Contents lists available at ScienceDirect



International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac



MDbDMRP: A novel molecular descriptor-based computational model to identify drug-miRNA relationships



Amit Daroch^{a,b}, Rituraj Purohit^{c,*}

^a Structural Bioinformatics Lab, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, Palampur, HP 176061, India
^b The Himalayan Centre for High-throughput Computational Biology, (HiCHiCoB, A BIC supported by DBT, India), Palampur, HP 176061, India
^c Academy of Scientific and Innovative Research, Ghaziabad 201002, India

ARTICLE INFO

Keywords: miRNAs Molecular descriptors Computational model Machine learning Molecular docking Drug discovery

ABSTRACT

MicroRNAs (miRNAs) are important in gene expression regulation and many other biological processes and have emerged as promising therapeutic targets. Identifying potential drug-miRNA relationships is helpful in disease therapy and pharmaceutical engineering in medical research. However, accurately predicting these relationships remains a significant computational challenge. This study introduces MDbDMRP, a novel molecular descriptorsbased drug-miRNA relationship prediction computational model designed to address this challenge. MDbDMRP leverages the power of machine learning to predict new drug-miRNA associations and non-associations. The model achieves exceptional performance, exceeding an average score of 0.92 across various evaluation metrics, including accuracy, precision, recall, and F1-score. Furthermore, it demonstrates a remarkable ability to distinguish between positive and negative interactions, as evidenced by an outstanding AUC-ROC score of 0.9864. The results obtained from MDbDMRP were further validated through molecular docking, reinforcing its performance. These results position MDbDMRP as a valuable tool for researchers aiming to unlock the potential of miRNAs in drug discovery.

1. Introduction

A class of non-coding RNAs, microRNAs are small, 22 to 25 nucleotides long, involved in gene regulation and RNA silencing [1]. Since the discovery of the first miRNA in *Caenorhabditis elegans* by Lee et al. [2], a wealth of research has unveiled the crucial functions of miRNAs in orchestrating diverse physiological processes. These include regulating cell growth [3], differentiation [4], and death [5], orchestrating immune responses [6], and fine-tuning gene expression levels [7]. Moreover, mounting evidence demonstrates a strong link between the dysregulation of key miRNAs and the development of various complex human diseases. The pervasive influence of miRNAs across various biological processes, both normal and abnormal, makes them a captivating class of potential drug targets. While traditional experimental methods for miRNA analysis, such as miRNA sequence analysis (using high-throughput sequencing), real-time qPCR, and Northern blot, provide valuable insights, they come with significant drawbacks. These methods are often labor-intensive, time-consuming, and expensive due to the involvement of specialized equipment, reagents, and personnel. This raises the need for faster, more cost-effective alternatives for identifying potential small molecule-miRNA associations Throughout the arduous stages of drug development, computational models predicting drug-miRNA associations offer invaluable assistance. By illuminating potential interactions, these models can guide researchers toward the most efficacious drugs, significantly reducing the financial burden and uncertainty associated with extensive experimentation [8].

Notably, within this realm, predicting drug-miRNA associations represents a pivotal step in advancing drug research and development. With the high cost and time required for traditional experimental validation, efficient computational tools hold immense potential to accelerate the discovery of promising drug-miRNA associations. As research into predicting drug-miRNA associations intensifies, valuable resource repositories are emerging. Databases like ncRNADrug [9], SM2miR [10], NoncoRNA [11], mTD [12], and NRDTD [13] offer a wealth of information for exploring these interactions. This wealth of data fuels the development of increasingly accurate and effective drug-miRNA association prediction models. Researchers utilize diverse computational models to predict potential drug-miRNA interactions, primarily categorized into three approaches: biological network-based models, machine learning-based models, and other predictive methods [14].

* Corresponding author. *E-mail addresses:* rituraj@ihbt.res.in, riturajpurohit@gmail.com (R. Purohit).

https://doi.org/10.1016/j.ijbiomac.2024.138580

Received 9 October 2024; Received in revised form 20 November 2024; Accepted 7 December 2024 Available online 8 December 2024 0141-8130/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Table 1

Selected features used to train the MDbDMRP model.

