-221, and miR-222.[6] Hence, understanding the relationship between miRNAs and diseases might shed light on pathogenesis, promoting miRNA-based applications such as biomarkers or drugs.[7–9] Currently, a significant number of disease-related miRNAs are experimentally confirmed and collected in multiple databases.[10–12] Despite these significant efforts, large-scale exploration of the potential disease-miRNA associations is unfeasible, since experimental validation is laborious and costly. In this context, effective computational methods are urgently needed to suggest potential associations and guide experimental efforts.

Diverse machine-learning models have been extensively implemented to assist in exploring miRNA-disease relationships.[13–22] From the widely accepted assumption that phenotypically similar diseases and functionally equivalent miRNAs tend to be associated, experimentally confirmed associations can be used to identify novel associations. One model in particular, miRNA target-dysregulated network (MTDN), has been built to unveil potential cancer-related miRNAs.[13] One of the posterior advances is the random forest for miRNA-disease association (RFMDA),[14] which is based on miRNA functional similarity (MISIM)[23] and disease semantic similarity,[23,24] as features to perform the miRNA-disease-association predictions.

Despite the remarkable effort of currently available methods, model performance was still limited by miRNA and disease similarity estimations that did not directly reflect miRNA mechanisms and disease pathogenesis. The performance improvement obtained

Received 14 May 2021; accepted 19 August 2021;
https://doi.org/10.1016/j.omtn.2021.08.016.

**Correspondence:** David B. Ascher, Structural Biology and Bioinformatics, Department of Biochemistry, University of Melbourne, Parkville 3052, VIC, Australia.
**E-mail:** david.ascher@unimelb.edu
**Correspondence:** Douglas E.V. Pires, Systems and Computational Biology, Bio21 Institute, University of Melbourne, Parkville 3052, VIC, Australia.
**E-mail:** douglas.pires@unimelb.edu.au

**Table 1. Selected features and corresponding biological meaning**

| Feature | Category | Meaning |
|---|---|---|
| 1 | miRNA functional similarity (MISIM) | similarity with "hsa-miR-1180-3p" |
| 2 | miRNA functional similarity (MISIM) | similarity with "has-miR-3179" |
| 3 | miRNA functional similarity (MISIM) | similarity with "hsa-miR-320c" |
| 4 | miRNA functional similarity (MISIM) | similarity with "hsa-miR-376b-3p" |
| 5 | miRNA functional similarity (MISIM) | similarity with "hsa-miR-487a-3p" |
| 6 | target-based miRNA similarity | similarity with "hsa-miR-127-3p" |
| 7 | target-based miRNA similarity | similarity with "hsa-miR-184" |
| 8 | target-based miRNA similarity | similarity with "hsa-miR-516a-5p" |
| 9 | symptom-based disease similarity | similarity with "Alopecia (D000505)" |
| 10 | symptom-based disease similarity | similarity with "Biliary Atresia (D001656)" |
| 11 | symptom-based disease similarity | similarity with "Atopic dermatitis (D003876)" |
| 12 | symptom-based disease similarity | similarity with "Myelodysplastic Syndromes (D009190)" |
| 13 | symptom-based disease similarity | similarity with "Tourette Syndrome (D005879)" |

by two additional methods, latent feature extraction for miRNA-disease association (LFEMDA)[15] and distance-based sequence similarity for miRNA-disease association (DBMDA),[16] emphasize that the introduction of biological features, such as miRNA sequence, into similarity calculation is important. A lack of actual negative samples was also a significant challenge, where various methods randomly selected negative samples from miRNA-disease pairs without confirmed associations.[14,16,21] This approach likely leads to false negatives. Two previous models, non-negative samples extraction (NSEMDA)[17] and negative sample selection strategy and multi-layer perceptron (NMLPMDA),[18] have proposed alternative approaches to select reliable negative samples. NSEMDA iteratively filtered unknown samples with positive-unlabeled (PU) learning, an algorithm designed to deal with a labeling issue, where only a single class is available.[25,26] Alternatively, NMLPMDA utilized the miRNA-gene-disease network to remove likely associations.[18]

Here we propose a novel machine-learning model that employs target- and symptom-based similarity for miRNA-disease-association prediction (TSMDA). In this study, miRNA target genes and disease symptoms were introduced to enhance similarity calculation, coupled with reliable negative sample selections based on extended miRNA-gene-disease network and modified PU learning.

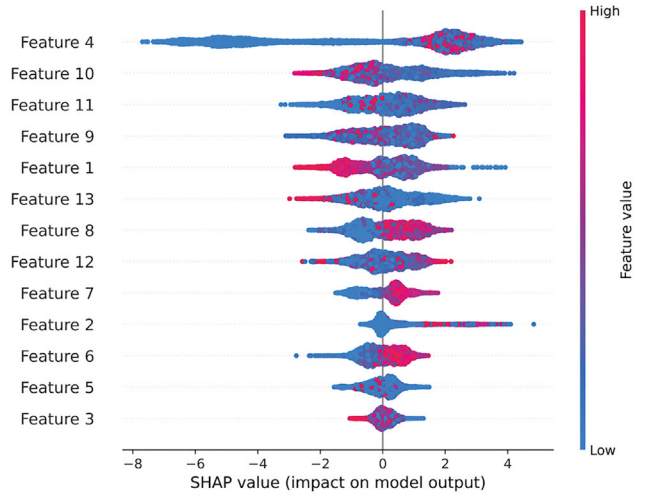


**Figure 1. Feature 4 is the most contributing feature to a prediction, showing a distinct positive correlation with a miRNA-disease association**

The SHAP value for each feature in the XGBoost model was calculated. The features are ranked based on the average impact on a model prediction. One dot represents one miRNA-disease association. The values of features are represented by color, red indicating high values and blue indicating low values.

## RESULTS

### Feature selection

In this study, two feature selection methods, a correlation-based and forward stepwise greedy feature selection,[27,28] were employed to select the minimal effective subset from 1,373 features to train a highly accurate model. As a result, 13 features were chosen. This subset consists of five miRNA functional similarities, three target-based miRNA similarities, and five symptom-based disease similarities (Table 1). It is adopted to train and validate the extreme gradient boosting (XGBoost) model.[29]

### Interpretation of the XGBoost model

Model interpretability is one of the essential aspects to consider before putting a machine learning model to use.[30–32] It is crucial for explaining the accuracy of model prediction and guiding performance improvement. Despite achieving high accuracy, popular complex models, such as XGBoost and neural networks,[29–33] are excessively complex for human interpretation. Different methods have been introduced to help understand the predictions in response to a lack of interpretability.[30–32] SHapley Additive exPlanations (SHAP) is one of the methods designed to explain a model by examining the contribution of each feature in terms of SHAP value to a prediction.[30] SHAP value is a measure of feature importance, calculated to exhibit the distribution of each feature's impact on a prediction. The benefits of SHAP values are computational efficiency and consistency with human explanations.[30]

