RESEARCH Open Access

Utilizing RNA-seq data in monotone iterative generalized linear model to elevate prior knowledge quality of the circRNA-miRNA-mRNA regulatory axis

Alikhan Anuarbekov[1] and Jiří Kléma[1*]

*Correspondence:
klema@fel.cvut.cz

[1] Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague, Technicka 2, 16627 Prague, Czech Republic

Abstract

Background: Current experimental data on RNA interactions remain limited, particularly for non-coding RNAs, many of which have only recently been discovered and operate within complex regulatory networks. Researchers often rely on in-silico interaction detection algorithms, such as TargetScan, which are based on biochemical sequence alignment. However, these algorithms have limited performance. RNA-seq expression data can provide valuable insights into regulatory networks, especially for understudied interactions such as circRNA-miRNA-mRNA. By integrating RNA-seq data with prior interaction networks obtained experimentally or through in-silico predictions, researchers can discover novel interactions, validate existing ones, and improve interaction prediction accuracy.

Results: This paper introduces Pi-GMIFS, an extension of the generalized monotone incremental forward stagewise (GMIFS) regression algorithm that incorporates prior knowledge. The algorithm first estimates prior response values through a prior-only regression, interpolates between these prior values and the original data, and then applies the GMIFS method. Our experimental results on circRNA-miRNA-mRNA regulatory interaction networks demonstrate that Pi-GMIFS consistently enhances precision and recall in RNA interaction prediction by leveraging implicit information from bulk RNA-seq expression data, outperforming the initial prior knowledge.

Conclusion: Pi-GMIFS is a robust algorithm for inferring acyclic interaction networks when the variable ordering is known. Its effectiveness was confirmed through extensive experimental validation. We proved that RNA-seq data of a representative size help infer previously unknown interactions available in TarBase v9 and improve the quality of circRNA disease annotation.

Keywords: Bayesian network, Structure inference, Circular RNA, Penalized regression, Functional annotation

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

The advent of modern high-throughput methods for RNA profiling (RNA-seq) has made it possible to efficiently sequence enormous amounts of transcript data. The collection of such large quantities of data enables the modelling of a more comprehensive and complex view of the molecular system of a cell [1–3]. This cellular system, composed of interacting genes and transcript products, forms a dynamic interaction network of co-regulating RNA elements [4].

Competing endogenous RNA (ceRNA) networks encode how various types of RNAs compete for shared micro RNAs (miRNAs), aiding in the understanding of gene regulation and cellular systems. circular RNAs (circRNAs) are the least explored RNA type in these networks. Current RNA-seq biology research has focused predominantly on circRNAs because of their emerging roles in gene regulation and disease mechanisms [5]. Although circRNAs were first discovered more than 20 years ago, the functional role of circRNAs in an organism is still not fully understood [6, 7]. Notably, they have been found to be a part of multiple pathways, including the RNA trap for messenger RNA (mRNA) production [6, 7] or miRNA selective regulation by sponging [7]. The latter pathway in the form of a circRNA-miRNA-mRNA regulatory interaction network, a special case of a ceRNA network [8], plays a crucial pathogenic role [7, 9, 10].

The miRNA-mRNA segment of the ceRNA network is supported by a substantial body of experimentally validated evidence accumulated over several decades. The most robust validation methods for demonstrating miRNA-mRNA associations include luciferase reporter assays, quantitative reverse transcription polymerase chain reaction (qRT-PCR), and western blots [11]. Consequently, current databases, such as TarBase v9 [12], are founded on strong empirical evidence. However, these experimental techniques are both expensive and time-consuming, creating a demand for computational in-silico tools to help narrow down potential targets [13, 14].

The importance of in-silico tools has increased further for less experimentally explored circRNA-miRNA interactions. Most circRNA-miRNA interaction databases rely solely on in-silico tools such as TargetScan [15] or miRanda [16]. Although there has been a recent increase in experimentally validated circRNA-miRNA interactions, as noted in [17], the overall quantity of such interactions remains insufficient for a complete shift to reliance on validated interactions alone.

RNA-seq profile data, along with other existing knowledge sources such as sequence alignment predictions and few experimentally validated results, can help improve interaction predictions. In this study, we consider non-RNA-seq data as prior knowledge. This prior knowledge acts as a benchmark for evaluating prediction tools and can help reduce the need for large RNA-seq sample sizes. This approach is particularly useful in cases of data sparsity, such as circRNA-miRNA-mRNA studies, which often follow the “large p, small n” paradigm (where p is the number of genes/RNAs and n is the number of samples) [18].

The recent increase in the number of RNA-seq datasets, despite their heterogeneous nature, may facilitate the combination of prior knowledge with diverse samples. This could enhance the quality of initial prior knowledge predictions and assist in identifying candidate RNA-RNA interactions for further experimental validation. The procedure of transforming the collection of RNA-seq samples into an enhanced interaction network of superior quality is referred to as the training process.

Despite these advancements, no existing study, to our knowledge, has introduced an algorithm that remains numerically stable when handling heterogeneous RNA-seq data while also integrating prior knowledge. The primary contribution of this paper is the development of such an algorithm, Pi-GMIFS. We evaluated its performance using a large RNA-seq dataset and analyzed the relationship between prior knowledge quality and the required RNA-seq sample size.

Finally, we applied Pi-GMIFS to infer a circRNA-miRNA interaction network in a setting where no experimentally validated prior knowledge was available. Additionally, we propose a method for assessing prediction quality when a ground-truth reference structure is absent.

Related work

Initially, the sequence alignment approach was one of the first attempts to describe the enigmatic regulatory mechanism between circRNAs and miRNAs [19]. Unsurprisingly, the sequence alignment algorithms, such as the abovementioned TargetScan or miRanda, initially developed for miRNA-mRNA target prediction [13, 20], were generalized to account for a new type of interaction. Nevertheless, despite providing experiment-free predictions for potential targets, their usage has been limited due to non-canonical region alignments [21] or high levels of false positive predictions [22].

Concurrently, methods based on high-throughput RNA-seq analysis have become accessible because of the emergence of numerous databases with RNA-seq profiles. Unlike experiments that concentrate on specific miRNAs or circRNAs of interest, these samples prepare a full-scale profile of all types of transcripts [23].

The foremost research started from correlation matrix analysis and by evaluating the correlation of each transcript pair separately. The problem, however, lies in the combinatorial effect, such as multiple miRNAs regulating the same mRNA simultaneously [24]. The solution was found in modelling the whole regulatory network, as it is employed in the general graphical framework of Bayesian probabilistic networks [25–28]. The algorithms designed for inferring the structure of general Bayesian networks were initially developed to address problems with unknown topological orderings. Despite improvements in the scalability of these methods-such as the tiled-ALARM network, which has approximately 10,000 variables and is inferred via the *MMPC* algorithm [27], or the 5,000-variable Gaussian network inferred via the *sparsebn* framework [29], the inference challenge remains significant. The persistence is expected, as structural inference in acyclic Bayesian networks without known topological ordering has been proven to be NP-hard [30].

An active research area involves applying Large Language Models (LLMs) to integrate both expression and sequence information of RNAs to infer RNA or protein structural interactions [31]. A key focus is developing scalable approaches that can efficiently tackle NP-hard problems. However, despite these potential scalability benefits, there remains an ongoing debate regarding the reliability and generalizability of LLM outputs, particularly in mathematical and logical problem solving [32].

Alternatively, the isolated analysis of individual subgraphs, e.g. regulatory axes, of the whole biological network is a common approach [33, 34] and thus is employed in this study in a specific case of the circRNA-miRNA-mRNA axis. The known ordering allows for the decomposition of an inference per RNA, reducing its complexity to a polynomial. A typical approach to integrate multiple regulatory RNAs is the modelling with a probabilistic regression-based model. To account for the non-linear nature of the relationships, Generalized Linear Models (GLMs) are used.

The idea of GLM usage in the RNA-seq data has been employed in [35–37], although instead of directly using regression to determine the interactions, the focus is shifted to metadata analysis such as the effect of experimental setup or tissue type. The inference of interactions is conducted on the basis of differential expression analysis similar to correlation analysis [36, 38].

Alternatively, modern state-of-the-art inference in a context of regression is performed via LASSO [39, 40] or ElasticNet [41] penalization. A specific Negative-binomial or Poisson type of GLMs designed for RNA-seq data was first implemented as penalized inference in [42, 43]. However, it suffers from numeric instabilities and our experiments have shown that it cannot be used in enormous heterogeneous datasets of RNA-seq profiles. A study [44] reported a similar issue, although it was attributed to the small proportion of their RNA-seq dataset, which allowed them to disregard the problem.

Recent work from [45] has introduced a stable iterative solution path, generalized monotone incremental forward stagewise (GMIFS), which mitigates the problem of numeric instabilities in heterogeneous and enormous RNA-seq data. The idea is to shift from large optimization steps to smaller steps to avoid numerically unstable regions of high-dimensional space and follow the solution path only.

The previously mentioned LASSO GLM inference is known to have a method of direct prior incorporation in the form of varying penalty hyperparameter weights on the basis of the amount and quality of evidence supporting a particular interaction [42]. However, the absence of direct hyperparameter modelling prevents GMIFS from directly incorporating the prior in the same way.

We propose a solution based on an alternative approach first introduced in [46]. It combines predictions from a prior-only ideal GLM network with the actual RNA-seq input data.

The novelty of the proposed algorithm, Prior-incorporated GMIFS (Pi-GMIFS), lies in its ability to overcome numerical instability through a forward stagewise procedure and its unconventional method of incorporating prior information. While the first approach allows the use of highly heterogeneous datasets, the number of RNA-seq samples required without prior information makes this approach impractical. The key reason is the lack of a sufficient number of RNA-seq samples in fields where the algorithm is most needed, such as in circRNA-miRNA interaction inference. Moreover, when many small-sample datasets are added to reach the required amount, the information from each additional dataset may be invalidated by batch effects due to different conditions and tissues.

Conversely, by incorporating prior information, the threshold for the minimum RNA-seq sample size is reduced, as information is implicitly extracted from the prior. This reduction is demonstrated empirically by comparing the performance and the relationship between

RNA-seq sample size and prior information in experimentally validated scenarios. Furthermore, the combination of prior information and RNA-seq data has been shown to improve the performance of the secondary task of automated disease annotation, indicating an indirect benefit.

Materials and methods

In the following section, we first present the formal problem definition, then introduce the publicly available data, and finally describe the proposed algorithm, Pi-GMIFS.

Problem statement

Let $j = 1, ..., M$ index the RNA-seq samples, and $i = 1, ..., N$ index the RNA transcripts quantified in each sample. The outcome of the RNA-seq profiling is represented by the read count matrix $X \in \mathbb{N}_0^{M \times N}$, where each column corresponds to the i-th RNA transcript and each row represents an j-th sample. The RNA transcripts are categorized into three types: circular RNAs (circRNAs), microRNAs (miRNAs), and messenger RNAs (mRNAs). Let $\boldsymbol{c} = \{\boldsymbol{c}^1, ..., \boldsymbol{c}^{N_c}\}$, $\boldsymbol{m} = \{\boldsymbol{m}^1, ..., \boldsymbol{m}^{N_m}\}$ and $\boldsymbol{r} = \{\boldsymbol{r}^1, ..., \boldsymbol{r}^{N_r}\}$ represent the sets of circRNA, miRNA, and mRNA variables, respectively. The entire count matrix is then $X = (\boldsymbol{c}^T, \boldsymbol{m}^T, \boldsymbol{r}^T)^T$ and $N = N_c + N_m + N_r$.