Molecule	Module	Selected Features
Drug	Adjacency Matrix	SpMAD_A
Molecules	Autocorrelation	AATS8p, AATS2i, AATS4i, AATS6i, ATSC2s, ATSC3s, ATSC5z, ATSC5v, ATSC5p, ATSC6i, AATSC4c, AATSC7p, AATSC5i, MATS8c, MATS1d, GATS5c, GATS6c, GATS2d, GATS8se, GATS8se
	Burden-CAS-Unlike-Topological (BCUT)	BCUTc-1 l, BCUTs-1 h, BCUTp-1 h
	Charged Partial Sub Area (CPSA)	RPCS
	Electrotopological State (Estate)	MAXsCH3, MINdssC
	Molecular Representation of Structure based on	Mor24, Mor11m, Mor19m, Mor24v, Mor25v, Mor32v, Mor09se, Mor24se, Mor27se, Mor32se, Mor08p,
	Electron Diffraction (MORSE)	MOTI IP EState VSA2
	(MoeTypes)	Lotate_voit2
	MolecularDistanceEdge	MDEO-12
	TopologicalCharge	JGI3
miRNA	Composition Based Features	CDK_AC, CDK_AG, CDK_CG, CDK_GC, NRI_C, NRI_U, DDON_A, RDK_AA, RDK_AG, RDK_CC, RDK_GA, ENT NL A, ENT NL C, ENT NL G, PDNC AA, PDNC GA,
	Cross Correlation Based Features	DCC_p1_p10_lag1, DCC_p1_p13_lag1, DCC_p1_p17_lag1, DCC_p1_p20_lag1, DCC_p2_p1_lag1,
		DCC_p2_p13_lag1, DCC_p2_p14_lag1, DCC_p2_p16_lag1, DCC_p5_p12_lag1, DCC_p5_p21_lag1,
		DCC_p10_p11_lag1, DCC_p10_p17_lag1, DCC_p10_p21_lag1, DCC_p12_p6_lag1, DCC_p13_p1_lag1,
		DCC_p13_p5_lag1, DCC_p13_p17_lag1, DCC_p13_p22_lag1, DCC_p14_p2_lag1, DCC_p14_p21_lag1, DCC_p14_p21_la
		DCC p18 p16 lag1, DCC p18 p21 lag1, DCC p17 p10 lag1, DCC p17 p10 lag1, DCC p18 p16 lag1, DCC p20 p16 lag1.
		DCC p20 p18 lag1, DCC p20 p22 lag1, DCC p21 p2 lag1, DCC p21 p5 lag1, DCC p21 p10 lag1,
		DCC_p21_p11_lag1, DCC_p21_p13_lag1, DCC_p21_p18_lag1, DCC_p21_p20_lag1, DCC_p22_p1_lag1,
		DCC_p22_p10_lag1, DCC_p22_p12_lag1, DCC_p22_p13_lag1,DCC_p1_p15_lag2, DCC_p1_p17_lag2,
		DCC_p1_p18_lag2, DCC_p2_p5_lag2, DCC_p2_p13_lag2, DCC_p2_p15_lag2, DCC_p2_p19_lag2,
		DCC_p2_p22_lag2, DCC_p10_p11_lag2, DCC_p11_p17_lag2, DCC_p12_p10_lag2, DCC_p13_p6_lag2, DCC_p12_p12_lag2, DCC_p14_p20_lag2, DCC_p15_p12_lag2
		DCC_p15_p18_iag2, DCC_p15_p20_iag2, DCC_p15_p21_iag2, DCC_p14_p20_iag2, DCC_p15_p12_iag2, DCC_p15_p12_
		DCC p16 p17 lag2, DCC p17 p16 lag2, DCC p17 p20 lag2, DCC p17 p21 lag2, DCC p17 p22 lag2,
		DCC_p18_p14_lag2, DCC_p18_p21_lag2, DCC_p19_p2_lag2, DCC_p19_p13_lag2, DCC_p19_p23_lag2, DCC_p19_p20_lag2,
		DCC_p20_p2_lag2, DCC_p20_p18_lag2, DCC_p20_p21_lag2, DCC_p21_p2_lag2, DCC_p21_p6_lag2,
		DCC_p21_p12_lag2, DCC_p21_p18_lag2, DCC_p21_p20_lag2, DCC_p22_p2_lag2,
	Auto Cross Correlation Based Features	DACC_p13_lag1, DACC_p19_lag1, DACC_p21_lag1, DACC_p1_lag2, DACC_p2_lag2, DACC_p21_lag2,
		DACC $p_1 p_1 a_{g_1}$, DACC $p_2 p_1 a_{g_1}$, DACC $p_2 p_1 p_1 a_{g_1}$, DACC $p_2 p_2 p_2 a_{g_1}$, DACC $p_2 p_2 p_2 a_{g_1}$, DACC $p_1 p_1 p_1 a_{g_1}$, DACC $p_1 p_1 p_1 p_2 p_2 p_1 a_{g_1}$, DACC $p_1 p_1 p_1 p_2 p_2 p_2 p_1 a_{g_1}$, DACC $p_2 p_2 p_2 p_2 p_1 a_{g_1}$, DACC $p_1 p_1 p_2 p_2 p_2 p_2 p_2 p_2 p_2 p_2 p_2 p_2$
		DACC p13 p2 lag1, DACC p13 p10 lag1, DACC p13 p14 lag1, DACC p13 p16 lag1, DACC p14 p19 lag1,
		DACC_p15_p13_lag1, DACC_p15_p14_lag1, DACC_p16_p1_lag1, DACC_p16_p6_lag1, DACC_p17_p11_lag1,
		DACC_p17_p15_lag1, DACC_p19_p1_lag1, DACC_p19_p10_lag1, DACC_p19_p17_lag1, DACC_p20_p2_lag1,
		DACC_p20_p5_lag1, DACC_p20_p13_lag1, DACC_p20_p17_lag1, DACC_p20_p19_lag1, DACC_p21_p15_lag1,
		DACC_p21_p16_lag1, DACC_p22_p17_lag1, DACC_p22_p21_lag1, DACC_p1_p6_lag2, DACC_p1_p13_lag2,
		DACC_p1_p10_iag2, DACC_p1_p20_iag2, DACC_p2_p10_iag2, DACC_p2_p10_iag2, DACC_p2_p17_iag2, DACC_p10_iag2, DACC_p10_iag2, DACC_p10_iag2, DACC_p10_iag2, DACC_p10_iag2, DACC_p10_iag2, DACC_p10_iag2, DACC_p2_p17_iag2, DACC_p2_p17_ia
		DACC p10 p18 lag2, DACC p12 p16 lag2, DACC p14 p6 lag2, DACC p14 p16 lag2, DACC p15 p21 lag2.
		DACC_p16_p20_lag2, DACC_p17_p15_lag2, DACC_p18_p1_lag2, DACC_p18_p2_lag2, DACC_p18_p12_lag2,
		DACC_p18_p20_lag2, DACC_p19_p21_lag2, DACC_p20_p1_lag2, DACC_p20_p5_lag2, DACC_p21_p1_lag2,
		DACC_p21_p10_lag2, DACC_p21_p16_lag2
	Pseudo Correlation Based Features	PC PDNC GG