In this work, we implemented SHAP to analyze how the trained XGBoost model makes a prediction. SHAP values of 13 selected features were calculated and displayed in Figure 1, where features are

**Table 2. The results of TSMDA based on a blind test, 5-fold, 10-fold, and 20-fold cross-validation in HMDD v.2.0**

| Methods | AUC | MCC | bACC | F1 |
|---|---|---|---|---|
| Blind test | 0.982 | 0.965 | 0.982 | 0.982 |
| 5-fold cross-validation | 0.989 ± 0.003 | 0.978 ± 0.005 | 0.989 ± 0.003 | 0.989 ± 0.003 |
| 10-fold cross-validation | 0.989 ± 0.004 | 0.978 ± 0.008 | 0.989 ± 0.004 | 0.989 ± 0.004 |
| 20-fold cross-validation | 0.989 ± 0.005 | 0.978 ± 0.010 | 0.989 ± 0.005 | 0.989 ± 0.005 |

ranked based on the average impact on model output in descending order. The most important feature is feature 4, representing the MISIM functional similarity with hsa-miR-376b. This miRNA is experimentally supported to be associated with a wide type of diseases, including adrenocortical carcinoma,[34] cerebral ischemia,[35] Graves' disease,[36] myocardial ischemia,[37] Parkinson's disease,[38] and prostate neoplasms.[39] According to a widely accepted assumption that similar miRNAs tend to be associated with phenotypically similar diseases, miRNAs with high feature 4 values will be more likely to be associated with these diseases or related conditions. This assumption is in accord with a remarkable positive correlation between feature 4 values and miRNA-disease associations in the figure. Similar trends can be clearly observed in features 6, 7, and 8 that represent target-based miRNA similarity.

Features 10, 11, and 9 are the 2nd, 3rd, and 4th most critical features, accounting for symptom-based disease similarities with biliary atresia, atopic dermatitis, and alopecia. In this case, they present an unclear correlation with miRNA-disease associations. This finding well accords with expectations, as many disease similarities are needed to be considered as a group to represent a particular disease.

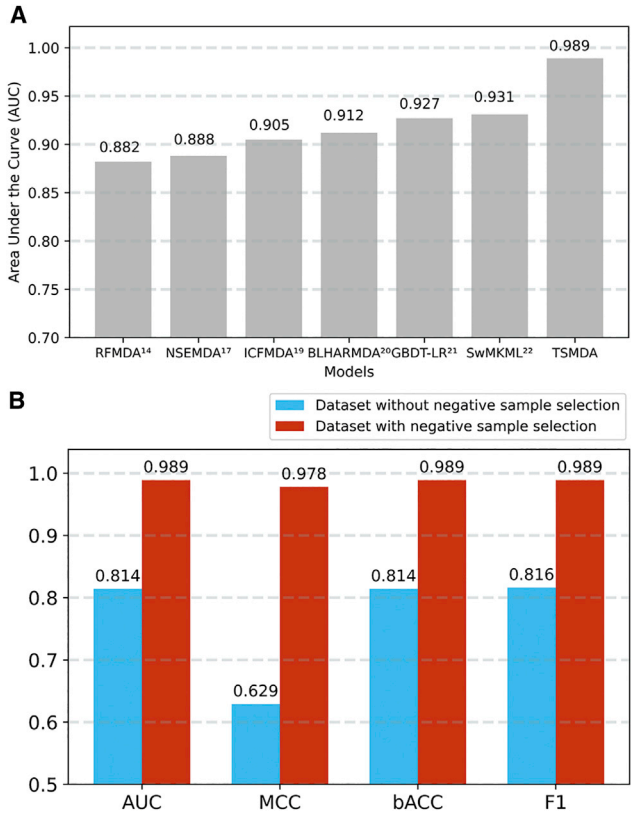


**Figure 2. Predictive performance of TSMDA**

(A) TSMDA considerably outperformed six recent miRNA-disease-association predictive models in terms of area under the curve (AUC). (B) Two negative sample selections, a miRNA-gene-disease network and modified PU learning, substantially enhance the performance of TSMDA. AUC, Matthews correlation coefficient (MCC), balanced accuracy (bACC), and F1 of TSMDA model with and without negative sample were assessed in 5-fold cross-validation with an extreme gradient boosting (XGBoost) classifier.

### Performance of TSMDA

We started by assessing the ability of TSMDA to predict miRNA-disease associations using The Human microRNA Disease Database (HMDD) v.2.0 database,[10] assessed under different cross-validation schemes. Under 5-fold cross-validation, our model achieved an AUC of 0.989, as well as Matthews correlation coefficient (MCC), balanced accuracy (bACC), and F1 scores of 0.978, 0.989, and 0.989, respectively (Table 2). The method obtained comparable outcomes from 10-fold and 20-fold cross-validation, further demonstrating the robustness of the TSMDA predictive model (Table 2). Taking a closer look at misclassified entries in a blind test and cross-validation, we noticed that the majority are false negatives. The investigation exhibits that 27 out of 31 entries in the blind test are false negatives. However, no particular miRNA or disease is found predominantly. We further examined the contribution of each feature to misclassified predictions in a blind test with individual SHAP values (Table S1). Unsurprisingly, the result suggested the features with high feature importance, especially feature 4, tend to be the main contributors to a misclassification.

Diverse computational models have been proposed to fill the missing knowledge of miRNA-disease relationships during the past 10 years.[13–22] In this study, we compare the performance of TSMDA with six recent miRNA-disease-association predictors: RFMDA,[14] NSEMDA,[17] ICFMDA,[19] BLHARMDA,[20] GBDT-LR,[21] and SwMKML.[22] The selected methods are based on the same dataset, HMDD v.2.0, enabling an adequate comparison. As most methods are not publicly available for replication, only the AUC values reported in the original article were used for a comparison. As a result, our model considerably outperformed all six recent predictive models (Figure 2A).

We believe one of the reasons behind the performance of TSMDA lies in the novel procedure to measure miRNA and disease similarity by considering target genes and symptoms, which directly reflect the biological nature of miRNAs and diseases. Moreover, unlike previous research that randomly selected negative samples from unknown associations,[14,16,21] TSMDA utilizes a miRNA-gene-disease network, followed by a modified PU learning, to construct more reliable negative samples (Figure 2B).

### Blind test

To evaluate the generalization capabilities of TSMDA, we assessed its performance on an independent blind test of experimentally validated miRNA-disease associations from HMDD, providing an unbiased evaluation of the trained model. The model reached an AUC, MCC, bACC, and F1 of 0.982, 0.965, 0.982, and 0.982, respectively, which were consistent with the performance obtained under cross-validation (Table 2).