This leads to the joint probability distribution:

$$\mathcal{P}(\boldsymbol{c}^1, ..., \boldsymbol{c}^{N_c}, \boldsymbol{m}^1, ..., \boldsymbol{m}^{N_m}, \boldsymbol{r}^1, ..., \boldsymbol{r}^{N_r}) = \mathcal{P}(\boldsymbol{c}, \boldsymbol{m}, \boldsymbol{r}) \tag{1}$$

We adopt a tripartite ceRNA [7, 9, 10] Bayesian network model to represent the dependencies between these three distinct RNA variable sets. This structure reflects the known regulatory relationships between RNA types for each variable, as described by the following factorization:

$$\mathcal{P}(\boldsymbol{c}, \boldsymbol{m}, \boldsymbol{r}) = \prod_{i=1}^{N_c} \mathcal{P}(\boldsymbol{c}^i \mid \emptyset) \prod_{i=1}^{N_m} \mathcal{P}(\boldsymbol{m}^i \mid \boldsymbol{c}) \prod_{i=1}^{N_r} \mathcal{P}(\boldsymbol{r}^i \mid \boldsymbol{m}) \tag{2}$$

For the remainder of the paper, the formulas are provided for the miRNA-mRNA terms only, e.g., $\mathcal{P}(\boldsymbol{r}^i \mid \boldsymbol{m})$, although they can be generalized to other terms straightforwardly.

The task of interaction network learning, also known as feature selection [40, 46], involves finding a subset of all possible conditional RNAs, e.g., the parent set $Pa(\boldsymbol{r}^i) \subseteq \boldsymbol{m}$, which actually contributes to the probability distribution of this particular RNA, e.g., $\mathcal{P}(\boldsymbol{r}^i \mid \boldsymbol{m}) = \mathcal{P}(\boldsymbol{r}^i \mid Pa(\boldsymbol{r}^i))$. Owing to the decomposition of the Bayesian network model in Eq. 2, the inference could be performed per-RNA independently.

A summary of the notation is provided in Fig. 1.

The proposed inference algorithm then predicts the parent set according to the count matrix:

$$f : X, \boldsymbol{r}^i \mapsto Pa(\boldsymbol{r}^i) \tag{3}$$

In addition to the count matrix X, prior knowledge about structural predictions from previous studies is also incorporated. In this study, prior knowledge is simplified to include only the presence of interaction entries, without any evaluation of interaction quality. This simplification yields the prior parent subset $Pa^{\pi}(\boldsymbol{r}^i)$, which represents the

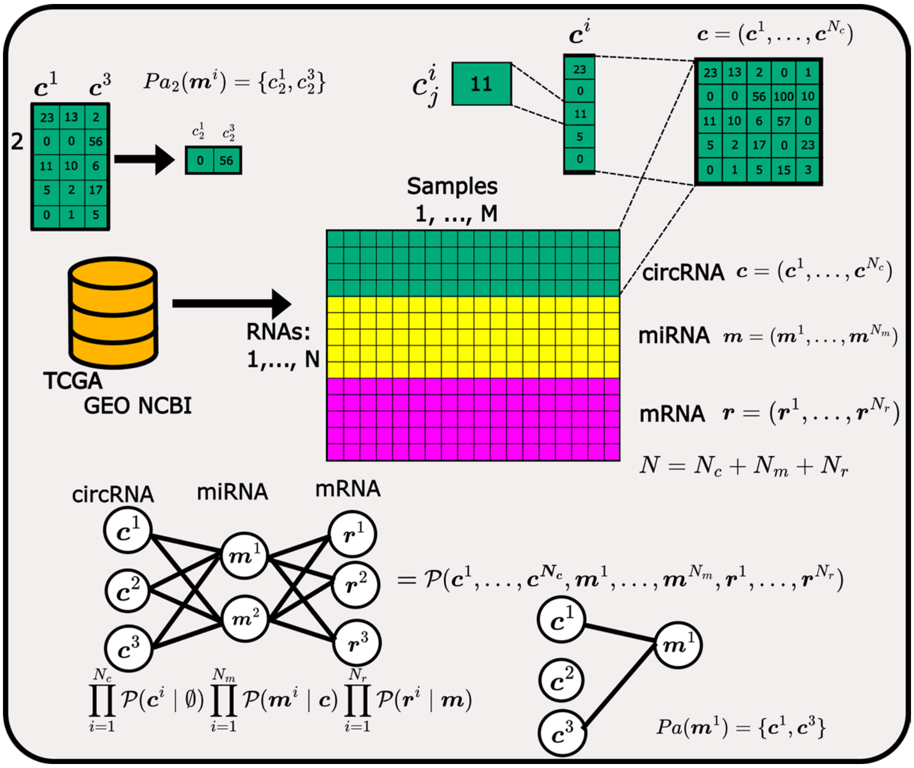

c^1 c^3
$Pa_2(m^i) = \{c_2^1, c_2^3\}$
2
23 13 2
0 0 56
11 10 6
5 2 17
0 1 5
c_2^1 c_2^3
0 56
c_j^i
11
c^i
23 0 11 5 0
$c = (c^1, \ldots, c^{N_c})$
23 13 2 0 1
0 0 56 100 10
11 10 6 57 0
5 2 17 0 23
0 1 5 15 3
Samples 1, ..., M
TCGA
GEO NCBI
RNAs: 1,..., N
circRNA $c = (c^1, \ldots, c^{N_c})$
miRNA $m = (m^1, \ldots, m^{N_m})$
mRNA $r = (r^1, \ldots, r^{N_r})$
$N = N_c + N_m + N_r$
circRNA
miRNA
mRNA
c^1 c^2 c^3 m^1 m^2 r^1 r^2 r^3
$= \mathcal{P}(c^1, \ldots, c^{N_c}, m^1, \ldots, m^{N_m}, r^1, \ldots, r^{N_r})$
$\prod_{i=1}^{N_c} \mathcal{P}(c^i \mid \emptyset) \prod_{i=1}^{N_m} \mathcal{P}(m^i \mid c) \prod_{i=1}^{N_r} \mathcal{P}(r^i \mid m)$
c^1 c^2 c^3 m^1
$Pa(m^1) = \{c^1, c^3\}$

Fig. 1 A summary of the tripartite ceRNA Bayesian network probabilistic model and the RNA-seq count data notation introduced above. An example of parent subset notation is also provided

set of prior interactions for each transcript $\boldsymbol{r}^i$. Consequently, the prediction function changes to:

$$f : X, Pa^{\pi}(\boldsymbol{r}^i), \boldsymbol{r}^i \mapsto Pa(\boldsymbol{r}^i) \tag{4}$$

The goal of the proposed algorithm is to overcome the prior parent prediction Pa^{π} in terms of the correctness of the parent subset. The correctness is evaluated in comparison with the assumed ground-truth parent set, Pa^{true}, which is defined by the state-of-the-art database of experimentally validated interactions. A metric to measure the correctness used in this study is the F1-score, although it can be any structure comparison metric such as the accuracy, Hamming distance [27], etc. We denote the metric as the function $S : Pa \times Pa^{true} \mapsto \mathbb{R}_0^+$. The goal of the algorithm could then be formulated as:

$$S(Pa, Pa^{true}) > S(Pa^{\pi}, Pa^{true}) \tag{5}$$

Data

In our study, we work with three principal types of datasets. First, we compiled a representative pool of RNA-seq expression profiles. Second, we utilize existing computationally predicted or experimentally validated RNA interaction data. Finally, we collected a set of available circRNA-disease annotations for the indirect evaluation of our algorithm.

RNA-seq data collection

We collected a sample pool from the Gene Expression Omnibus (GEO) [47](https://www.ncbi.nlm.nih.gov/geo/). The database provides a large catalogue of high-throughput gene and RNA expression profiles. Among these profiles, large enough RNA-seq datasets with measurements of all three RNA types from Homo Sapiens were selected and downloaded, and a total of 434 samples were obtained. Furthermore, 1,658 samples from The Cancer Genome Atlas (TCGA) [48] were added. The only concern with the TCGA dataset is the absence of circRNA amplification experiments, e.g. circRNA-seq entries, making the circRNA counts in TCGA datasets unusable in the circRNA-miRNA part. The collected RNA-seq datasets are summarized in Table 1.

Mature miRNAs were annotated according to the miRBase [49] database. The mRNAs were annotated according to the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) [50] gene symbol convention. The circRNA mapping was taken from [51], where the raw data were mapped to the GRCh37/hg19 genome and annotated according to the circBase [52] database. Lastly, only those circRNAs with at least one entry in circInteractome [53] were annotated.

To partially mitigate the effects of heterogeneity, including differences in raw read counts and gene lengths across individual RNAs, we employed two straightforward normalization methods: Counts per Million (CPM) and Transcripts Per Million (TPM). The pre-normalized value distribution is shown in Fig. 2.

In subsequent experiments, we employ two collections of RNA-seq data. The first collection consists of 434 samples from MDS, RNA Atlas, and BT, which include all three RNA types: circRNA, miRNA, and mRNA. The second collection comprises 2,092 samples from all five datasets, containing only miRNA and mRNA. The larger collection is used primarily to empirically assess the impact of increasing sample size, while the smaller collection is applied to the clinically relevant task of inferring novel circRNA-miRNA interactions, a scenario where the algorithm is most needed.