DAESTB [15] combines an autoencoder to extract hidden patterns with gradient-boosting trees to identify small molecule - miRNA pairs. EKRRSMMA [16], using feature reduction and ensemble learning, accurately predicts interactions between small molecules and micro-RNAs, potentially aiding drug development and disease treatment. The GISMMA [17] model uses network analysis and graphlet interactions to predict small molecule-miRNA associations, achieving high accuracy and potential benefits for disease therapy. Yichen Zhong et al. presented a multitask joint learning framework (MTJL) [18] by constructing a comprehensive similarity network for drugs and miRNAs, leveraging a graph autoencoder (GAE) to derive distinct embedding representations for both. DLP [19] uses subspace segmentation and DLSR algorithm to identify potential miRNA drug targets. PUDT [20] and HGCNMDA [21] are computational models designed to enhance drug discovery and disease understanding by predicting drug-target interactions and miR-NA-disease associations, respectively. PUDT improves interaction predictions by categorizing unknown samples, while HGCNMDA leverages a multi-relational network to capture complex miRNA-disease links. SMMARTs [22] employs graph regularization techniques within heterogeneous networks to discover associations between small molecules and microRNAs. GCFMCL [23] and GCNNMMA [24] leverage graph neural networks to model molecular structures and utilize contrastive learning or CNNs to enhance the prediction. The SMAJL [25] framework

employs a joint learning approach that uses a Restricted Boltzmann Machine to integrate variety of pharmacology, network information, structural and sequence information to predict association scores. MHCLMDA [26] and KBMF-MDI [27] are computational methods for predicting miRNA-disease associations to support disease understanding and treatment. MHCLMDA uses hypergraph contrastive learning to capture complex multi-view interactions, while KBMF-MDI employs Bayesian matrix factorization on miRNA and disease similarities, with both approaches outperforming existing models in accuracy for uncovering unknown associations. GNMFDMA [28] and DCMF [29] utilize matrix factorization methods to predict associations between drugs and miRNAs. SMANMF [30] aims to reveal unknown relationships between small molecules and miRNAs using non-negative matrix factorization. MFIDMA [31] and DMR-PEG [32] highlight the significance of capturing complex relationships in drug-miRNA networks through advanced neural architectures.

Here, we present a new methodology to predict drug-miRNA associations using machine-learning methods based on the molecular descriptors of drugs and miRNAs. Molecular descriptors bridge the gap between complex molecular structures and numerical data, enabling computational tools to decipher physicochemical properties, predict biological activities, and guide drug discovery. By capturing diverse structural information, tailored molecular descriptors empower the in-



Fig. 1. Workflow of MDbDMRP.



Fig. 2. Confusion Matrix displaying the model's prediction results.

silico modelling of molecules, accelerating research and unlocking hidden secrets in the realm of chemical and biological interactions. MDbDMRP serves as a powerful tool for accelerating drug discovery efforts by enabling the identification of potential drug repurposing opportunities and novel therapeutic targets. By predicting drug-miRNA interactions with high confidence, the model can help prioritize compounds for further experimental validation and clinical development. This opens exciting possibilities for designing targeted therapies that leverage the regulatory power of miRNAs, ultimately paving the way for more effective and personalized treatments.

2. Materials and methods

2.1. Dataset

We collected known drug-miRNA associations from the ncRNADrug database [9]. This initial dataset contained 13,503 interactions between 2122 miRNAs and 630 drugs. We then cleaned the data, removing missing or invalid entries (like dead miRNAs or drugs without 3D information). This resulted in a final set of 11,446 observations. To create this list of potential associations, we downloaded the sequence



Fig. 3. Violin Plot displaying density curves for each evaluation method.

information on all human miRNAs from miRBase [33]. We prepared the data for further analysis to explore potential new interactions and divided it into two parts. The first section lists interactions between miRNAs and drugs that have already been documented in research. The second section explores possibilities beyond currently known interactions. It includes all the remaining miRNAs (not found in the known associations) paired with each drug from the known associations using random sampling. To construct the negative samples, we followed the common approach of pairing drugs and miRNAs that are not known to be associated based on current biological knowledge and databases. Specifically, we generated random drug-miRNA pairs by ensuring that these pairs did not overlap with the known positive associations.

2.2. Descriptors computation

Molecular descriptors are numerical representations capturing various aspects of a molecule's structure for computational analysis. These mathematical values are used quantitatively to describe molecules' chemical and physical properties, which can be a valuable knowledge set for computational calculations. Mordred [34] is a Pythonbased tool for calculating molecular descriptors, designed to integrate seamlessly with RDKit. It offers over 1800 descriptors covering both 2D and 3D properties, making it highly versatile for cheminformatics and machine learning applications. Mordred's compatibility with python libraries like pandas and scikit-learn simplifies data processing and model building, while its scalability on cloud and high-performance platforms makes it ideal for handling large datasets efficiently. Nfeature [35] is a versatile package for nucleic acid analysis. It empowers users to explore the composition, distribution, and correlation patterns within nucleotide sequences, facilitating insightful biological interpretations. Nfeature calculates 14,385 features for RNA sequences and the same was used to calculate the miRNA features.