### Predicting miRNA-disease associations in cancer

Three case studies involving prevalent cancer types (breast, prostate, and lung cancer) were employed to evaluate the capability of TSMDA of predicting potential miRNA-disease associations in a real-world scenario.

The statistics reported in the 2020 annual report of the American Cancer Society show that these cancers are among the top five cancers with the highest estimated new cases and deaths in the US population.[40] Breast cancer is widely known as the most prevalent cancer in females, accounting for 30% of the cases.[40] Similarly, prostate cancer is the most commonly found male cancer, responsible for one-fifth of the cases, while lung cancer is the second most common type of cancer in both genders.[40]

In the first case study, the general predictive performance of TSMDA was assessed by its ability to identify the breast, prostate, and lung cancer-related miRNAs for experimentally validated associations in dbDEMC and miRCancer.[11,12] Known associations in HMDD v.2.0 were chosen as a training dataset. The top 50 cancer-related miRNAs were ranked based on TSMDA scores and listed in Tables S2–S4. Using TSMDA scores, 49, 50, and 50 of the predicted miRNAs associated with breast, prostate, and lung cancer, respectively, were experimentally confirmed by other databases.

The ability of TSMDA to predict potential associations for diseases without verified associated miRNAs was evaluated in the second case study. Known associations between the three cancer types and miRNAs in the training set of HMDD v.2.0 were removed, one cancer at a time. As a result, 49, 49, and 49 of the top 50 were validated with known associations in dbDEMC and miR2Cancer (Tables S5–S7).[11,12]

In the third case study, miR2Disease containing 3,273 known associations between 349 miRNAs and 163 diseases was used to demonstrate our model performance on different datasets.[41] miR2Disease was used to train the model, and the top 50 potential associated miRNAs predicted were investigated in dbDEMC and miR2Cancer (Tables S8–S10).[11,12] All associations were confirmed, indicating the robustness of TSMDA to uncover potential miRNA-disease associations when considering different datasets.

### TSMDA web server

We have made TSMDA available as an easy-to-use web server. the TSMDA web server works according to the following procedures. First, users are required to manually provide a list of miRNAs in miRBase format and a list of disease Medical Subject Heading (MeSH) IDs. This list can be provided as a file. Users also have the possibility to fill a single string for either miRNA or MeSH ID. The example can be downloaded in the TSMDA server (Figure 3A). After running TSMDA, prediction results will be provided as a table, which can be downloaded as a comma-separated file. For each pair of miRNA and disease, an association confidence is shown. A higher score indicates a higher potential of association between miRNA and disease. Moreover, related evidence is given as a PMID for a pair of miRNA and disease with existing experimental support in Mammalian ncRNA-Disease Repository (MNDR) or dbDEMC.[11,42] The TSMDA web server is available at http://biosig.unimelb.edu.au/tsmda/.

## DISCUSSION

The utilization of miRNAs as diagnostic biomarkers or drugs has received growing attention,[7–9] due to their significant regulatory roles in various physiological processes. To enable the development of miRNA-based therapeutic applications, a wide range of studies has validated a large number of relationships between miRNAs and disease, which have provided a better understanding of miRNA regulatory mechanisms.[3–5] A significant proportion of potential miRNA-disease associations are yet to be explored, and computational methods play an essential role in assisting on this task.

The proposed TSMDA prediction model has led to three major improvements for miRNA-disease-association prediction in terms of (1) miRNA similarity calculation, (2) disease similarity calculation, and (3) negative sample selection strategies. First, an approach for miRNA similarity calculation called target-based miRNA similarity was introduced. Unlike sequence or associated-disease information used in many previous methods,[13–22] individual miRNAs' target genes directly reflect their unique function in molecular pathways. TSDMA has shown that by combining this method with MISIM miRNA functional similarity, they can help improve the model's prediction power and reliability (Figure 4). Second, the symptom-based approach was utilized to calculate disease similarity. Several studies indicated the remarkable predictive capability of symptom-based similarity as it is associated with several molecular mechanisms,[43–45] including shared genes, protein interactions, and molecular origins. Finally, we designed modern negative sample selection approaches on TSMDA. A lack of actual negative samples has been a limitation of miRNA-disease-association studies for an extended period. In this work, two reliable methods proposed in previous research, miRNA-gene-disease network[18] and traditional PU learning,[17,25,26] were adopted and modified. A more comprehensive network was obtained in comparison with previous methods by integrating two datasets from miRTarbase and Tarbase.[46,47] The modified PU learning approach was introduced to relieve the strong dependence on the chosen criteria of selecting reliable negative samples in the original method.[48]

To verify the performance of TSMDA, the method was assessed under different cross-validation schemes, as well as through an independent