Experimentally validated prior knowledge collection

Next, since the RNA-seq profiling is limited in number in case of circRNA or any other emerging field of the RNA-seq analysis, the sample size of the RNA-seq could not be increased indefinitely. In practice, the use of RNA-seq samples only could not achieve

Table 1 Comparison of all the RNA-seq datasets collected

Dataset	GEO/BioProject ID	Samples	circRNAs[1]	miRNAs[1]	mRNAs[1]
MDS [54]	PRJNA896500, PRJNA679200	77	3,009	2,507	17,232
RNA Atlas[55]	GSE138734	294	2,242[2]	1,646	17,903
BT [56]	GSE236933, GSE236932	63	2,150[2]	2,349	18,784
TCGA-COAD [48]	—–	460	1	2,142	19,178
TCGA-BRCA [48]	—–	1,198	36	2,221	19,223
Total table[3]	—–	2,092	3,009	2,479	18,778

[1] Numbers represent RNAs with at least one non-zero raw read count in the dataset

[2] These circRNAs were mapped from the chromosome location (chr:start end) to the circBase ID (hsa_circ_ID) [52] (http://www.circbase.org/) using appropriate human genome version mapping

[3] Total numbers after removing RNAs with zeroes in all datasets

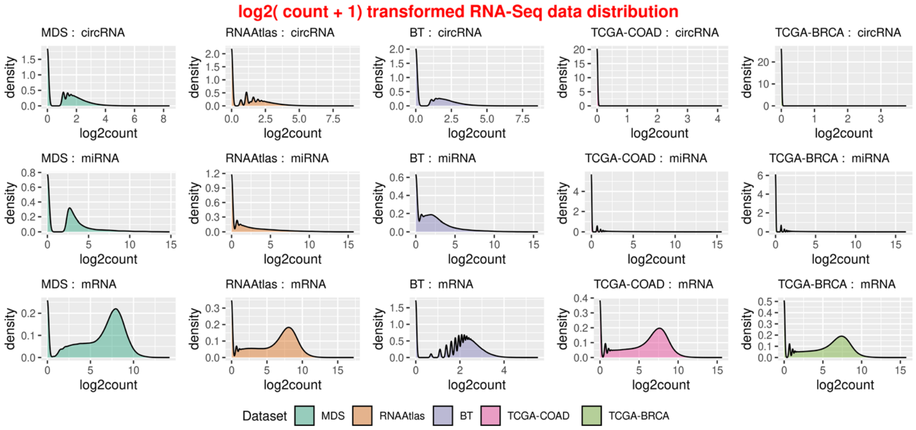


Fig. 2 A density plot of all datasets raw read counts for three types of RNAs in the ceRNA network. Log-scale transformation is applied to visualize exponentially differing counts only

Table 2 A table of prior knowledge databases used

Source	Interaction type	Num entries	Precision	Recall	Method
TarBase v9 [12]	miRNA-mRNA	1,292,326	100 % (89%) [1]	100% [1]	Experimental in-vitro
multiMiR[1] [57]	miRNA-mRNA	732,112	34%[1]	19%[1]	Experimental in-vitro
circInteractome[53]	circRNA-miRNA	55,148	12.5%[58] [2]	25%[58][2]	TargetScan in-silico

[1] TarBase v9 prior knowledge is used as a ground truth reference for the multiMiR evaluation. Formally, the reference is assumed to be perfectly correct, as reported by the 100% entries. However, realistically, it is limited by biological evaluation methods, such as PAR-CLIP, which are estimated to have an approximately 11% false positive (FP) rate [59]. Thus, the second value in brackets corresponds to this estimation

[2] Performance on miRNA-mRNA experiments in [58] is used as a reference

any reasonable result by itself. To overcome such issue, the RNA-seq data is enhanced with the additional (prior knowledge) sources of interaction information.

First type of these prior knowledge sources are the known experimentally validated interactions databases. The most recent and largest of them, TarBase v9 [12], was used as a ground-truth, whereas its previous version TarBase v8 with several additional databases, combined into a single multiMiR [57] package, was provided as an input to the algorithm. A summary of these prior knowledge databases is presented in Table 2.

TargetScan prior source and its performance evaluation

Another source of prior knowledge is synthetic computational algorithm predictions. In this study, biochemical alignment algorithms (e.g., TargetScan and miRanda) are considered. This is because the circInteractome database [53] has been regarded as one of the main reference sources for circRNA-miRNA interactions in several studies [7, 10, 60, 61]. Consequently, in the absence of experimentally validated information, predictions from these algorithms are the only available prior source.

Evaluating the performance of computational tools such as TargetScan and miRanda for predicting circRNA-miRNA interactions is challenging due to the recent discovery of these interactions [7] and the lack of validated references. One approach to evaluate these tools is to compare their performance on miRNA-mRNA interactions using available experimentally validated data and then extrapolate the results to the circRNA-miRNA case.

Several studies have taken different approaches to this evaluation. One study compared the predictions with the miRTarBase database [58]. Another study used an experimental reference dataset and measured changes in protein levels in upregulated or downregulated miRNAs to calculate precision values [20, 62]. A third study assessed precision using the positive predictive value (PPV) by comparing predictions with interactions from miRTarBase [58].

The results of these evaluation studies are summarized in Table 3. In subsequent experiments, a reference performance of 12.5% precision and 25% recall is used as the benchmark level for the circRNA-miRNA part of the network. To our knowledge, no experimentally validated database with sufficient entries is currently available. The main reason for choosing the evaluation in [58] is that it employs non-canonical miRNA-mRNA interaction types that are most similar to those observed in the circRNA-miRNA case.

Experimentally validated circRNA-disease databases

Finally, when experimentally validated interactions are insufficient or unavailable-such as for circRNA-miRNA interactions within the tripartite ceRNA network-an indirect evaluation approach can be employed. Instead of relying solely on limited experimental structural evidence, we utilize a broader set of databases that provide robust experimental support for circRNA-disease relationships. These databases include evidence from in vitro knockdown studies, qRT-PCR, and microarray expression analyses.

Importantly, experimental validations predominantly favor entries from the circInteractome/TargetScan databases [15, 53]. This bias arises because the candidate set of potential interactions tested in these studies is often generated using TargetScan [63]. As a result, many validated interactions are skewed toward the circInteractome prior rather than reflecting an objective evaluation.

For instance, consider the reference experiment presented in [64] from Circ2Disease [65]. The study first identified a set of circRNAs that are differentially expressed in hepatic steatosis. Subsequently, the researchers constructed a network comprising TargetScan-derived circRNA-miRNA interactions and the miRDB 5.0 experimentally

Table 3 Different TargetScan performance evaluation methods

Source	Reference validation	TargetScan	TargetScan PPV	TargetScan recall
2009 [20]	pSILAC experiment	v5.0	0.51	0.12
2016 [58]	miRTarBase v6	v7.0	0.125[2]	0.25[2]
2020 [62][4]	miRWalk2	v6.2	0.8	0.7
2020 [62][4]	miRDIP high conf. experiment	v6.2	0.9	0.26
2022 [21]	CLASH experiment	v7.2	0.95	0.12
2022 [21]	CLEAR-CLIP experiment	v7.2	0.99	0.23
2022 [21]	miRTarBase v8	v7.2	—[3]	0.76 (0.3)[1]

[1] Values in brackets represent non-functional target sites of miRNAs

[2] Values were computed according to supplementary Figure SD 4, average across the UTR length range is used

[3] miRTarBase provides only positive examples so the authors did not provide any precision evaluations

[4] Supplementary Table S21 provided in [62] used for reference

validated miRNA-mRNA interactions. An automated gene annotation tool was then used to generate a set of connected genes along with their functional annotations from this network. Within this set, a specific gene was demonstrated to play a key role in hepatic steatosis. Although this discovery is confirmed, the reliability of the approach depends directly on the quality of the circRNA-miRNA network and it certainly favours the predictions from TargetScan.

A summary of all collected circRNA-disease associations and their respective databases is presented in Table 4.

Since both circRNA-disease associations and computational outputs are available for the circRNA-miRNA part of the network, we designated one as the input and the other as the reference. Specifically, the circInteractome TargetScan data-provided in a direct structural form-was selected as the input prior knowledge. In contrast, circRNA-disease pairs, which are provided in an implicit form, are used solely as a reference for evaluating the predicted structure. The exact procedure of indirect evaluation is described later in Sect. Indirect structure evaluation via disease annotation.

Regression framework

On the basis of the problem statement, we assume a Bayesian network decomposition term to follow the Negative Binomial distribution owing to its overdispersed count-data nature:

$$\begin{gathered}\mathcal{P}(\boldsymbol{r}^i \mid Pa(\boldsymbol{r}^i)) \sim \text{NegativeBinomial}(\boldsymbol{\mu}^i, \boldsymbol{\phi}^i)\\ \text{where } \boldsymbol{\mu}^i \text{ is a mean, and } \boldsymbol{\phi}^i \text{ is a dispersion parameter}\end{gathered} \tag{6}$$

We then model the Generalized Linear Model (GLM) regression as the log-linear relationship of its mean parameter to the parent set:

$$\begin{gathered}\boldsymbol{\mu}^i = \exp\left[\beta_{i0} + \sum_{\boldsymbol{m}^k \in Pa(\boldsymbol{r}^i)} \beta_{ik} \log \boldsymbol{m}^k\right]\\ \text{Shorthand notation: } \boldsymbol{r}^i \sim Pa(\boldsymbol{r}^i)\end{gathered} \tag{7}$$

where log is the natural logarithm. Note that the logarithm and all subsequent transformations are applied to predictors only, the dependent variable remains as a raw read count.

Table 4 A table of experimentally validated circRNA-disease associations

Name	CircRNAs[1]	Diseases[2]
Circ2 disease [65]	5	11
CircR2Disease v2.0 [63]	57	36
circRNADisease v2.0 [17]	415	38
Total overlap	469	50

[1] Only the overlap with 3,009 previously collected CircRNAs is considered

[2] Only the overlap with 60 miRNA/mRNA diseases from GPACDA [66] is considered

Structure inference

One of the most convenient ways to perform feature selection in the GLM setting is to use all possible parents $\boldsymbol{m}$ and apply a penalty to coefficient values, which encourages sparsity by forcing some coefficients to zero. This penalized regression approach has been applied in several studies [39, 40, 42] in the form of the LASSO [40] L_1 penalty or ElasticNet [41] combination of L_1 and L_2 penalties. In the case of the LASSO penalty, the previous GLM framework can be extended as:

$$\begin{aligned} \boldsymbol{\mu}^i &= \exp\left[\beta_{i0} + \sum_{\boldsymbol{m}^k \in \boldsymbol{m}} \beta_{ik} \log \boldsymbol{m}^k\right] \\ &\arg\min_{\boldsymbol{\beta}}\ l_i(X, \boldsymbol{\beta}) + \lambda \sum_{\boldsymbol{m}^k \in \boldsymbol{m}} |\beta_{ik}| \\ &\quad \text{Short-hand notation: } \boldsymbol{r}^i \sim \boldsymbol{m} \end{aligned} \tag{8}$$

where $l_i(X, \boldsymbol{\beta})$ represents the Negative Binomial loss (deviance), which is proportional to the log-likelihood. The final prediction output will then be a set of non-zero coefficients: $Pa(\boldsymbol{r}^i) = \{\boldsymbol{m}^k : \beta_{ik} \neq 0\}$.

However, owing to the heterogeneity and imbalance in the RNA-seq dataset, the range of values across different RNA types varies significantly across samples, which could lead to an issue where larger counts are driven to zero. To reduce this bias, a normalization procedure is introduced:

$$\begin{aligned} r_j^i \text{ or } m_j^k &= \text{Original raw read count of mRNA/miRNA} \\ o_j^k &= \text{Normalization factor of } \text{ k-th RNA and } \text{ j-th sample} \\ \widetilde{m}_j^k &= \frac{m_j^k}{o_j^k} = \text{Normalized expression level} \\ \boldsymbol{\mu}^i &= \exp\left[\beta_{i0} + \sum_{\boldsymbol{m}^k \in \boldsymbol{m}} \beta_{ik} \log \widetilde{\boldsymbol{m}}^k + \log \boldsymbol{o}^i\right] \end{aligned} \tag{9}$$

The normalization factors remain the same, e.g. $\boldsymbol{o}^k$, for the $\boldsymbol{m}^k$ miRNA variable in case of circRNA-miRNA regression. Also, the index is different for mRNAs, miRNAs and circRNAs and is denoted as k for simplicity.