2.3. Data preprocessing and feature selection

Various techniques come into play in the quest to identify the most informative features of machine learning models. In the initial preprocessing, features with a high proportion of missing values were removed. The threshold was set at 0.5, meaning that if over 50 % of the values for a given feature were missing, that feature was excluded from the analysis. Features with zero variance were removed as they lack discriminatory power. Specifically, columns with constant values across all samples were identified and excluded from the dataset since they do not contribute meaningful information to the model's predictions. To address redundancy, pairwise correlation analysis was applied to identify and remove highly correlated features. The threshold for "high correlation" was set at 0.8. For pairs of features where the correlation exceeded this threshold, we compared each feature's correlation with the target variable. The feature with a lower correlation to the target was removed, retaining the one that provided the most relevant predictive information. RFECV (Recursive Feature Elimination with Cross-Validation) is an iterative process that ranks features by their importance and recursively removes the least impactful ones to optimize model performance. RFECV was used with a RandomForestClassifier to select impactful features, employing a StratifiedKFold with 10 splits $(n_{splits} = 10)$ to ensure balanced class distribution across folds. We set step size as 1 to remove one feature per iteration for precise selection. The final feature set was chosen based on accuracy across crossvalidation folds. Through these feature selection methods, the following features were identified as relevant for training the ML model: Adjacency Matrix, Autocorrelation, Burden-CAS-Unlike-Topological (BCUT), Charged Partial Sub Area (CPSA), Electrotopological State (Estate), Molecular Representation of Structure based on Electron Diffraction (MoRSE), Molecular Operating Environment Types (Moe-Types), MolecularDistanceEdge, TopologicalCharge for drugs and Composition Based Features, Cross Correlation Based Features, Auto Cross Correlation Based Features, Pseudo Correlation Based Features for miRNAs. A detailed description of the features used is given in Table.1.

2.4. Machine learning algorithms

This study used the tree-based pipeline optimization tool TPOT [36] and lazy predict project separately to identify the best algorithm for the ML model. TPOT uses genetic programming to generate the most

Receiver Operating Characteristic (ROC) Curve



Fig. 4. The AUC Curve of the 5-fold Cross-Validation of the model MDbDMRP.

suitable pipeline for the given dataset, thus decreasing computational costs and saving time. In TPOT, the generations were set to 5, which means TPOT goes through five rounds of refining and improving different model pipelines. The population size was set to 20, allowing TPOT to test 20 models in each round, so it has a wide range of options to explore. The cv (cross-validation) was set to 5, which splits the data into five parts, giving a balanced and accurate performance assessment for each model. Lastly, random state was set to 42 to ensure that results are consistent and reproducible by keeping data splits and random choices the same each time. Lazy Predict is a Python library that can provide useful features such as model selection and hyperparameter optimization, which help researchers get the most out of the ML model. In Lazy Predict, custom metric was set to None, so Lazy Predict uses default metrics like accuracy to quickly compare which models perform best. The combination of TPOT and Lazy Predict provided a thorough overview of various models and pipelines. TPOT's genetic programming allowed for deep exploration and optimization, while Lazy Predict's quick comparisons helped validate XGBoost's potential as the most suitable model for our dataset. Both tools were instrumental in selecting and finalizing XGBoost based on performance, accuracy, and computational cost, aligning with the objectives of the study.

2.5. Model implementation

To build our classification model, we leveraged the powerful XGBClassifier from the xgboost library (version 1.4.2), known for its efficiency and accuracy. We split our dataset into a 70 % training set and a 30 % testing set, ensuring reproducibility with a fixed random state. We configured the XGBClassifier with key parameters like use_label_encoder as False to avoid warnings and eval_metric as 'logloss' to guide optimization. After training the model on the training set, we made predictions on the unseen test set and assessed its performance using multiple metrics (Fig. 1). To ensure future usability, we saved the trained model using Python's pickle library.

2.6. Performance evaluation metrics

We set up ways to test how well the new model works. We employed leave-one-out cross-validation (LOOCV) and 5-fold cross-validation to solidify our confidence in the model's performance. LOOCV ensures that every single data point is utilized for testing exactly once [37]. The remaining data points are then combined to form the training set. In 5fold cross-validation, we split the dataset evenly into five groups. We tested the model with one group each time while using the others to make predictions. Additionally, we looked at accuracy (Acc.), sensitivity

Precision-Recall Curves for 5-Fold Cross-Validation



Fig. 5. The Precision-recall Curve of MDbDMRP.

(Sen), specificity (Spec.), and Matthews Correlation Coefficient (MCC) to get a picture of the effectiveness of our model. The calculation formulas for these metrics are as follows:

$$Acc = rac{TP + TN}{TP + TN + FP + FN}, Sen = rac{TP}{TP + FN}, Spec = rac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Along with these metrics, the AUC-ROCs curve demonstrates the model's performance. The confusion matrix, a tool for evaluating the performance and accuracy of a computational model, displays true positives, false positives, true negatives, and false negatives, representing the model's predictive power.

3. Predictive results and validation of MDbDMRP

We employed several feature selection techniques and model hyperparameter optimization methods to develop a machine learning model that perfectly balances accuracy and interpretability. This meticulous process allowed us to identify the optimal set of features. Leveraging TPOT [36], a powerful tool designed to explore various machine learning pipelines, we efficiently identified the XGBClassifier algorithm as the most suitable architecture for all our investigated properties.

3.1. Leave-one-out cross-validation (LOOCV)

To thoroughly assess the performance of MDbDMRP, we opted for Leave-One-Out Cross-Validation (LOOCV). This robust validation technique, a specialized form of k-fold cross-validation, ensures that every single data point is utilized for testing exactly once [37]. The remaining data points are then combined to form the training set. MDbDMRP achieved stellar performance across all evaluated metrics, including accuracy, precision, recall, F1 score, and balanced accuracy, with an outstanding average score of 0.92 (Fig. 2).