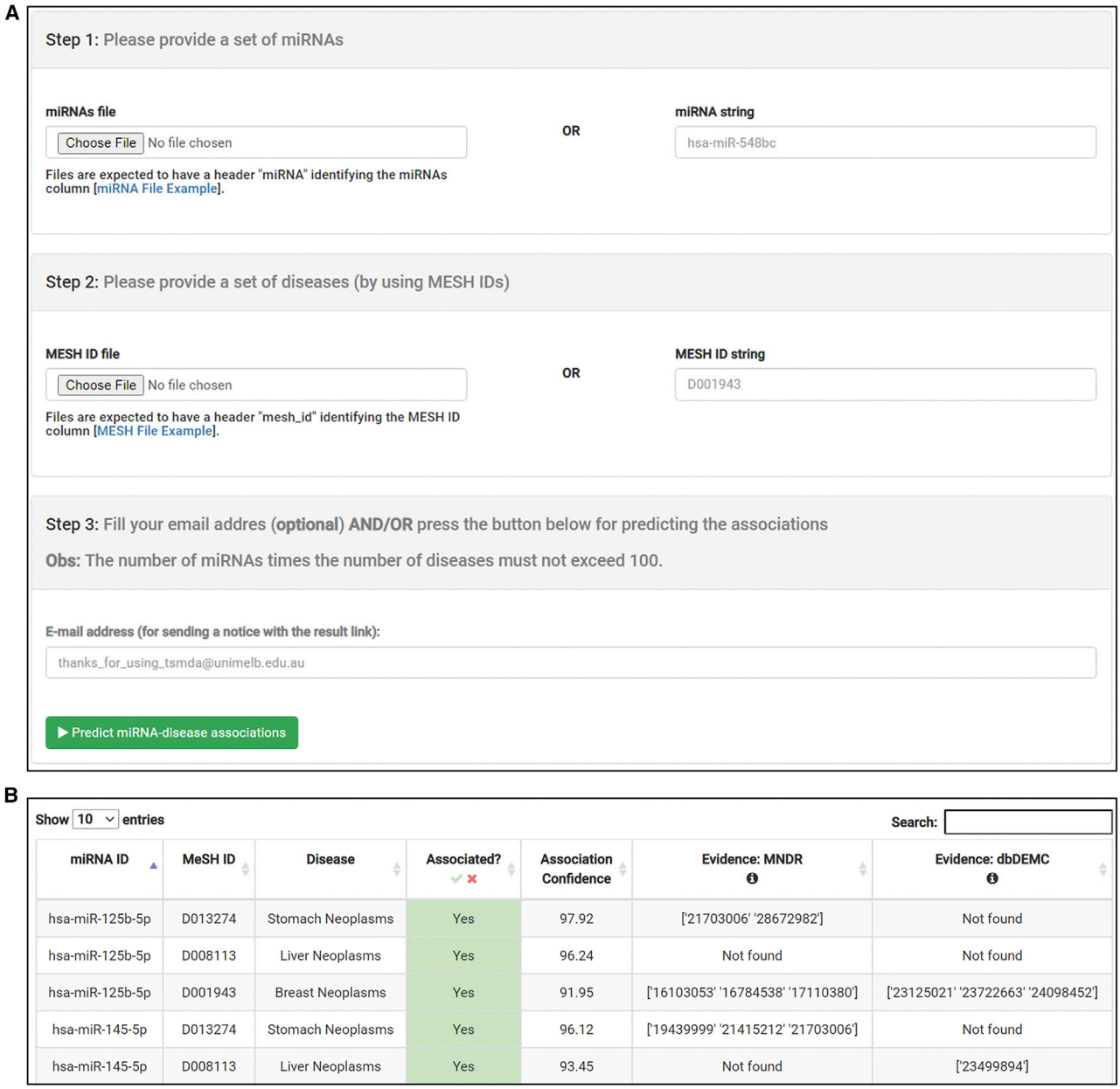


**Figure 3. The TSMDA web server interface**

(A) A list of miRNAs in miRBase IDs and diseases in MeSH IDs are required as input for the TSMDA web server. (B) The result from TSMDA is provided as a table. A higher prediction score indicates a higher probability for miRNA-disease association. If a miRNA-disease association is experimentally supported by MNDR[31] or dbDEMC,[11] evidence is provided as a PMID.

blind test and three case studies. The performance levels and consistency under different validation scenarios illustrate the robustness of the method in prioritizing potential miRNA-disease associations. Furthermore, we showed TSMDA has outperformed alternative state-of-the-art methods (Figure 2A),[14,17,19–22] indicating a substantial improvement from previous efforts. The model's reliability in a real-world application was supported by the case studies on the three common cancer types. To facilitate access to the method's capabilities and enable reproducibility, we developed a user-friendly web server to allow easy access by other researchers.

In future works, miRNA-disease-association predictions might be improved in many directions. One of the limitations of the current model is the bias in data availability. A significant proportion of

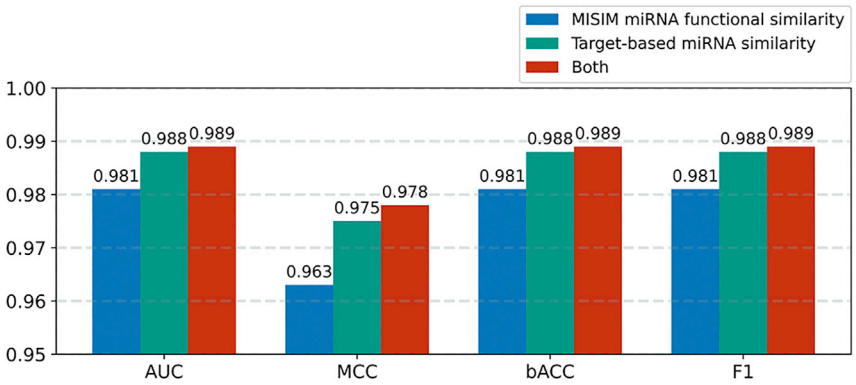


**Figure 4. The introduction of miRNA functional similarity (MISIM) with target-based miRNA similarity moderately enhances TSMDA performance**

AUC, MCC, bACC, and F1 in TSMDA models with three sets of features—3 target-based miRNA similarities (T) with 5 symptom-based similarities, 5 MISIM similarities (M) with 5 symptom-based similarities, and 8 target-based and MISIM miRNA similarities (T + M) with 5 symptom-based similarities—were assessed in 5-fold cross-validation with XGBoost classifier.

experimentally validated miRNA-disease associations as well as miRNA-target gene interactions has not been confirmed. Although TSMDA has attempted to overcome this bias by introducing a unique weighting scheme, more informative data sources, such as miRNA expression profiles, should be taken into consideration. On the other hand, other molecular properties of diseases, such as related biochemical pathways, could be introduced to enhance predictive accuracy. However, the disease similarity estimation is restrained by the limitation of HMDD v.2.0, where some diseases are not found in the Disease Ontology,[49] a standardized ontology for human diseases generally used for diverse disease similarity calculations.[50,51]

Data quality is a significant hurdle in determining the success of miRNA-disease-association prediction models. As future work, a practical method that utilizes other biological information to guide a reliable negative sample selection may be proposed to increase the model effectiveness. Furthermore, miRNA expression profiles retrieved from public databases, such as The Cancer Genome Atlas, can be utilized to improve data quality. Removing confirmed miRNA-disease associations with low confidence according to differential expression analysis may significantly improve data reliability.

## MATERIALS AND METHODS

### TSMDA general workflow

The proposed pipeline consists of five main steps (Figure 5). First, confirmed miRNA-disease associations were obtained from HMDD v.2.0.[10] In the following step, feature engineering is performed and three sets of similarities constructed: MISIM,[23] target-based miRNA similarity, and symptom-based disease similarity. These were integrated into feature vectors, representing pairs of miRNA-disease associations. Subsequently, reliable negative samples were selected using miRNA-gene-disease network and modified PU learning. Following that, a subset of relevant features is chosen by correlation-based and forward stepwise greedy feature selection.[27,28] An extreme gradient boosting classifier (XGBoost) was employed to create a prediction model for potential associations. The method's performance was assessed using both internal (5-fold, 10-fold, and 20-fold cross-validation) and external validation (blind test and three case studies).[52]

### Data collection: Human miRNA-disease associations

Experimentally validated human miRNA-disease associations were retrieved from HMDD v.2.0.[10] The dataset contains 5,430 associations between 495 miRNAs and 383 diseases. Given this dataset, a vector $V$ was built to describe the associations between miRNA and disease as follows:

$$V = \left(A_{i,j}, A_{i+1,j}, A_{i+2,j}, \ldots, A_{M\times D}\right), \quad \text{(Equation 1)}$$

where $M$ and $D$ are the number of miRNAs and diseases in HMDD v.2.0, respectively, and $A_{i,j}$ is equal to one (1) if miRNA $i$ and disease $j$ are experimentally associated, and zero (0), otherwise.