This ensures that our pipeline benefits from both the original raw read count Negative Binomial distribution modeling, as described in [36, 37], and the advantages of normalized continuous-value approaches, where predictor coefficients are numerically stable, as seen in microarray studies such as [40]. Note that the normalization is generic and all CPM/TPM normalizations, as well as more advanced methods such as TMM [36] could be represented in this formulation by simply dividing the transformed count by the raw read one. Note that the normalization is implicitly applied to the dependent variable via offset $\boldsymbol{o}^i$ as well.

Furthermore, this approach allows for straightforward incorporation of prior knowledge. In this context, prior knowledge refers to external information about the relationships between the candidate predictors $\boldsymbol{m}^k$ and the target variable $\boldsymbol{r}^i$. To reflect the reliability of the external source, it is reasonable to adjust the penalty accordingly. In its simplest form, prior knowledge affects only the presence or absence of a relationship. Thus, we introduce two penalty hyperparameters: a larger penalty for candidates without prior knowledge, and a smaller penalty for those supported by prior knowledge. Formally:

$$\lambda_{ik} = \begin{cases} \lambda_{prior} \text{ if } \boldsymbol{m}^{\mathrm{k}} \in \mathrm{Pa}^{\pi}(\boldsymbol{r}^{\mathrm{i}}) \\ \lambda_{non-prior} \text{ otherwise} \end{cases} \text{s.t. } \lambda_{non-prior} \geq \lambda_{prior} \tag{10}$$

The LASSO objective function is then changed accordingly:

$$l_i(X, \boldsymbol{\beta}) + \lambda_{prior} \sum_{\boldsymbol{m}^k \in Pa^{\pi}(\boldsymbol{r}^i)} |\beta_{ik}| + \lambda_{non-prior} \sum_{\boldsymbol{m}^k \notin Pa^{\pi}(\boldsymbol{r}^i)} |\beta_{ik}| \tag{11}$$

We employed this approach via the open-source R library *mpath* [42] on our collected heterogeneous RNA-seq dataset. However, the direct solution of the LASSO, with or without prior knowledge, results in numeric divergence, such as infinite gradient overflow. This issue was found to be theoretical in nature rather than a flaw in the implementation [43, 44].

Specifically, the issue arises from the curvature of the exponential parameter space. Similar issues with the Negative Binomial LASSO were encountered in [44], further highlighting the instability caused by the curvature of the heterogeneous RNA-seq collection. This issue cannot be resolved via the coordinate-descent-based optimization algorithm commonly employed for minimization [43]. For reference, the percentage of hyperparameters that could not be estimated, even after multiple manual restarts with different initial coefficient values, is presented in Table 5.

Prior-incorporated generalized monotone incremental forward stagewise (Pi-GMIFS)

The alternative optimization algorithm, generalized monotone incremental forward stagewise (GMIFS), was used. Instead of direct optimization of the GLM objective function for a given λ via a second-order derivative, gradient optimization via the constant step is performed. The algorithm then only follows the LASSO solution path and does not expand to numerically unstable outer regions. The drawback of this simplification is the limited model capability, as the current GMIFS model is equivalent to the constrained monotone version of the LASSO algorithm, not allowing for ElasticNet-like L_2 penalty addition.

Another potential source of the divergence may be the numerically unstable second derivative, as discussed in [67]. To address this, GMIFS uses an expanded space instead of directly computing the second derivative, as is done in coordinate descent. For each penalized parameter β_i, GMIFS introduces two parameters – β_i^+ and β_i^-,

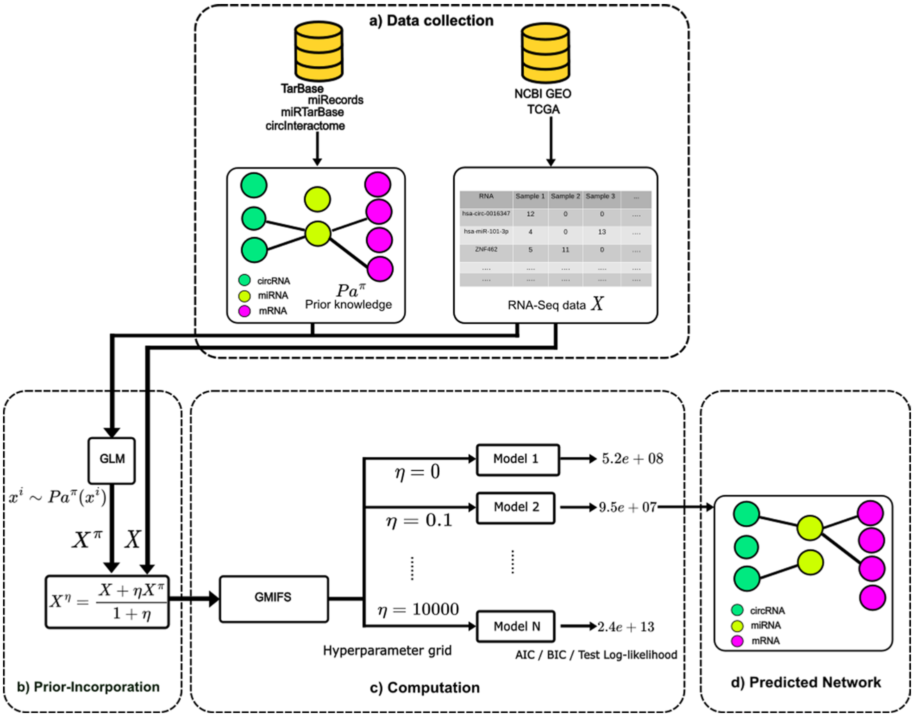
Table 5 Numeric issues of the *mpath* library [42] with penalized GLM applied to the collected RNA-seq dataset

Model	Total hyperparameters	Successfully converged	Convergence
NegBin LASSO (97 mRNAs)	5335	894	16.7%
Poisson LASSO (97 mRNAs)	5335	1834	34.38%
NegBin ElasticNet (2 mRNAs)	550	248	45.09%

each with different signs and coefficient estimations. Next, it performs small gradient steps with the most correlated signed parameter in contrast to the optimal step size computed via the second derivative. After the iterative fitting process, the resulting variable value is derived as $\beta_i = \beta_i^+ - \beta_i^-$.

The primary challenge with the GMIFS approach lies in incorporating prior knowledge. In previous algorithms, it was possible to directly modify the penalty values; however, in GMIFS, the hyperparameter grid is absent and is replaced by the gradient step and tolerance threshold. Then, the solution path generated during the iterative procedure approximates the path that would be obtained using an infinitely fine hyperparameter grid. The optimal value is then selected along this solution path based on the selected criterion. Consequently, an alternative method of prior incorporation, first proposed in [46], is used. The approach, named Pi-GMIFS, can then be summarized in the diagram in Fig. 3.

The key concept is response mixing: instead of directly modifying the hyperparameters of the prior and non-prior coefficients, a perfectly linearized response value of prior-parent-only non-penalized regression is mixed with the actual input data to force the algorithm to use the prior parents. Such a modification introduces a new hyperparameter η, which is selected according to the predefined hyperparameter grid and the selection is performed via four alternative metrics – BIC, AIC, train log-likelihood and test-set log-likelihood. For each criteria we take a maximum criterion value observed on the solution path for a given η.



RNA	Sample 1	Sample 2	Sample 3	...
hsa-circ-0016347	12	0	0	....
hsa-miR-101-3p	4	0	13	....
ZNF462	5	11	0	....
....	....	....	....	....
....	....	....	....	....

Fig. 3 An illustration of the Pi-GMIFS algorithm. The prior-incorporation step is the extension on top of the GMIFS algorithm. The hyperparameter grid for a mixture of prior-incorporation response value and original response value, in this case, is selected on a log scale between 0 and 10,000

The AIC, BIC and test-set log-likelihood (in form of cross-validation) criteria are taken from the original GMIFS algorithm from [45]. The train log-likelihood is added as a non-regularized form of AIC and BIC. The proposed algorithm is summarized in Algorithm 1.

The only detail to be discussed is the sample division. For the subsequent experiments we selected a 95:5 train-test split for the test-set log-likelihood criterion. The division is performed uniformly across all samples, ensuring that each dataset (MDS, BT, RNA Atlas) has its samples represented in both the training and test sets.

All other criteria that do not require the test set are given the same training set as the test-set-based criterion. The reason is computational, using a different training set for the test-set-based criterion would necessitate running an entire separate GMIFS computation pipeline for a single criterion. The impact of additional samples is also of a particular interest in the studied setup, as a lack of performance improvement would indicate that the sample size is already sufficient.

Algorithm 1 Prior-incorporated GMIFS

```
Ensure:
    $X$ – Input RNA-seq data matrix
    $Pa^{\pi}$ – Input prior interactions extracted from prior knowledge database
    $\eta$ – A hyperparameter controlling the truthfulness of prior, $\eta \geq 0$
for $\boldsymbol{r}^i \in \boldsymbol{r}$ do
    Fit an unpenalized GLM regression using only prior parents: $\boldsymbol{r}^i \sim Pa^{\pi}(\boldsymbol{r}^i)$
    Extract estimated response values $\boldsymbol{r}^{i,\pi}$
    Interpolate between estimated prior response and the input data response: $\boldsymbol{r}^{i,\eta} = \frac{(\boldsymbol{r}^i + \eta \boldsymbol{r}^{i,\pi})}{1+\eta}$
    Run a penalized GMIFS regression with unchanged predictor data $\boldsymbol{r}^{i,\eta} \sim \boldsymbol{m}$
    Select optimal model according to one of 4 criteria (AIC, BIC, train log-lik., test log-lik.)
end for
```

Indirect structure evaluation via disease annotation

We additionally propose an alternative approach to evaluating the quality of predictions, rather than directly comparing them to a ground-truth reference. Our method involves running an automated disease annotation pipeline that uses the predicted structure as input. If the structure is improved, the accuracy of the secondary annotation task is also expected to increase.

The key reason for the introduction of this alternative evaluation method is the late discovery of circRNAs. Compared with the well-established mRNA and miRNA databases, experimentally validated circRNA-miRNA interactions remain scarce, making their use impractical. Instead, a much more extensive collection of circRNA-disease associations, as shown in Table 4, was used to indirectly evaluate the credibility of the inferred structure, hence the term “indirect evaluation”.

To systematically compare large-scale networks, we adopted the framework from circGPA/GPACDA [60, 66]. This approach allows us to annotate and evaluate each selected circRNA and its associated disease terms in a structured manner and in polynomial time.

In contrast to the aforementioned framework, several approaches have been explored for circRNA-disease annotation. Benchmarking results from [68] indicate that deep learning-based methods outperform simpler techniques such as circGPA/GPACDA.

Nevertheless, among the evaluated deep learning alternatives [69–72], only [72] addresses the problem within the context of a tripartite circRNA-lncRNA-miRNA network. However, the fact that the exact circRNA-miRNA-mRNA network is not directly utilized necessitates additional training of these networks for our specific problem. Given that the indirect approach serves solely as a secondary validation task, we have opted for a simpler tool, although deep learning models could potentially enhance the accuracy of the indirect evaluation.