3.2. Cross-validation results

To further solidify our confidence in the model's ability to perform well on unseen data, a standard evaluation process known as 5-fold cross-validation was implemented. This method divides the data into training and testing sets, allowing us to assess the model's generalizability [38]. We compared MDbDMRP with several other models, namely: DLP [19], GCFMCL [23], SMMART [22], SMAJL [25],



Fig. 6. Interaction and binding pose of Lidocaine, Metformin and ZINC457488 with precursor miR-21.

SMANMF [30], GCNNMMA [24], GNMFDMA [28], MFIDMA [31], DMR-PEG [32] and DCMF [29]. The models were tested on their own specified dataset and MDbDMRP was tested on its own dataset. The AUCs of DLP, GCFMCL, SMMART, SMAJL, SMANMF, GCNNMMA, GNMFDMA, MFIDMA, DMR-PEG and DCMF are 0.8729, 0.9528, 0.8588, 0.8746, 0.8429, 0.9384, 0.9193, 0.9444, 0.9475 and 0.9868 where average AUC of MDbDMRP is 0.9864. MDbDMRP maintained an average score of 0.9199 in evaluation metrics, including accuracy, precision, recall, F1 score, balanced accuracy, and Matthews Correlation Coefficient (Fig. 3). This reinforces the model's effectiveness in tackling real-world applications with a high degree of certainty. Additionally, the exceptional average AUC-ROC score of 0.9864 indicates the strong ability of the model to distinguish between positive and negative classes (Fig. 4). The precision-recall curve also validates the performance of model (Fig. 5). This approach enables researchers to ensure the model performs well not only on the specific training data but also on unseen data, thereby enhancing the model's performance credibility.

3.3. Case study

To further validate the performance of MDbDMRP, we conducted a case study. miR-21 is one of the extensively studied miNRA. Many studies have shown the role of miR-21 in different types of cancer [39–42]. In later studies, miR-21 was established to be an oncogenic microRNA [43–50]. So we used MDbDMRP to predict drugs that may bind to miR-21. The model has shown that Metformin (CID4091) and Lidocaine (CID3676) can bind to miR-21 (pdb id:2mnc). To validate our predictions, we performed molecular docking to show the binding of these predicted drug molecules to miR-21. ZINC4574788 was identified as the potential molecule for docking against miR-21 [51]. CDOCKER a molecular docking technique which is built on CHARMM, can deliver extremely precise docking outcomes [52]. We analyzed the interaction energy and profile of Metformin, Lidocaine and ZINC4574788 were -33

kcal/mol, -13 kcal/mol and -11 kcal/mol respectively. The drug molecules predicted using MDbDMRP showed improved performance in docking experiments and yielded better results, highlighting the model's effectiveness in predicting drug-miRNA associations.

4. Concluding remarks and future perspectives

Many researchers are looking for computational methods to decipher new drug-miRNA associations. This paper represented a molecular descriptor-based method, MDbDMRP, that deduced the potential relationship between miRNAs and drug molecules by combining known drug-miRNA associations with molecular descriptors and machinelearning-based methods. We constructed a framework that integrated molecular descriptors, feature selection methods, and machine-learning algorithms to train the model and predict new potential relationships with unseen data. Since MDbDMRP incorporated all the miRNAs for Homo sapiens available in the miRBase [33], it can show the most accurate predictions. The stellar performance of MDbDMRP can be attributed to the following points. First, various molecular descriptors were calculated for drugs and miRNAs, the mathematical representations of molecular properties generated by algorithms. Second, feature selection techniques were employed to select the relevant and important feature subsets required to train an ML model. Feature selection techniques helped remove redundant and noisy data, which added to decreased computational complexity and better prediction accuracy for the model. Third, using genetic programming, the tree-based pipeline optimisation tool (TPOT)18 was used to generate the most suitable pipeline for our dataset, thus reducing the computational cost and time. Fourth, XGBClassifier, a special implementation of XGBoost (eXtreme Gradient Boosting), was used to train the model, leveraging the benefits of XGBoost like performance, accuracy, flexibility and regularization. The efficient data handling, parallel processing and out-of-core computation reduce the computational cost, and regularization prevents overfitting. Finally, MDbDMRP can predict new drug-miRNA

A. Daroch and R. Purohit

relationships on unseen datasets.

The performance of MDbDMRP was evaluated with leave-one-out cross-validation and 5-fold cross-validation. The model achieved AUC-ROC score of 0.9864. The drug molecules predicted by MDbDMRP demonstrated superior performance in docking experiments and produced favorable results, underscoring the model's effectiveness in predicting drug-miRNA associations. MDbDMRP can be used to predict the relationship of any drug molecule with miRNAs, which could be valuable information for scientists studying drug repositioning.

However, MDbDMRP has some limitations also. For example, users need to calculate molecular descriptors from the tools used in the study, i.e. Mordred [34] and NFeature [35] for drugs and miRNAs, respectively. Users must also generate new files with their molecules of interest and their respective molecular descriptors to predict the relationships. We anticipate that more computational models combined with molecular descriptors could be developed to predict new drug-miRNA relationships for better drug repositioning and miRNA-targeted drug development.

Software information

Python version: 3.11.5. Anaconda Jupyter Lab: 3.6.3. xgBoost version: 1.4.2.