### miRNA functional similarity

The MISIM used in this research was proposed by Wang et al.[23] due to its relative simplicity and decent capability to represent miRNA similarity in a number of studies.[14–22] The data of known miRNA-disease associations was utilized to assess miRNA similarity based on the assumption that miRNAs with similar functions are more likely to be associated with pathologically similar diseases. We retrieved miRNA functional similarity of miRNAs found in HMDD v.2.0 from the Cui Lab repository. The miRNA functional similarity matrix (MFS) describing the pairwise similarities among 495 miRNAs was constructed.

### Target-based miRNA similarity

Despite a satisfactory contribution to miRNA-disease predictions, incomplete data of validated associations still limited the performance of MISIM. To address this limitation, other data types should be considered to enhance miRNA similarity representation and mitigate biases. Two modern methods, LFEMDA and DBMDA, proposed sequence-based approaches to estimate miRNA similarity. The improved accuracy indicated the usefulness of biological features.[15,16]

In this work, biological information of miRNA targets was introduced to determine miRNA similarity. miRNAs perform a regulatory function via complementary base pairing with several mRNAs. Thus, miRNAs with similar target genes are more likely to have similar functions in molecular pathways. Here, we utilized the numbers of shared target genes to assess miRNA similarity. The experimentally validated miRNA-target interactions were available at miRTarBase

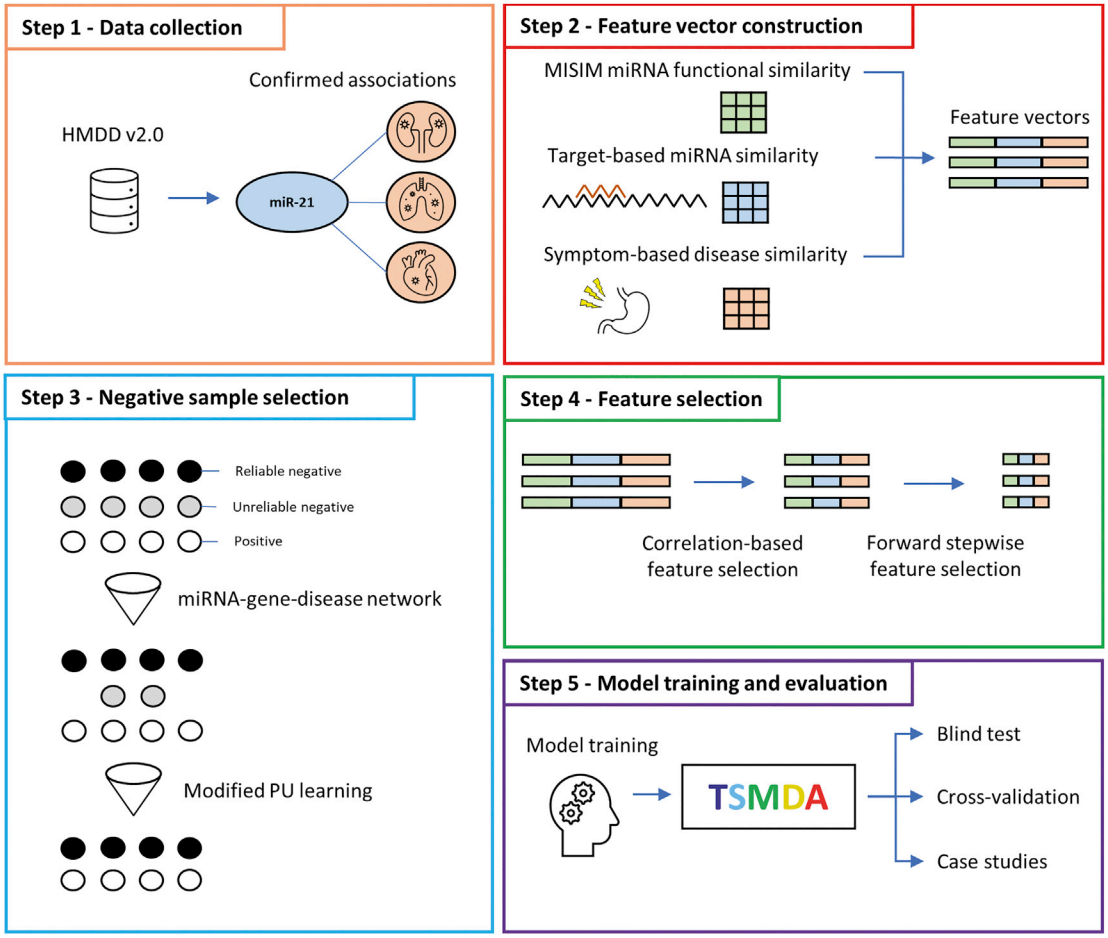


**Figure 5. TSMDA: Predicting miRNA-disease associations**

The development of TSMDA is divided into five steps: (1) data collection, (2) feature vector construction, (3) negative sample selection, (4) feature selection, and (5) model training and evaluation.

and TarBase.[46,47] miRTarBase consists of 553,168 interactions between 3,775 miRNAs and 22,336 target genes, whereas TarBase contains 422,614 interactions between 1,084 miRNAs and 20,790 target genes. The interactions related to miRNAs found in HMDD v.2.0 were extracted and merged, producing the dataset of 397,402 interactions between 489 miRNAs and 21,284 genes. Across all 495 miRNAs in the HMDD v.2.0, six missing miRNAs were proved by miRBase to be experimental errors.[53]

The information of shared target genes between miRNAs was utilized to calculate miRNA similarity. The 21,284-dimensional vector $M$ described target genes for miRNA $i$ was created as:

$$M_i = \left(s_{i,1}, s_{i,2}, s_{i,3}, \ldots, s_{i,j}\right), \quad \text{(Equation 2)}$$

where $s_{i,j}$ denotes the strength of the interaction between miRNA $i$ and target gene $j$. It is calculated by taking the prevalence of target genes in the dataset into consideration. The strength of interaction between a pair of miRNA $i$ and target gene $j$ is equal to log 2 of term frequency of target gene if they are interacting, otherwise equal to zero as follows:

$$s_{i,j} = \begin{cases} log_2 F_j & M_i \text{ and } T_j \text{ are interacting} \\ 0 & otherwise \end{cases}. \quad \text{(Equation 3)}$$

In the equation, $F_j$ is a term frequency of a target gene. $M_i$ and $T_j$ refer to miRNA $i$ and target gene $j$.

In the end, cosine similarity was employed to assess the target-based miRNA similarity between the arrays representing the miRNAs.[54] Cosine similarity is a standard metric used to compute the directional similarity between two vectors by capturing orientational differences. The advantage of the cosine similarity is the computation irrespective of vectors' sizes. miRNA similarity was calculated as stored in a target-based miRNA similarity matrix ($TMS$).