In this evaluation, the GPACDA computed the annotation score and its p-value for each circRNA-disease pair. The score was derived from the interaction network of the corresponding circRNA and was based on known associations between the disease and mRNA, as well as the disease and miRNA. The GPACDA tool works with the assumption of *guilt-by-association* [60, 73] which suggests that interacting RNA molecules are likely to have similar functions or roles. The resulting scores were evaluated against the known circRNA-disease associations introduced in Sect. Experimentally validated circRNA-disease databases that remained hidden during the automated annotation. The process of indirect evaluation is summarized in Fig. 4.

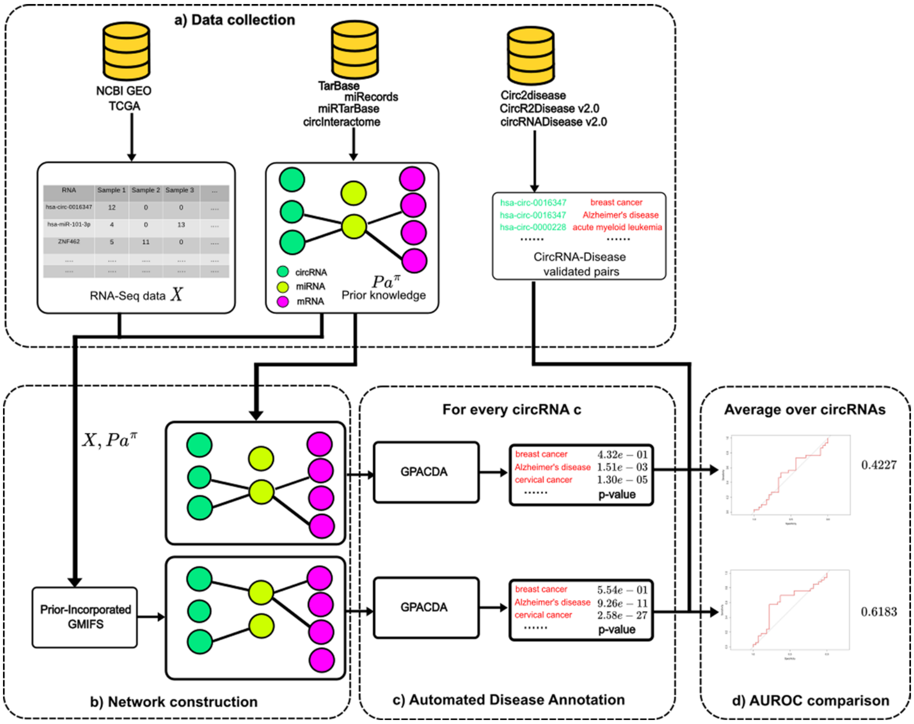


Fig. 4 An illustration of the indirect evaluation pipeline. The GPACDA annotation tool is run with different input interaction networks, and the quality of the resulting circRNA-disease annotations is compared among the networks

Since the annotation task is different from the original annotation with a fixed validated network [60, 66], the modification of score statistics, named as *varying network statistics score*, was newly implemented into the GPACDA tool. The difference is that instead of the annotation, the network structure is varying. The detailed implementation is available in Appendix B.

Area under the receiver operating characteristic curve (AUROC) metric for indirect structure evaluation via disease annotation

The primary challenge in evaluation lies in selecting the appropriate metric for the generated statistics and their corresponding p-values on the basis of the validated pairs. The absolute values of the statistic and p-value are impractical; they are inversely related to the number of edges. As the network density for a given circRNA increases, the p-value decreases. This could be illustrated by the relationship $pvalue_{\|\mathbf{a}_i^{m,c}\|_1}(s(c^i, \mathbf{a}_i^{m,c})) \propto \frac{1}{\|\mathbf{a}_i^{m,c}\|_1}$ (see Appendix B for the detailed notation). Consequently, only the p-values for all diseases of a single circRNA are comparable because they are of the same order of magnitude given by same network part, allowing for robust comparisons within a single circRNA. Thus, rather than evaluating absolute p-values across circRNAs, the analysis focuses on the sorted order of terms on the basis of p-values for a single circRNA. If a specific circRNA has validated associations with a set of diseases, the goal is to predict lower p-values for this set than for the remaining diseases.

A suitable criterion for assessing the order of p-values, rather than their absolute values, is the Area Under the Receiver Operating Characteristic (AUROC) [74]. The AUROC metric operates similarly to the F1-score, specifically by calculating the True Positive Rate (TPR) and False Positive Rate (FPR), as defined in Eq. 12.

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ FPR &= \frac{FP}{FP+TN} \end{aligned} \tag{12}$$

The computation involves defining a varying decision threshold. Prediction scores below this threshold are classified as negative predictions, whereas those above are classified as positive. The estimated area under the TPR-FPR curve reflects the strength of the prediction [74].

In our AUROC context, the predicted score is defined as $1 - pvalue$, ensuring values range between 0 and 1. A higher score indicates a lower p-value and a greater likelihood of the annotation term. Positive predictions involve associating a correct disease-circRNA association with a $(1 - pvalue) < threshold$. The validated pairs provide the ground truth for computing the AUROC metric. The implementation of AUROC computation is performed via the **pROC** open-source R library [75].

Results

In this section, the proposed algorithm is experimentally evaluated. First, we demonstrate a direct evaluation using a known ground-truth reference. Next, we present an indirect evaluation for cases where the true underlying structure is unknown.

miRNA-mRNA interaction prediction with F1-score

The first experiment uses Pi-GMIFS to predict miRNA-mRNA interactions. The interactions recorded in TarBase v9, the latest experimentally verified interaction database, are taken as the ground truth. Pi-GMIFS algorithm takes as the input a set of high-throughput RNA-seq data alongside with prior knowledge on miRNA-mRNA interactions. In the experiment, the size of the RNA-seq dataset as well as the quality of the prior will be changed. We will work with two different RNA-seq sample sizes and MultiMir and synthetic priors. The goal is to approach the ground truth, this includes determining how many novel interactions were correctly identified and how many outdated interactions were successfully removed. To the best of our knowledge, no existing algorithm can compute the unstable RNA-seq dataset alongside the prior. Therefore, the prior level was designated as the benchmark baseline to be surpassed.

We use the F1-score as the score function $S()$ from the problem statement, which reflects the harmonic mean of precision and recall. Unlike the more straightforward Structural Hamming Distance (SHD) metric for structural assessment, the F1-score provides a balanced evaluation for both dense and sparse networks [76]. The equation for the metric is as follows:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{F1-Score} &= \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \end{aligned} \tag{13}$$

where TP, FP and FN represent true positive, false positive and false negative interactions, respectively.

For the computations, the Pi-GMIFS was configured as follows:

- $\tau = 10^{-5}, \epsilon = 10^{-4}$ (Default GMIFS initialization from [45])
- Negative Binomial distribution, the dispersion parameter $\boldsymbol{\phi}^i$ is estimated via Hilbe's method of moments as in [45]
- $\eta \in \{10^{-1}, 0, 1, 10, 10^2, 10^3, 10^4\}$, $I_{max} = 3000$ (For hyperparameter η refer to Sect. Regression framework, for hyperparameter I_{max} refer to Appendix A)

Instead of applying the pipeline to the whole miRNA-mRNA network, only the top 100 mRNAs with the greatest increase in the number of novel unique miRNA-mRNA interactions between TarBase v9 and multiMiR were selected. Among these 100 mRNAs, only 97 had at least one non-zero raw read count across all samples. This reduction is due to the time complexity involved in processing the full set of 16,000 mRNAs. An overview of the first 10 entries with the greatest increase in novel interactions from multiMir to TarBase v9 is summarized in Table 6. The first row shows that multiMiR achieved approximately 32% precision and 5% recall for MACF1 when evaluated against TarBase v9. When averaged across all the multiMiR mRNAs, they align with the values presented in the second row of Table 2. When the percentages are averaged over the 97 selected mRNAs, precision and recall correspond to the values reported in the first row of Table 7.

Table 6 The 10 mRNAs with the greatest increase in novel unique miRNA-mRNA interactions from multiMir (TarBase v8 and other databases) to TarBase v9

HGNC symbol	multiMir entries	TarBase v9 entries	Novel entries	Obsolete entries
MACF1	66	413	392	45
DYNC1H1	165	481	372	56
HUWE1	195	473	344	66
PRPF8	145	431	333	47
SPTBN1	140	413	315	42
MAP1B	241	466	307	82
CLTC	215	462	304	57
BIRC6	135	403	302	34
ZBTB20	151	346	296	101
UBR4	161	392	295	64

MultiMir prior

The first miRNA-mRNA experimental run aims to evaluate the performance of the Pi-GMIFS algorithm when it is applied to high-throughput RNA-seq data alongside the multiMiR prior, a precise prior with a relatively low recall. This experiment predominantly test the ability of Pi-GMIFS to infer hidden valid interactions.

Initially, the expression data comprised 434 GEO RNA-seq samples. The results are depicted in the top left plot of Fig. 5. Although none of the criteria and hyperparameter values resulted in an outcome surpassing the F1 threshold set by the multiMiR prior, this limitation is attributable to the sample size. When additional 1,658 samples from the TCGA database were incorporated into the RNA-seq data, the algorithm outperformed the multiMiR prior, as illustrated in bottom left plot in Fig. 5. Notably, with the increased sample size, the maximum test-set log-likelihood criterion proved less effective, though CPM normalization again emerged as a more suitable choice than TPM normalization method.

TargetScan synthetic prior

To estimate the performance of Pi-GMIFS with a suboptimal TargetScan prior, such as circInteractome, it is necessary to adjust the prior quality to align with benchmarks reported in [20, 58]. In our experiment, we selected a prior level of 0.25 recall and 0.125 precision, as described in [58]. This selection is based on the fact that circRNA-miRNA interactions were not considered in the original TargetScan alignment model, creating a bias towards conserved seed regions of miRNA [21]. Additionally, circInteractome specifically estimates interactions for a limited subset of miRNAs (193 out of 2,656), which potentially reduces its recall if the “80-20 rule” is not applied [60].