CRediT authorship contribution statement

Amit Daroch: Writing – original draft, Validation, Software, Resources, Investigation, Formal analysis, Data curation. **Rituraj Purohit:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Funding

The work was carried out under the aegis of The Himalayan Centre for High-throughput Computational Biology (HiCHiCoB), a BIC supported by DBT, Govt. of India [BT/PR40122/BTIS/137/30/2021].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was carried out under The Himalayan Centre for Highthroughput Computational Biology (HiCHiCoB), a BIC supported by DBT, Govt. of India. We gratefully acknowledge CSIR-Institute of Himalayan Bioresource Technology, Palampur, for providing the facilities to carry out this work. The CSIR support in the form of project MLP: 0201 for bioinformatics studies is highly acknowledge. This manuscript represents CSIR-IHBT Communication No. 5709.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2024.138580.

Data availability

The authors declare that all the miRNAs used in the model are human-related. The model, required files, and dataset can be accessed publically at: https://github.com/Daroch-Amit/MDbDMRP.

References

- D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2004) 281–297, https://doi.org/10.1016/S0092-8674(04)00045-5.
- [2] R.C. Lee, R.L. Feinbaum, V. Ambros, The C. Elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75 (1993) 843–854, https://doi.org/10.1016/0092-8674(93)90529-Y.
- [3] A.M. Cheng, M.W. Byrom, J. Shelton, L.P. Ford, Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis, Nucleic Acids Res. 33 (2005) 1290–1297, https://doi.org/10.1093/NAR/GK1200.
- [4] E.A. Miska, How microRNAs control cell division, differentiation and death, Curr. Opin. Genet. Dev. 15 (2005) 563–568, https://doi.org/10.1016/J. GDE.2005.08.005.
- [5] P. Xu, M. Guo, B.A. Hay, MicroRNAs and the regulation of cell death, Trends Genet. 20 (2004) 617–624, https://doi.org/10.1016/J.TIG.2004.09.010.
- [6] N. Stern-Ginossar, N. Elefant, A. Zimmermann, D.G. Wolf, N. Saleh, M. Biton, E. Horwitz, Z. Prokocimer, M. Prichard, G. Hahn, D. Goldman-Wohl, C. Greenfield, S. Yagel, H. Hengel, Y. Altuvia, H. Margalit, O. Mandelboim, Host immune system gene targeting by a viral miRNA, Science (80-.). 317 (2007) 376–381, https://doi. org/10.1126/SCIENCE.1140956/SUPPL_FILE/STERN-GINOSSAR.SOM.REV1.PDF.
- [7] R.A. Shivdasani, MicroRNAs: regulators of gene expression and cell differentiation, Blood 108 (2006) 3646–3653, https://doi.org/10.1182/BLOOD-2006-01-030015.
- [8] Z. Zhou, L. Zhuo, X. Fu, J. Lv, Q. Zou, R. Qi, Joint masking and self-supervised strategies for inferring small molecule-miRNA associations, Mol. Ther. - Nucleic Acids 35 (2024) 102103, https://doi.org/10.1016/J.OMTN.2023.102103.
- [9] X. Cao, X. Zhou, F. Hou, Y. Huang, M. Yuan, M. Long, S. Chen, W. Lei, J. Zhu, J. Chen, T. Zhang, A.-Y. Guo, W. Jiang, ncRNADrug: a database for validated and predicted ncRNAs associated with drug resistance and targeted by drugs, Nucleic Acids Res. 52 (2024) D1393–D1399, https://doi.org/10.1093/NAR/GKAD1042.
- [10] X. Liu, S. Wang, F. Meng, J. Wang, Y. Zhang, E. Dai, X. Yu, X. Li, W. Jiang, SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression, Bioinformatics 29 (2013) 409–411, https://doi.org/10.1093/ bioinformatics/bts698.
- [11] L. Li, P. Wu, Z. Wang, X. Meng, C. Zha, Z. Li, T. Qi, Y. Zhang, B. Han, S. Li, C. Jiang, Z. Zhao, J. Cai, NoncoRNA: a database of experimentally supported non-coding RNAs and drug targets in cancer, J. Hematol. Oncol. 13 (2020) 1–4, https://doi. org/10.1186/S13045-020-00849-7/FIGURES/1.
- [12] X. Chen, W. Bin Xie, P.P. Xiao, X.M. Zhao, H. Yan, mTD: a database of microRNAs affecting therapeutic effects of drugs, J. Genet. Genomics 44 (2017) 269–271. doi: https://doi.org/10.1016/J.JGG.2017.04.003.
- [13] X. Chen, Y.Z. Sun, D.H. Zhang, J.Q. Li, G.Y. Yan, J.Y. An, Z.H. You, NRDTD: a database for clinically or experimentally supported non-coding RNAs and drug targets associations, Database 2017 (2017), https://doi.org/10.1093/DATABASE/ BAX057.
- [14] J. Li, H. Lin, Y. Wang, Z. Li, B. Wu, Prediction of potential small molecule—miRNA associations based on heterogeneous network representation learning, Front. Genet. 13 (2022) 1079053, https://doi.org/10.3389/FGENE.2022.1079053/ BIBTEX.
- [15] L. Peng, Y. Tu, L. Huang, Y. Li, X. Fu, X. Chen, DAESTB: inferring associations of small molecule–miRNA via a scalable tree boosting model based on deep autoencoder, Brief. Bioinform. 23 (2022) 1–14, https://doi.org/10.1093/BIB/ BBAC478.
- [16] C.C. Wang, C.C. Zhu, X. Chen, Ensemble of kernel ridge regression-based small molecule-miRNA association prediction in human disease, Brief. Bioinform. 23 (2022) 1–11, https://doi.org/10.1093/BIB/BBAB431.
- [17] N.N. Guan, Y.Z. Sun, Z. Ming, J.Q. Li, X. Chen, Prediction of potential small molecule-associated microRNAs using graphlet interaction, Front. Pharmacol. 9 (2018) 394684, https://doi.org/10.3389/FPHAR.2018.01152/BIBTEX.
- [18] Y. Zhong, C. Shen, X. Xi, Y. Luo, P. Ding, L. Luo, Multitask joint learning with graph autoencoders for predicting potential MiRNA-drug associations, Artif. Intell. Med. 145 (2023) 102665, https://doi.org/10.1016/J.ARTMED.2023.102665.
- [19] K. Zheng, G. Duan, Q. Zhao, M. Yang, X. Liang, Y. Liu, J. Wang, DLP: duplex link prediction via subspace segmentation for predicting drug-MiRNA associations, IEEE/ACM Trans. Comput. Biol. Bioinforma. (2024), https://doi.org/10.1109/ TCBB.2024.3402215.
- [20] W. Lan, J. Wang, M. Li, J. Liu, Y. Li, F.X. Wu, Y. Pan, Predicting drug-target interaction using positive-unlabeled learning, Neurocomputing 206 (2016) 50–57, https://doi.org/10.1016/J.NEUCOM.2016.03.080.
- [21] W. Peng, Z. Che, W. Dai, S. Wei, W. Lan, Predicting miRNA-disease associations from miRNA-gene-disease heterogeneous network with multi-relational graph convolutional network model, IEEE/ACM Trans. Comput. Biol. Bioinforma. 20 (2023) 3363–3375, https://doi.org/10.1109/TCBB.2022.3187739.
- [22] C. Shen, J. Luo, W. Ouyang, P. Ding, H. Wu, Identification of small moleculemiRNA associations with graph regularization techniques in heterogeneous networks, J. Chem. Inf. Model. 60 (2020) 6709–6721, https://doi.org/10.1021/ ACS_JCIM.0C00975/ASSET/IMAGES/LARGE/CI0C00975 0006.JPEG.
- [23] L. Zhuo An Associate Professor, J. Wei, L. Zhuo, Z. Zhou, X. Lian, X. Fu, X. Yao, Problem solving protocol GCFMCL: predicting miRNA-drug sensitivity using graph collaborative filtering and multi-view contrastive learning, Brief. Bioinform. 2023 (2023) 1–11, https://doi.org/10.1093/bib/bbad247.
- [24] Z. Niu, X. Gao, Z. Xia, S. Zhao, H. Sun, H. Wang, M. Liu, X. Kong, C. Ma, H. Zhu, H. Gao, Q. Liu, F. Yang, X. Song, J. Lu, X. Zhou, Prediction of small molecule drugmiRNA associations based on GNNs and CNNs, Front. Genet. 14 (2023) 1201934, https://doi.org/10.3389/FGENE.2023.1201934/BIBTEX.
- [25] C. Shen, J. Luo, Z. Lai, P. Ding, Multiview joint learning-based method for identifying small-molecule-associated MiRNAs by integrating pharmacological,