### Symptom-based disease similarity

Several studies demonstrated a close correspondence between the resemblance of molecular pathogenesis (e.g., shared gene, protein-protein interactions, and molecular origin) and the phenotypic similarity in clinical symptoms.[55,56] On this basis, Zhou et al.[43] proposed the novel symptom-based disease similarity calculation that can be applied to create a phenotype network profile for discovering molecular targets for drug repurposing.[44,45] This approach has displayed a robust correlation between calculated similarity and molecular-level disease components. The unique advantage of this method is a wide availability of directly observable clinical phenotypes in various diseases. For this reason, TSMDA aimed to implement a symptom-based approach to measure disease similarity.

The co-occurrences of diseases and symptoms in PubMed were used to characterize each disease in terms of clinical phenotypes. First, the 383 diseases from HMDD v.2.0 were mapped to 328 MeSH identifiers.[57] For each disease, its MeSH ID was used as a query to search for co-occurrences with 481 symptoms (2020th updated), categorized by PubMed. Disease $i$ can be described by a 481-dimensional vector as follows:

$$D_i = (w_{i,1}, w_{i,2}, w_{i,3}, \ldots, w_{i,481}). \quad \text{(Equation 4)}$$

$w_{i,j}$ quantifies the intensity of the co-occurrence between disease $i$ and symptom $j$. According to the bias where some symptoms such as pain are comparatively more abundant, the intensity was estimated considering the term frequency-inverse document frequency (TF-IDF).[43] It is calculated from absolute co-occurrence $W_{i,j}$ as the following equation:

$$w_{i,j} = W_{i,j}\, log\frac{N}{n_j}, \quad \text{(Equation 5)}$$

where $N$ denotes the number of diseases in HMDD v.2.0, while $n_j$ represents the number of diseases where symptom $j$ appears. Same as target-based miRNA similarity, the cosine similarity was also employed to measure the directional similarity between symptom-described vectors for each disease.[54] The symptom-based disease similarity among 495 diseases was represented as a symptom-based disease similarity matrix ($SDS$).

### miRNA and disease similarity integration

We obtained 1,373-dimensional feature vectors describing 189,585 possible pairs of miRNAs and diseases in HMDD v.2.0 from the integration of MISIM miRNA functional similarity, target-based miRNA similarity, and symptom-based disease similarity. The feature vectors $F_{i,j}$ representing miRNA $i$ and disease $j$ were constructed as follows:

$$F_{i,j} = \left(mms_{i,1}, \ldots, mms_{i,nM}, tms_{i,1}, \ldots, tms_{i,nM}, sds_{j,1}, \ldots, sds_{j,nD}\right). \quad \text{(Equation 6)}$$

Here, $mms_{i,m}$ and $tms_{i.m}$ denote MISIM and target-based miRNA similarity between miRNA $i$ and miRNA $m$, whereas $sds_{j,d}$ is the symptom-based disease similarity between disease $j$and disease $d$. $nM$and $nD$ are numbers of miRNAs and diseases in HMDD v.2.0.

### Negative sample selection

Negative sample selection is undeniably one of the most crucial processes in miRNA-disease-association modeling due to the absence of true negative samples in the database. A variety of negative sample selection strategies have been explored to address this issue.

The general standard procedure is to obtain negative samples by a random selection from unlabeled miRNA-disease associations.[14,16,21] This approach expects the ideal situation where unconfirmed pairs can be arbitrarily considered as not existing, which may not be valid, negatively affecting the reliability of negative samples. NSEMDA[17] has proposed alternative strategies that utilize a traditional PU learning model[25,26] to train the model and remove unreliable negative samples iteratively. In contrast, NMLPMDA suggested a distinct method that focused on the construction of a miRNA-gene-disease network.[18] Pairs of miRNA and disease that show no relationship were selected as reliable negative samples. The remarkable accuracy of these methods illustrates the potential to prioritize reliable negative samples. However, there is still room for improvement.

TSMDA employed a miRNA-gene-disease network, followed by modified PU learning to form a robust negative sample selection. The methods were further improved by extending the size of the network and replacing the original PU learning with a modified algorithm. In details, 115,891,964 verified gene-disease associations between 21,671 genes and 30,170 diseases were acquired from DisGENET v.7.0.[58] They were integrated with the aforementioned miRNA-target gene interactions from miRTarbase[46] and Tarbase,[47] forming the miRNA-gene-disease network. Pairs of miRNA and disease sharing the same gene in the network were considered as potential miRNA-disease associations. Unknown associations in our dataset were then mapped to the network to filter out the potential associations. From 184,155 unknown associations, only 20,716 associations (~10%) are selected as promising negative samples.

To increasingly refine the negative samples, modified PU learning[48] employing an iterative pruning strategy was introduced. It was initially proposed to mitigate the heavy dependence on the chosen criteria of reliable negative sample selection,[48] resulting in more reliable negative samples. In this work, 20% of known associations in HMDD v.2.0 were separated from the dataset and used as positive samples in PU learning to prevent overfitting from a bias toward a dataset, while the remaining negative samples were negative samples. Random forest (RF) classifier[59] was selected to train a model in an iterative manner because of the robustness to overfitting and less requirement for parameter tuning. Negative samples with low confidence scores were removed in each turn, otherwise retained in the dataset.

During the first loop, the RF classifier was trained to remove a large proportion of negative samples that were highly likely to be positive samples. Merely 1% of negative samples classified as positives or negatives, but with a probability lower than 95%, they were eliminated. Due to this strict condition, the remaining negative samples will be comparatively more reliable and suitable for training subsequent

models. In the following loops, we aimed for a slight reduction of negative samples in each loop. An RF classifier was similarly implemented; however, the hyperparameter was set in order to limit the model complexity, allowing iterative pruning. The numbers of estimators and maximum depth were reduced to 20 and 3. Only negative samples classified as positives were removed each step. The process was run until the number of reliable samples was the same as known associations.

### Feature selection

After the negative sample selection, feature selection was used to define a better set of features, so redundancy and noise are removed or diminished, computation time and model complexity are reduced, and overfitting is less likely to happen.[52] In several miRNA-disease-association models, employing a proper feature selection technique leads to a substantially increased predictive performance.[60–62] TSMDA utilizes two feature selection means, a correlation-based[27] and forward stepwise greedy feature selection.[28,63–65]

Initially, Pearson's correlation coefficients (PCCs) between every pair of features were calculated and represented as a heatmap in Figure S1. It was apparent that multiple features are redundant, so some can be discarded without reducing model accuracy. We conducted a performance evaluation to examine the optimal cutoff for PCC values (Figure S2). As a result, the cutoff of 0.6 was selected. If a PCC between features is higher than 0.6, only one feature is randomly retained. Consequently, the number of features was drastically reduced from 1,373 to 97.