A synthetic prior was constructed to incorporate these aspects. First, the adjacency matrix of TarBase v9 was used, from which edges were subsampled, and some false edges were randomly introduced. The number of removed and added edges was adjusted to achieve a synthetic prior with 0.25 recall and 0.125 precision. Furthermore, a constraint was applied such that edges in the synthetic prior came from a restricted subset

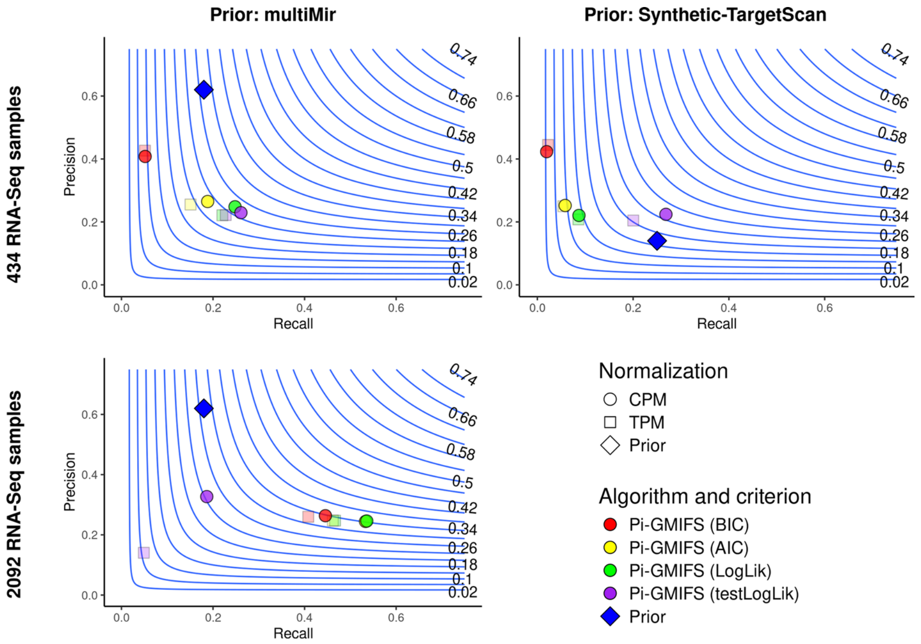


Fig. 5 Precision-recall plot with F1 isolines. The Pi-GMIFS scores are shown for two priors: multiMiR (TarBase v8, etc.) in the left column and a synthetically subsampled TarBase v9, designed to imitate TargetScan performance, in the right column. Two RNA-seq sample collections were used: 434 samples in the first and 2,092 samples in the second row. The numbers near the contour lines correspond to F1-score levels. Only the CPM normalization method is shown, as it outperformed the other TPM normalization method. Each model selects the optimal hyperparameter η on the basis of 4 criteria (AIC, BIC, train log-lik, and test log-lik). The choice of η is made for each RNA node separately. The experiment with 2,092 RNA-seq samples using the synthetic TargetScan prior was not performed for its redundancy, see the explanation in Sect. Performance in miRNA-mRNA experiments

of mRNAs. In particular, only 40 mRNAs had at least one prior interaction, while the remaining mRNAs were empty, thus simulating the circInteractome.

The synthetic prior, along with the same 434 GEO RNA-seq samples, was then input into Pi-GMIFS. The results are summarized in the top right plot of Fig. 5. The use of CPM normalization, and the test-set log-likelihood criterion resulted in an improvement of two F1 levels over the prior.

Performance in miRNA-mRNA experiments

A technical overview of all the experiments is shown in Table 7. Pi-GMIFS demonstrated its ability to refine erroneous prior knowledge and improve its quality in terms of the F1-score. Specifically, Pi-GMIFS outperformed the multiMir prior (0.62 precision, 0.18 recall, 0.28 F1-score) when 2,092 RNA-seq samples were used, achieving a precision of 0.24, a recall of 0.54 and an F1-score of 0.34. The performance metrics indicate that Pi-GMIFS successfully detected numerous new interactions while maintaining a precision level that exceeds the typical precision achieved by computational methods. However, 434 RNA-seq samples were insufficient to surpass the prior. In contrast, Pi-GMIFS outperformed the synthetic prior (0.125 precision, 0.25 recall, 0.17 F1-score) even with 434 RNA-seq samples, improving both precision (0.22) and recall (0.27) to achieve an F1-score of 0.24.

Table 7 Summary of all 97-mRNA experiments

Algorithm	Prior	Sample size	Precision [1]	Recall [1]	F1 [1]
Prior only	multiMiR	—	0.62	0.18	**0.28**
Pi-GMIFS, CPM, AIC	multiMiR	434	0.26	0.19	0.22
Pi-GMIFS, CPM, BIC	multiMiR	434	0.40	0.05	0.09
Pi-GMIFS, CPM, log-likelihood	multiMiR	434	0.25	0.25	0.25
Pi-GMIFS, CPM, test log-likelihood	multiMiR	434	0.23	0.26	0.25
Pi-GMIFS, TPM, AIC	multiMiR	434	0.25	0.15	0.19
Pi-GMIFS, TPM, BIC	multiMiR	434	0.43	0.05	0.09
Pi-GMIFS, TPM, log-likelihood	multiMiR	434	0.22	0.22	0.22
Pi-GMIFS, TPM, test log-likelihood	multiMiR	434	0.22	0.23	0.22
Prior only	multiMiR	—	0.62	0.18	0.28
Pi-GMIFS, CPM, AIC	multiMiR	2092	0.25	0.54	0.34
Pi-GMIFS, CPM, BIC	multiMiR	2092	0.26	0.45	0.33
Pi-GMIFS, CPM, log-likelihood	multiMiR	2092	0.24	0.54	**0.34**
Pi-GMIFS, CPM, test log-likelihood	multiMiR	2092	0.33	0.19	0.24
Pi-GMIFS, TPM, AIC	multiMiR	2092	0.25	0.46	0.32
Pi-GMIFS, TPM, BIC	multiMiR	2092	0.26	0.41	0.32
Pi-GMIFS, TPM, log-likelihood	multiMiR	2092	0.25	0.47	0.32
Pi-GMIFS, TPM, test log-likelihood	multiMiR	2092	0.14	0.05	0.07
Prior only	TargetScan[2]	—	0.125	0.25	0.17
Pi-GMIFS, CPM, AIC	TargetScan	434	0.25	0.06	0.09
Pi-GMIFS, CPM, BIC	TargetScan	434	0.42	0.02	0.04
Pi-GMIFS, CPM, log-likelihood	TargetScan	434	0.22	0.09	0.13
Pi-GMIFS, CPM, test log-likelihood	TargetScan	434	0.22	0.27	**0.24**
Pi-GMIFS, TPM, AIC	TargetScan	434	0.25	0.06	0.09
Pi-GMIFS, TPM, BIC	TargetScan	434	0.44	0.02	0.04
Pi-GMIFS, TPM, log-likelihood	TargetScan	434	0.21	0.09	0.12
Pi-GMIFS, TPM, test log-likelihood	TargetScan	434	0.21	0.20	0.20

For each input setup, the best GMIFS hyperparameter choice performance is highlighted as bold along with its F1-score. The values are rounded to 2 digit precision

[1] Performance computed with the assumption of TarBase v9 being the ground truth

[2] A TargetScan prior was constructed by synthetically subsampling the TarBase v9 network edges to have the same properties as TargetScan estimation from [58] with 0.125 precision and 0.25 recall

Note that the experiment combining the 2,092 RNA-seq sample collection with the synthetic TargetScan prior was not conducted. This decision was based on the observation from the experiment with MultiMir, 2,092 RNA-seq samples, and the $\eta = 0$ configuration, which excluded any prior information. In this RNA-seq-only case the performance of the model dropped slightly–for example, the best train log-likelihood had a precision of 0.18, a recall of 0.51 and an F1-score of 0.26, which already outperformed the TargetScan prior F1-score level of 0.17. In other words, Pi-GMIFS with a sufficiently large RNA-seq sample set overcomes the synthetic prior even without incorporating any prior information. The interaction prediction based purely on RNA-seq data works better than computationally-based interaction prediction with TargetScan.

Finally, an additional analysis of proposed η hyperparameter on the performance has been evaluated in Appendix C.

Comparison with other methods

In order to benchmark the performance of Pi-GMIFS, we compared it against several alternative frameworks using an identical dataset comprised of 434 RNA-seq samples. Our evaluation spans methods implemented in the *glmnet* library [77], the SPONGE framework [78], and various correlation-based approaches from the *miRLAB* R library [79]. Each of these methods processes log CPM-transformed data (Poisson uses non-logarithmic CPM) to ensure consistency in data representation among different input formats.

A major limitation of the *glmnet* library is that it does not offer a penalized Negative Binomial family implementation; instead, its closest alternative is Poisson regression. While Poisson regression can serve as an approximation for overdispersed count data under specific conditions, it inherently lacks the flexibility of a negative binomial model in accurately capturing the overdispersion characteristics typically observed in RNA-seq experiments. To overcome numerical instability in this context, glmnet employs a step-halving technique [77]. Although this approach enhances stability by adjusting the optimization step sizes in case of infinite gradient detection, it does not fully overcome divergence issues linked to the curvature of the second derivative of the loss function.

Notably, when running *glmnet*'s LASSO without incorporating prior information, the algorithm achieved a convergence rate of more than 95%, rendering it a useful reference model, especially in comparison to a much lower *mpath* convergence rates from Table 5. Any prior incorporation decreased this percentage under the 95% and it was decided not to bechmark them.

In contrast, the theoretical framework of Pi-GMIFS is designed to inherently solve these convergence issues without any numerical techniques. This theoretical advantage becomes particularly critical in heterogeneous datasets where batch effects are significant and experimental evidence for novel RNA interactions, such as in the circRNA-miRNA setting, is scarce. The seemingly negligible difference in convergence rates increases with each additional few-sample RNA-seq experiment.

Another comparison concerns the *SPONGE* framework [78], which utilizes *glmnet* configured with the Gaussian family. By processing log CPM values – similar to methodologies traditionally employed for microarray data – SPONGE avoids the convergence challenges with count-based models. Despite this advantage, the Gaussian formulation sacrifices count-specific information from RNA-seq data, thereby reducing prediction accuracy. Thus, while SPONGE exhibits stability due to its reliance on a distribution that is less sensitive to divergence, this comes at the cost of losing the performance of RNA interaction modelling.

In addition to regression-based frameworks, we also evaluated several correlation-based approaches implemented in the *miRLAB* R library [79]. These methods employ measures such as mutual information, Pearson's/Spearman's correlation coefficients, and z-score-based metrics to quantify the associations between RNA pairs. While the simplicity and interpretability of these correlation metrics offer an accessible means for preliminary exploration of RNA interactions, their inability to discriminate between direct and indirect regulatory effects within the complex RNA networks reinforces the superiority of regression-based approaches in this context. Only the top 3 performance methods from *miRLAB* library were shown

Notably, all methods under evaluation – including the regression-based approaches and the correlation-based methods derived from the miRLAB R library-were executed using default parameter settings. In our experiments, we deliberately avoided extensive hyperparameter fine-tuning, particularly regarding the tolerance and learning step hyperparameters in the GMIFS component of Pi-GMIFS. This decision ensures that any performance differences are attributable to the inherent characteristics of the algorithms rather than to parameter optimization.

The use of more advanced methods, whether the Bayesian methods that are based on MCMC sampling or posterior integral computation [26, 80, 81] suffer from the scalability issue and could not model thousand of RNAs as needed in this context.