genomics, and network knowledge, J. Chem. Inf. Model. 60 (2020) 4085–4097, https://doi.org/10.1021/ACS.JCIM.0C00244/ASSET/IMAGES/LARGE/ CI0C00244 0006_JPEG.

- [26] W. Peng, Z. He, W. Dai, W. Lan, MHCLMDA: multihypergraph contrastive learning for miRNA–disease association prediction, Brief. Bioinform. 25 (2023), https://doi. org/10.1093/BIB/BBAD524.
- [27] W. Lan, J. Wang, M. Li, J. Liu, F.X. Wu, Y. Pan, Predicting MicroRNA-disease associations based on improved MicroRNA and disease similarities, IEEE/ACM Trans. Comput. Biol. Bioinforma. 15 (2018) 1774–1782, https://doi.org/10.1109/ TCBB.2016.2586190.
- [28] M.N. Wang, Y. Li, L.L. Lei, D.W. Ding, X.J. Xie, Combining non-negative matrix factorization with graph Laplacian regularization for predicting drug-miRNA associations based on multi-source information fusion, Front. Pharmacol. 14 (2023) 1132012, https://doi.org/10.3389/FPHAR.2023.1132012/BIBTEX.
- [29] S.H. Wang, C.C. Wang, L. Huang, L.Y. Miao, X. Chen, Dual-network collaborative matrix factorization for predicting small molecule-miRNA associations, Brief. Bioinform. 23 (2022), https://doi.org/10.1093/BIB/BBAB500.
- [30] J. Luo, C. Shen, Z. Lai, J. Cai, P. Ding, Incorporating clinical, chemical and biological information for predicting small molecule-microRNA associations based on non-negative matrix factorization, IEEE/ACM Trans. Comput. Biol. Bioinforma. 18 (2021) 2535–2545, https://doi.org/10.1109/TCBB.2020.2975780.
- [31] Y.J. Guan, C.Q. Yu, Y. Qiao, L.P. Li, Z.H. You, Z.H. Ren, Y.C. Li, J. Pan, MFIDMA: A Multiple Information Integration Model for the Prediction of Drug-miRNA Associations, Biol. 2023, Vol. 12, Page 41 12 (2022) 41. doi:https://doi.org/10.33 90/BIOLOGY12010041.
- [32] C. Zhao, H. Wang, W. Qi, S. Liu, Toward drug-miRNA resistance association prediction by positional encoding graph neural network and multi-channel neural network, Methods 207 (2022) 81–89, https://doi.org/10.1016/J. YMFTH 2022 09 005
- [33] A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, miRBase: from microRNA sequences to function, Nucleic Acids Res. 47 (2019) D155–D162, https://doi.org/ 10.1093/NAR/GKY1141.
- [34] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator, J. Chemother. 10 (2018) 1–14, https://doi.org/10.1186/S13321-018-0258-Y/FIGURES/6.
- [35] M. Mathur, S. Patiyal, A. Dhall, S. Jain, R. Tomer, A. Arora, G.P.S. Raghava, G.P.S. Raghava, Nfeature: a platform for computing features of nucleotide sequences, BioRxiv (2021) 2021.12.14.472723. doi:https://doi.org/10.1101/2021.12.14. 472723.
- [36] T.T. Le, W. Fu, J.H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector, Bioinformatics 36 (2020) 250–256, https://doi.org/10.1093/BIOINFORMATICS/BTZ470.
- [37] T.T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recogn. 48 (2015) 2839–2846, https://doi. org/10.1016/J.PATCOG.2015.03.009.
- [38] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat. Comput. 21 (2011) 137–146, https://doi.org/10.1007/S11222-009-9153-8/ METRICS.