Forward stepwise greedy feature selection was used to scale down the remaining dimensions by selecting the best combination of features.[28] The process begins with zero features selected. The most useful feature contributing the most to the performance was included one at a time. In each step, 10-fold cross-validation with XGBoost[29] was performed, then evaluated with MCC (Figure S3). At the end, 13 features (Table 1) were chosen as the best combination required to train a highly accurate model. The subset of features contained five miRNA functional similarities, three target-based miRNA similarities, and five symptom-based disease similarities.

### XGBoost classifier

XGBoost[29] is one of the most widely used tree-based boosting algorithms, where a set of weak classifiers are combined to form a strong classifier sequentially. In each iteration, misclassification errors of a previous classifier were corrected to create a more accurate model. In contrast to other boosting algorithms, XGBoost has several enhancements in regularization, parallelization, handling missing values, dropout methods, and others.

In this work, this algorithm has been shown to be the one with best performances in terms of miRNA-disease-association predictions in preliminary experiments (see Table S11). The final feature vectors represented by the selected 13 features are adopted to train and validate the XGBoost classification model.

### Availability of data and materials

The datasets used in this work are available at http://biosig.unimelb.edu.au/tsmda/data.

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2021.08.016.

## ACKNOWLEDGMENTS

K.U. was supported by the Melbourne Research Scholarship. A.G.C.d.S. acknowledges the Joe White Bequest Fellowship for its support. D.B.A. and D.E.V.P. were funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MR/M026302/1). D.B.A. was supported by the Wellcome Trust (grant 093167/Z/10/Z), the Jack Brockhoff Foundation (JBF 4186, 2016), and an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405). Supported in part by the Victorian Government's Operational Infrastructure Support Program.

## AUTHOR CONTRIBUTIONS

K.U. prepared the dataset, designed and conducted the experiment, and wrote the manuscript with support and advice from A.G.C.d.S., A.A., D.E.V.P., and D.B.A. The web server was designed and established by A.G.C.d.S. The project was conceived, designed, and supervised by D.B.A. All the authors read and approved the final manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## REFERENCES

1. Wahid, F., Shehzad, A., Khan, T., and Kim, Y.Y. (2010). MicroRNAs: synthesis, mechanism, function, and recent clinical trials. Biochim. Biophys. Acta *1803*, 1231–1243.

2. Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. (2005). Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. Cell *122*, 553–563.

3. Deng, S., Calin, G.A., Croce, C.M., Coukos, G., and Zhang, L. (2008). Mechanisms of microRNA deregulation in human cancer. Cell Cycle *7*, 2643–2646.

4. Gurha, P. (2016). MicroRNAs in cardiovascular disease. Curr. Opin. Cardiol. *31*, 249–254.

5. Xu, B., Hsu, P.K., Karayiorgou, M., and Gogos, J.A. (2012). MicroRNA dysregulation in neuropsychiatric disorders and cognitive dysfunction. Neurobiol. Dis. *46*, 291–301.

6. Kochman, M. (2007). MicroRNA Expression Patterns to Differentiate Pancreatic Adenocarcinoma From Normal Pancreas and Chronic Pancreatitis. Yearbook of Gastroenterology *2007*, 63–64.

7. Schwarzenbach, H., Milde-Langosch, K., Steinbach, B., Müller, V., and Pantel, K. (2012). Diagnostic potential of PTEN-targeting miR-214 in the blood of breast cancer patients. Breast Cancer Res. Treat. *134*, 933–941.

8. Mar-Aguilar, F., Mendoza-Ramírez, J.A., Malagón-Santiago, I., Espino-Silva, P.K., Santuario-Facio, S.K., Ruiz-Flores, P., Rodríguez-Padilla, C., and Reséndez-Pérez, D. (2013). Serum circulating microRNA profiling for identification of potential breast cancer biomarkers. Dis. Markers *34*, 163–169.

9. Rupaimoole, R., and Slack, F.J. (2017). MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. Nat. Rev. Drug Discov. *16*, 203–222.

10. Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Res. *42* (D1), D1070–D1074.

11. Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., and Teschendorff, A.E. (2017). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. Nucleic Acids Res. *45* (D1), D812–D818.

12. Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics *29*, 638–644.

13. Xu, J., Li, C.X., Lv, J.Y., Li, Y.S., Xiao, Y., Shao, T.T., Huo, X., Li, X., Zou, Y., Han, Q.L., et al. (2011). Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. Mol. Cancer Ther. *10*, 1857–1866.

14. Chen, X., Wang, C.C., Yin, J., and You, Z.H. (2018). Novel Human miRNA-Disease Association Inference Based on Random Forest. Mol. Ther. Nucleic Acids *13*, 568–579.

15. Che, K., Guo, M., Wang, C., Liu, X., and Chen, X. (2019). Predicting MiRNA-Disease Association by Latent Feature Extraction with Positive Samples. Genes (Basel) *10*, 80.

16. Zheng, K., You, Z.H., Wang, L., Zhou, Y., Li, L.P., and Li, Z.W. (2020). DBMDA: A Unified Embedding for Sequence-Based miRNA Similarity Measure with Applications to Predict and Validate miRNA-Disease Associations. Mol. Ther. Nucleic Acids *19*, 602–611.

17. Wang, C.C., Chen, X., Yin, J., and Qu, J. (2019). An integrated framework for the identification of potential miRNA-disease association based on novel negative samples extraction strategy. RNA Biol. *16*, 257–269.

18. Li, N., Duan, G., Yan, C., Wu, F.X., and Wang, J. (2020). MiRNA-Disease Associations Prediction Based on Negative Sample Selection and Multi-layer Perceptron. In Bioinformatics Research and Applications. ISBRA 2020, Volume 12304, Z. Cai, I. Mandoiu, G. Narasimhan, P. Skums, and X. Guo, eds., Lecture Notes in Computer Science (Cham: Springer).

19. Jiang, Y., Liu, B., Yu, L., Yan, C., and Bian, H. (2018). Predict MiRNA-Disease Association with Collaborative Filtering. Neuroinformatics *16*, 363–372.

20. Chen, X., Cheng, J.Y., and Yin, J. (2018). Predicting microRNA-disease associations using bipartite local models and hubness-aware regression. RNA Biol. *15*, 1192–1205.

21. Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020). Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. Comput. Biol. Chem. *85*, 107200.

22. Pan, Z., Zhang, H., Liang, C., Li, G., Xiao, Q., Ding, P., and Luo, J. (2019). Self-Weighted Multi-Kernel Multi-Label Learning for Potential miRNA-Disease Association Prediction. Mol. Ther. Nucleic Acids *17*, 414–423.

23. Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics *26*, 1644–1650.

24. Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., Teng, Z., et al. (2013). Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors. PLoS ONE *8*, e70204.

25. Liu, B., Lee, W.S., Yu, P.S., Heights, Y., and Li, X. (1998). Partially Supervised Classification of Text Documents, https://www.cs.uic.edu/~liub/S-EM/unlabelled.pdf.