As summarized in Table 8, the performance of Pi-GMIFS is superior to all above-mentioned methods that are still runnable in the given setting.

circRNA-miRNA inference assessed with GPACDA

To validate the correctness of the structural inference for circRNAs and assess its performance on a clinically relevant task, GMIFS regression was applied in the circRNA functional annotation pipeline. The primary objective was to demonstrate that structure inference enhances the accuracy of automated circRNA-disease annotation. Again, the prior network is selected as the benchmark baseline to be surpassed.

circRNA-miRNA inference by Pi-GMIFS

Before the annotation process, the first step involved computing the circRNA-miRNA network via Pi-GMIFS, with the miRNA-mRNA subnetwork fixed to the experimentally validated TarBase v9. All circInteractome database entries were used as prior knowledge, and 434 RNA-seq samples from the GEO database were used for structure inference. The selection of the optimal η hyperparameter was performed using the maximum test set likelihood. This approach mirrors the investigation in the miRNA-mRNA section, where 434 RNA-seq samples (divided in a 95:5 ratio as mentioned in Sect. Regression framework) achieved the best performance in Synthetic-TargetScan prior with the test-set log-likelihood criterion.

After learning, the initial circInteractome network consisting of 55,148 interactions between 3,009 circRNAs and 193 miRNAs was expanded into the GEO network with 153,171 interactions between 1,791 circRNAs and 2,336 miRNAs. As a result, the network became less dense per miRNA, despite the total increase in the number of interactions.

Network evaluation by GPACDA

After the learning process, the indirect evaluation procedure introduced in Sect. Indirect structure evaluation via disease annotation was run. To address network scarcity nuances, only 237 circRNAs that had at least one known disease annotation and were connected to at least one miRNA in the estimated network were annotated and evaluated. In case when Pi-GMIFS returns an empty parent set for a non-empty prior set, the recommended procedure is to use the prior.

Table 8 Summary of the performance comparison between Pi-GMIFS and alternative inference frameworks

Algorithm	**Prior**	**Precision**	**Recall**	**F1**
Pi-GMIFS, CPM, AIC[1]	No Prior	0.26	0.14	**0.18**
Pi-GMIFS, CPM, BIC[1]	No Prior	0.44	0.05	0.08
Pi-GMIFS, CPM, log-likelihood[1]	No Prior	0.23	0.20	**0.22**
Pi-GMIFS, CPM, test log-likelihood[1]	No Prior	0.22	0.22	**0.22**
Sponge/Gaussian Elastic glmnet	No Prior	0.15	0.01	0.02
glmnet Poisson LASSO (AIC)[2]	No prior	0.21	0.16	0.18
glmnet Poisson LASSO (BIC)[2]	No prior	0.21	0.16	0.18
glmnet Poisson LASSO (logLik)[2]	No prior	0.21	0.16	0.18
glmnet Poisson LASSO (testLogLik)[2]	No prior	0.23	0.10	0.14
Mutual Information	No prior	0.14	0.04	0.07
Z-Score	No prior	0.15	0.05	0.08
Hoeffding	No prior	0.15	0.05	0.07
Pi-GMIFS, CPM, AIC	multiMiR	0.26	0.19	**0.22**
Pi-GMIFS, CPM, BIC	multiMiR	0.40	0.05	**0.09**
Pi-GMIFS, CPM, log-likelihood	multiMiR	0.25	0.25	**0.25**
Pi-GMIFS, CPM, test log-likelihood	multiMiR	0.23	0.26	**0.25**
Sponge/Gaussian Elastic glmnet	multiMiR	0.38	0.01	0.02
Mutual Information	multiMiR	0.14	0.05	0.07
Z-Score	multiMiR	0.14	0.05	0.07
Hoeffding	multiMiR	0.14	0.05	0.07
Pi-GMIFS, CPM, AIC	Synthetic-TargetScan	0.25	0.06	0.09
Pi-GMIFS, CPM, BIC	Synthetic-TargetScan	0.42	0.02	0.04
Pi-GMIFS, CPM, log-likelihood	Synthetic-TargetScan	0.22	0.09	**0.13**
Pi-GMIFS, CPM, test log-likelihood	Synthetic-TargetScan	0.22	0.27	**0.24**
Sponge/Gaussian Elastic glmnet	Synthetic-TargetScan	0.11	0.004	0.01
Mutual Information	Synthetic-TargetScan	0.20	0.07	0.10
Z-Score	Synthetic-TargetScan	0.20	0.07	0.10
Hoeffding	Synthetic-TargetScan	0.20	0.07	0.10

For each input setup, Pi-GMIFS was proven to be superior with some criteria than all other frameworks. Pi-GMIFS criteria that surpassed all other frameworks are highlighted in bold, along with their corresponding F1-score. All values are rounded to 2 digit precision

[1] A previous result with 434 RNA-seq samples was taken with $\eta = 0$ only

[2] *glmnet* library's Poisson LASSO computation resulted in 97% convergence rate, e.g., 940 out of 970 lambda hyperparameters converged in a given default maximum iteration limit. Notably, no mRNA had full non-convergence, thus, making this particular setup usable

The reached AUROCs

The results are presented in Fig. 6, which displays the annotation AUROCs obtained for the individual circRNAs from 9 different circRNA-miRNA subnetworks. The plot includes the prior network as well as the networks inferred using four different criteria (AIC, BIC, train log-likelihood and test log-likelihood). The networks labelled **(Prior+criterion)** represent networks constructed by combining both prior interactions and edges generated by Pi-GMIFS under the specified criterion.

The main conclusion is that Pi-GMIFS, which maximizes the test set likelihood (the setting already recommended from the previous mRNA-miRNA experiment), generates networks that improve the quality of disease annotations with respect to the circInteractome prior. The mean AUROC increased from 0.70 to 0.72. The prior was surpassed

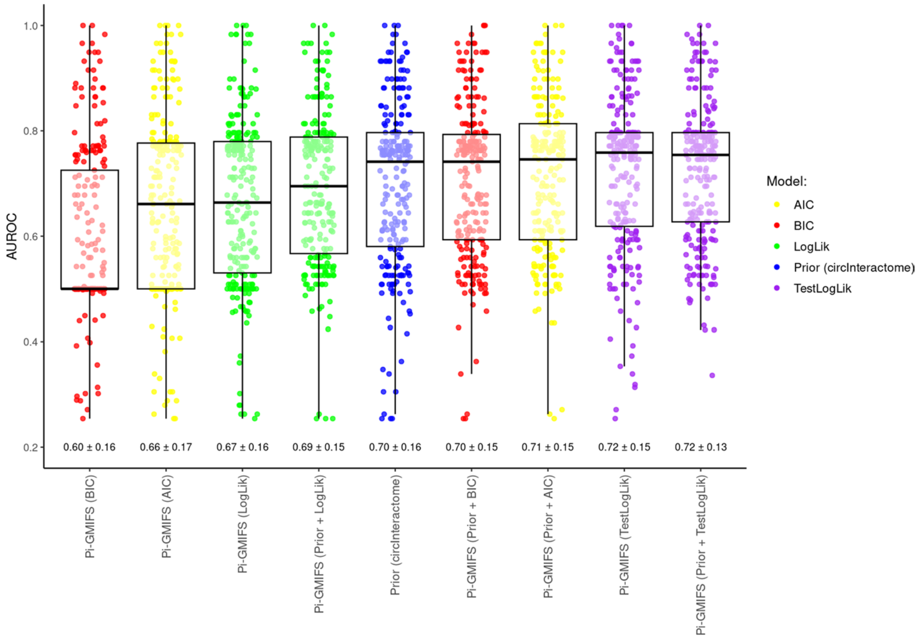


Fig. 6 A distribution of AUROC values for 239 out of 469 circRNAs that have at least one edge in any non-prior criterion model. The networks are sorted according to the average AUROC. The bottom values correspond to the mean and standard deviation. The absence of edges reduces the AUROC to 0.5, as indicated by the median line in the leftmost sparse BIC network

both when learning the structure from scratch and when extending the prior network. Consequently, the inference algorithm with this criterion successfully identified new interactions from the RNA-seq data that were not included in the prior set, and that are consistent with the existing knowledge on circRNA-disease associations.

Discussion

In the previous section, Pi-GMIFS was evaluated in two principled scenarios. First, we directly assessed the algorithm's performance using RNA-seq data by comparing the inferred miRNA-mRNA interactions with experimentally validated interactions. We employed the TarBase v9 as the gold standard interaction dataset. We experimentally confirmed that Pi-GMIFS improves with increasing RNA-seq sample size and it is important to collect a representative expression dataset. The minimum number of RNA-seq samples required increases with the quality of the prior. When dealing with a well-established interaction database such as MultiMiR, thousands of RNA-seq samples are needed. This suggests the practical applicability of the algorithm in real-world structure inference tasks. The given number of RNA-seq samples is undoubtedly available in public expression databases, although their number may decrease with increasing demands for completeness in terms of the representation of different types of RNA. The RNA-seq datasets used in our experiments are publicly available, see Sect. Conclusion.

The inferior synthetic prior, which is an estimate of circRNA-miRNA TargetScan performance, demonstrated that the Pi-GMIFS algorithm can be effective even for the circRNA-miRNA component of the network. However, owing to the limited number of experimentally validated circRNA-miRNA interactions, we proposed an alternative indirect testing approach for validation. This method allows researchers to assess the quality of predictions without relying on a direct ground truth network structure, particularly in more complex scenarios. The indirect testing approach involves evaluating the predictive performance of the estimated network within an RNA-disease functional annotation pipeline. The quality of disease prediction serves as an indirect metric to assess the structural accuracy of the network.

In terms of the second experimental scenario, we applied the Pi-GMIFS algorithm to a circRNA functional annotation task, where its performance efficiency was confirmed, as shown in Fig. 6. However, the algorithm's coverage was limited, as only 239 out of 469 circRNAs had at least one edge in the predicted network. On the other hand, the evaluation may have overestimated the quality of the circInteractome prior, since experimental research often uses TargetScan to generate potential interaction candidates, as shown in Sect. Experimentally validated circRNA-disease databases. Nevertheless, as demonstrated in the miRNA-mRNA experiment, an increase in sample size consistently improved performance and enhanced the statistical power of the interaction detection process.

The Pi-GMIFS algorithm effectively integrates prior knowledge and demonstrates its utility in large-scale RNA interaction networks. Our findings suggest that this enhanced algorithm offers a viable alternative to conventional in silico biochemical alignment methods, such as TargetScan, and has the potential to improve the performance of existing algorithms. Among the four evaluation criteria used (AIC, BIC, training log-likelihood, and testing log-likelihood), the test log-likelihood exhibited the best performance.

Compared with similar studies, the authors of [80] used the zero-inflated version of the Negative Binomial distribution to model the correlation, not the log-linear distribution. Additionally, their number of samples was at a maximum of 200 in comparison to the 400+ and 2,000+ samples used in our case. However, their approach has proven useful for use in transcription factor (TF) cases, where the amount of empirical evidence is as low as that in circRNA cases. Nevertheless, the correlation matrix approach, while being improved significantly, still has the drawback of being unable to distinguish between the direct and indirect effects of regulation [24, 81].

In [81], the authors proposed the use of variational inference to learn the approximation of the posterior distribution integral over all the parameters. Their approach was reported to outperform *mpath* [42] penalized regression, although it was shown in a specific time-series domain of Dynamic Bayesian networks, and is still reported to be used on the order of 40–50 nodes maximum. This limitation is not enough for the studied circRNA-miRNA-mRNA case.