- [39] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, J.R. Downing, T. Jacks, H.R. Horvitz, T.R. Golub, MicroRNA expression profiles classify human cancers, Nat. 2005 4357043 435 (2005) 834–838. doi:https://doi.org/10.1038/nature03702.
- [40] A. Esquela-Kerscher, F.J. Slack, Oncomirs microRNAs with a role in cancer, Nat. Rev. Cancer 2006 64 6 (2006) 259–269. doi:https://doi.org/10.1038/nrc1840.
- [41] W. Zhang, J.E. Dahlberg, W. Tam, MicroRNAs in tumorigenesis: a primer, Am. J. Pathol. 171 (2007) 728–738, https://doi.org/10.2353/AJPATH.2007.070070.
- [42] J.A. Chan, A.M. Krichevsky, K.S. Kosik, MicroRNA-21 is an Antiapoptotic factor in human glioblastoma cells, Cancer Res. 65 (2005) 6029–6033, https://doi.org/ 10.1158/0008-5472.CAN-05-0137.
- [43] S. Zhu, H. Wu, F. Wu, D. Nie, S. Sheng, Y.Y. Mo, MicroRNA-21 targets tumor suppressor genes in invasion and metastasis, Cell Res. 2008 183 18 (2008) 350–359. doi:https://doi.org/10.1038/cr.2008.24.
- [44] S. Zhu, M.L. Si, H. Wu, Y.Y. Mo, MicroRNA-21 targets the tumor suppressor gene tropomyosin 1 (TPM1), J. Biol. Chem. 282 (2007) 14328–14336, https://doi.org/ 10.1074/JBC.M611393200.
- [45] Z. Lu, M. Liu, V. Stribinskis, C.M. Klinge, K.S. Ramos, N.H. Colburn, Y. Li, MicroRNA-21 promotes cell transformation by targeting the programmed cell death 4 gene, Oncogene 2008 2731 27 (2008) 4373–4379. doi:https://doi.org/ 10.1038/onc.2008.72.
- [46] T. Li, D. Li, J. Sha, P. Sun, Y. Huang, MicroRNA-21 directly targets MARCKS and promotes apoptosis resistance and invasion in prostate cancer cells, Biochem. Biophys. Res. Commun. 383 (2009) 280–285, https://doi.org/10.1016/J. BBRC.2009.03.077.
- [47] L.B. Frankel, N.R. Christoffersen, A. Jacobsen, M. Lindow, A. Krogh, A.H. Lund, Programmed cell death 4 (PDCD4) is an important functional target of the MicroRNA miR-21 in breast Cancer cells, J. Biol. Chem. 283 (2008) 1026–1033, https://doi.org/10.1074/JBC.M707224200.
- [48] I.A. Asangani, S.A.K. Rasheed, D.A. Nikolova, J.H. Leupold, N.H. Colburn, S. Post, H. Allgayer, MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer, Oncogene 2008 2715 27 (2007) 2128–2136. doi:https://doi. org/10.1038/sj.onc.1210856.
- [49] J. Zheng, H. Xue, T. Wang, Y. Jiang, B. Liu, J. Li, Y. Liu, W. Wang, B. Zhang, M. Sun, miR-21 downregulates the tumor suppressor P12CDK2AP1 and stimulates cell proliferation and invasion, J. Cell. Biochem. 112 (2011) 872–880, https://doi. org/10.1002/JCB.22995.
- [50] R. Kumarswamy, I. Volkmann, T. Thum, Regulation and function of miRNA-21 in health and disease, RNA Biol. 8 (2011), https://doi.org/10.4161/rna.8.5.16154.
- [51] M. Junaid, R. Dash, N. Islam, J. Chowdhury, M.J. Alam, S.D. Nath, M.A.S. Shakil, A. Azam, S.M. Quader, S.M., Zahid hosen, molecular simulation studies of 3,3'-Diindolylmethane as a potent MicroRNA-21 antagonist, J. Pharm. Bioallied Sci. 9 (2017) 259–265, https://doi.org/10.4103/JPBS_JPBS_266_16.
- [52] G. Wu, D.H. Robertson, C.L. Brooks, M. Vieth, Detailed analysis of grid-based molecular docking: a case study of CDOCKER—A CHARMm-based MD docking algorithm, J. Comput. Chem. 24 (2003) 1549–1562, https://doi.org/10.1002/ JCC.10306.