26. Rochio, J.J. (1971). Relevant feedback in information retrieval. In The Smart Retrieval System: Experiments in Automatic Document Processing, G. Salton, ed. (Englewood Cliffs, NJ: Prentice Hall Inc.), pp. 313–323.

27. Hall, M.A. (2000). Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Proceedings of the Seventeenth International Conference on Machine Learning, pp. 359–366.

28. Deng, X., Li, Y., Weng, J., and Zhang, J. (2019). Feature selection for text classification: A review. Multimedia Tools Appl. *78*, 3797–3816.

29. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

30. Lundberg, S.M., and Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. In NeurIPS, *30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.).

31. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

32. Erik, S., and Igor, K. (2014). Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. *41*, 647–665.

33. Schmidhuber, J. (2015). Deep learning in neural networks: an overview. Neural Netw. *61*, 85–117.

34. Iliopoulos, D., Bimpaki, E.I., Nesterova, M., and Stratakis, C.A. (2009). MicroRNA signature of primary pigmented nodular adrenocortical disease: clinical correlations and regulation of Wnt signaling. Cancer Res. *69*, 3278–3282.

35. Li, L.J., Huang, Q., Zhang, N., Wang, G.B., and Liu, Y.H. (2014). miR-376b-5p regulates angiogenesis in cerebral ischemia. Mol. Med. Rep. *10*, 527–535.

36. Liu, R., Ma, X., Xu, L., Wang, D., Jiang, X., Zhu, W., Cui, B., Ning, G., Lin, D., and Wang, S. (2012). Differential microRNA expression in peripheral blood mononuclear cells from Graves' disease patients. J. Clin. Endocrinol. Metab. *97*, E968–E972.

37. Pan, Z., Guo, Y., Qi, H., Fan, K., Wang, S., Zhao, H., Fan, Y., Xie, J., Guo, F., Hou, Y., et al. (2012). M3 subtype of muscarinic acetylcholine receptor promotes cardioprotection via the suppression of miR-376b-5p. PLoS ONE *7*, e32571.

38. Vargas-Medrano, J., Yang, B., Garza, N.T., Segura-Ulate, I., and Perez, R.G. (2019). Up-regulation of protective neuronal MicroRNAs by FTY720 and novel FTY720-derivatives. Neurosci. Lett. *690*, 178–180.

39. Nam, R.K., Wallis, C.J.D., Amemiya, Y., Benatar, T., and Seth, A. (2018). Identification of a novel MicroRNA panel associated with metastasis following radical prostatectomy for prostate cancer. Anticancer Res. *38*, 5027–5034.

40. American Cancer Society (2020). Cancer Facts & Figures 2020, https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf.

41. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. *37* (*Suppl 1*), D98–D104.

42. Ning, L., Cui, T., Zheng, B., Wang, N., Luo, J., Yang, B., Du, M., Cheng, J., Dou, Y., and Wang, D. (2021). MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acids Res. *49* (D1), D160–D164.

43. Zhou, X., Menche, J., Barabási, A.L., and Sharma, A. (2014). Human symptoms-disease network. Nat. Commun. *5*, 4212.

44. Casas, A.I., Hassan, A.A., Larsen, S.J., Gomez-Rangel, V., Elbatreek, M., Kleikers, P.W.M., Guney, E., Egea, J., López, M.G., Baumbach, J., and Schmidt, H.H.H.W. (2019). From single drug targets to synergistic network pharmacology in ischemic stroke. Proc. Natl. Acad. Sci. USA *116*, 7129–7136.

45. Cheng, F., Desai, R.J., Handy, D.E., Wang, R., Schneeweiss, S., Barabási, A.L., and Loscalzo, J. (2018). Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat. Commun. *9*, 2691.

46. Huang, H.Y., Lin, Y.C.D., Li, J., Huang, K.Y., Shrestha, S., Hong, H.C., Tang, Y., Chen, Y.G., Jin, C.N., Yu, Y., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res. *48* (D1), D148–D154.

47. Karagkouni, D., Paraskevopoulou, M.D., Chatzopoulos, S., Vlachos, I.S., Tastsoglou, S., Kanellos, I., Papadimitriou, D., Kavakiotis, I., Maniou, S., Skoufos, G., et al. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic Acids Res. *46* (D1), D239–D245.

48. Hernández Fusilier, D., Montes-y-Gómez, M., Rosso, P., and Guzmán Cabrera, R. (2015). Detecting positive and negative deceptive opinions using PU-learning. Inf. Process. Manage. *51*, 433–443.

49. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. *40* (D1), D940–D946.

50. Yu, G., Wang, L.G., Yan, G.R., and He, Q.Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics *31*, 608–609.

51. Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., Li, C., Li, X., Rao, S., and Li, X. (2011). DOSim: an R package for similarity between diseases based on Disease Ontology. BMC Bioinformatics *12*, 266.

52. Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, Second Edition (Sebastopol, CA: O'Reilly Media).

53. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47* (D1), D155–D162.

54. Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull. *24*, 35–43.

55. Freudenberg, J., and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics *18* (*Suppl 2*), S110–S115.

56. Wang, Q., Liu, W., Ning, S., Ye, J., Huang, T., Li, Y., Wang, P., Shi, H., and Li, X. (2012). Community of protein complexes impacts disease association. Eur. J. Hum. Genet. *20*, 1162–1167.

57. Lipscomb, C.E. (2000). Medical subject headings (MeSH). Bull. Med. Libr. Assoc. *88*, 265–266.

58. Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. *48* (D1), D845–D855.

59. Ho, T.K. (1995). Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, *1*, pp. 278–282.

60. Peng, J., Hui, W., Li, Q., Chen, B., Jiang, Q., Shang, X., and Wei, Z. (2018). A learning-based framework for miRNA-disease association identification using neural networks. bioRxiv. https://doi.org/10.1101/276048.

61. Yao, D., Zhan, X., and Kwoh, C.K. (2019). An improved random forest-based computational model for predicting novel miRNA-disease associations. BMC Bioinformatics *20*, 624.

62. Chen, X., Zhu, C.C., and Yin, J. (2019). Ensemble of decision tree reveals potential miRNA-disease associations. PLoS Comput. Biol. *15*, e1007209.

63. Rodrigues, C.H.M., Pires, D.E.V., and Ascher, D.B. (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci. *30*, 60–69.

64. Pires, D.E.V., and Ascher, D.B. (2020). mycoCSM: Using Graph-based signatures to Identify Safe Potent hits against mycobacteria. J. Chem. Inf. Model. *60*, 3450–3456.

65. Myung, Y., Pires, D.E.V., and Ascher, D.B. (2020). mmCSM-AB: guiding rational antibody engineering through multiple point mutations. Nucleic Acids Res. *48* (W1), W125–W131.