Several extensions to the framework are planned, including batch correction to mitigate heterogeneity across multiple datasets, incorporation of advanced normalization techniques for predictor re-scaling, and the addition of a regularization term to the GMIFS optimization function to address multicollinearity, as in ElasticNet. Each of these modifications aims to address specific challenges encountered during the experiments and could improve overall performance, potentially reducing the minimum sample size required to outperform a given prior threshold.

Additionally, a much deeper review of the prior quality collected from the TarBase v9, multiMiR and other databases is planned. Not only that different miRNA-mRNA interactions are validated differently (directly or indirectly, via the protein products or via RNA itself, etc.), even the gold standard proof of the interaction is currently assumed to be a set of 2–3 experiments, every single one of which has a certain degree of uncertainty and false positive chance. Consequently, the unreliable ground truth and differently weighted prior quality input is a planned direction for future work.

A broader direction for future research involves exploring cyclic pathways or, in cases with unknown topological ordering, inferring the dominant regulatory axis within cyclic networks or allowing for cycles in the model.

Conclusion

In this study, we introduced an extension of the GMIFS algorithm, termed Prior-incorporated GMIFS (Pi-GMIFS), which is designed to solve a penalized Generalized Linear Model (GLM) regression. The proposed GLM framework leverages both the count-based nature of RNA-seq data used for target variables and the normalized values of parent RNAs used for predictor variables. The modified algorithm employs an iterative gradient approach and incorporates prior knowledge through response mixing [46], combining actual expression data with the estimated prior response.

We evaluated the Pi-GMIFS algorithm for RNA interaction refinement using RNA-seq data. Pi-GMIFS overcame the existing RNA interaction priors in terms of F1-score. Its performance improves with the amount of RNA-seq data, and our experiments demonstrated that the algorithm is applicable in real world scenarios, such as those involving expression sample sizes currently available in public databases. Overall, Pi-GMIFS effectively integrates prior knowledge and expression data, improving the reliability of both computationally predicted and experimentally validated ceRNA interaction databases.

Appendix A convergence criterion of Pi-GMIFS algorithm

Algorithm 2 Additional GMIFS convergence criterion

```
Ensure:
    $I_{max}$ – A maximum number of iterations allowed without exploration and exploitation
    $\tau$ – convergence threshold
    $\epsilon$ – gradient step size
procedure GMIFS
    $I \leftarrow 0$
    $L_{prev} \leftarrow$ the log-likelihood of the intercept-only model
    $L_{max} \leftarrow L_{prev}$
    $\boldsymbol{\theta}_{prev} \leftarrow (\boldsymbol{\beta} = \mathbf{0}, \phi =$ Hilbe's method of moments estimation$)$
    $n_{\mathrm{maxcoefs}} \leftarrow 0$
    for each iteration of GMIFS do
        Compute $L_{current}$ log-likelihood and $\boldsymbol{\theta}_{current}$ after $\epsilon$-step gradient
        $\boldsymbol{\beta}^{current} \leftarrow \boldsymbol{\beta} \in \boldsymbol{\theta}_{current}$
        $n_{\mathrm{coefs}} \leftarrow \sum_{i} [\boldsymbol{\beta}_i^{current} \neq 0]$
        if $|L_{prev} - L_{current}| \leq \tau$ then
            Convergence, terminate algorithm
        end if
        if $L_{max} \geq L_{current}$ and $n_{\mathrm{maxcoefs}} \geq n_{\mathrm{coefs}}$ then
            $I \leftarrow I + 1$
        else
            $I \leftarrow 0$
        end if
        $L_{max} \leftarrow \max(L_{max}, L_{current})$
        $n_{\mathrm{maxcoefs}} \leftarrow \max(n_{\mathrm{maxcoefs}}, n_{\mathrm{coefs}})$
        if $I \geq I_{max}$ then
            Terminate algorithm, possible oscillation or plateau
        end if
        $L_{prev} \leftarrow L_{current}$
        $\boldsymbol{\theta}_{prev} \leftarrow \boldsymbol{\theta}_{current}$
    end for
end procedure
```

In contrast to the conventional penalized regression solution, such as the coordinate descent, the GMIFS algorithm is missing a global optimum criterion and requires a convergence criterion. In [45], the authors of *countgmifs* library introduced a simple criterion in which the difference between two successive log-likelihoods is smaller than a pre-specified tolerance τ. However, during the computations, the oscillation behavior of the algorithm was observed near the optimum. Because of this, an additional criterion of oscillation avoidance is introduced. It is based on a strategy in which either the algorithm expands the space by introducing an additional non-zero coefficient variable, or it improves the maximum visited log-likelihood. If none of those are present in the sufficiently extensive number of iterations I_{max}, an oscillation or plateau is detected, and the algorithm is terminated. The alternative of adjusting the step size, which is common for a deep learning field [82], has been found to increase the time complexity severely, thus remaining for future work. Therefore, a cuttoff strategy has been found to be more suitable.

This strategy is summarized in Algorithm 2.

Appendix B proposed varying network statistic in GPACDA disease prediction

CircGPA/GPACDA [60] computes the score and p-value statistics for each circRNA-disease pair based on the permutation of miRNA/mRNA annotations, assuming a fixed network structure. This method was originally designed as an annotation algorithm rather than a tool for network comparison. In contrast, our study focuses on network modifications, with annotations assumed to be experimentally validated and fixed. To address this requirement, we propose a modification of the annotation statistic from GPACDA [66] to account for the changes in the network.

The notation is preserved as in [60, 66] with a minor modification of indices. The $\mathbf{a}^{m,c} \in \{0,1\}^{N_m \times N_c}$ is an adjacency matrix of the algorithm's predicted circRNA-miRNA parent set:

$$\mathbf{a}^{m,c}_{i,j} = \begin{cases} 1, \text{ if } m^i \in Pa(c^j) \\ 0, \text{ otherwise} \end{cases} \tag{B1}$$

Equivalently, $\mathbf{A}^{r,m} \in \{0,1\}^{N_r \times N_m}$ is the adjacency matrix of the miRNA-mRNA parent set. For the simplification only the $\mathbf{a}^{m,c}$ is predicted and $\mathbf{A}^{r,m}$ is assumed to be validated, e.g. TarBase v9.

The RNAs are compared against the set of N_d diseases $\mathbf{d} = \{d^1, d^2, ..., d^{N_d}\}$. In this study, a set of 60 diseases from [66] is used. Examples of these diseases are as follows: *hepatocellular carcinoma*, *acute myeloid leukemia*, *Alzheimer's disease*, etc.

The RNA-disease mapping vector for miRNAs $\mathbf{g}^m \in \{0,1\}^{N_m \times N_d}$, $\mathbf{g}^m = (\mathbf{g}^{m^1}, ..., \mathbf{g}^{m^{N_m}})$ is defined as:

$$\mathbf{g}^{m^k}(d^j) = \begin{cases} 1, \text{ if } m^k \text{ has an experimental evidence of association with disease } d^j \\ 0, \text{ otherwise} \end{cases} \tag{B2}$$

Equivalently, $\mathbf{g}^r \in \{0,1\}^{N_r \times N_d}$ and $\mathbf{g}^c \in \{0,1\}^{N_c \times N_d}$ are mapping vectors defined for mRNA-disease and circRNA-disease entries. For the remainder of the section the notation is simplified to $\mathbf{g}^m = \mathbf{g}^m(d^j)$, e.g., the formula below is to be repeated for each disease d^j and each circRNA c^i.

First – the statistic value (see Eq. B3) remains the same as in [60], as it represents the number of paths that connect a given circRNA c^i with RNAs with annotations of a particular disease term d^j. The network is given as $\mathbf{a}^{m,c}, \mathbf{A}^{r,m}$ and known miRNA/mRNA associations $\mathbf{g}^m, \mathbf{g}^r$.

$$s(c^i, \mathbf{a}^{m,c}_i) = (\mathbf{a}^{m,c}_i)^T \mathbf{g}^m + (\mathbf{a}^{m,c}_i)^T (\mathbf{A}^{r,m})^T \mathbf{g}^r. \tag{B3}$$

Next, we define a p-value of this statistic as the probability that a random circRNA-miRNA network part $\mathbf{a}^{m,c}_i$ of the same size $\|\mathbf{a}^{m,c}_i\|_1$ with the same miRNA/mRNA validated annotations $\mathbf{g}^m, \mathbf{g}^r$ and with the same fixed miRNA-mRNA network $\mathbf{A}^{r,m}$ reaches the same statistic value $s(c^i, \mathbf{a}^{\mu,c^i})$ or higher. Similar to [60], this means that the null hypothesis is that among all circRNA-miRNA networks of a given size, a circRNA c^i has no preference in interactions with miRNAs (or mediated interactions with mRNAs) annotated with disease d^j. The alternative hypothesis states that circRNA c^i should be annotated with disease d^j as it is overrepresented in the network.

This definition can be denoted as in Eq. B4.

$$pvalue_{\|\mathbf{a}_i^{m,c}\|_1}(s(c^i, \mathbf{a}_i^{m,c})) = P(s_{\|\mathbf{a}_i^{m,c}\|_1} \geq s(c^i, \mathbf{a}_i^{m,c})) \tag{B4}$$

where $s_{\|\mathbf{a}_i^{m,c}\|_1}$ is a statistic of a random c^i-miRNA network of size $\|\mathbf{a}_i^{m,c}\|_1$.

Consequently, the expected statistic value is now calculated differently as shown in Eq. B5.

$$\mathbb{E}\left(s(c^i, \mathbf{a}_i^{m,c})\right) = \frac{1}{\binom{N_m}{\|\mathbf{a}_i^{m,c}\|_1}} \sum_{\mathbf{a}_i^{m,c}} \left[(\mathbf{a}_i^{m,c})^T \mathbf{g}^m + (\mathbf{a}_i^{m,c})^T (\mathbf{A}^{r,m})^T \mathbf{g}^r\right] \tag{B5}$$

which can be solved analytically as in Eq. B6.

$$\mathbb{E}\left(s(c^i, \mathbf{a}_i^{m,c})\right) = \frac{\|\mathbf{a}_i^{m,c}\|_1}{N_m} \left[\mathbf{g}^m + (\mathbf{A}^{r,m})^T \mathbf{g}^r)\right] \tag{B6}$$

To compute a particular p-value for a given network $\mathbf{a}_i^{m,c}$, we can transform it into the same problem as in [60]. To show this, let us first rewrite the statistic from Eq. B3 to the form given in Eq. B7.

$$s(c^i, \mathbf{a}_i^{m,c}) = (\mathbf{a}_i^{m,c})^T \left[\mathbf{g}^m + (\mathbf{A}^{r,m})^T \mathbf{g}^r\right] \tag{B7}$$

It could be noted that the entire right term in square brackets is independent of $\mathbf{a}_i^{m,c}$ choice and could be calculated once. Let us denote it as $s(\mathbf{A}^{r,m}, \mathbf{g}^m, \mathbf{g}^r)$ in Eq. B8.

$$s(c^i, \mathbf{a}_i^{m,c}) = (\mathbf{a}_i^{m,c})^T s(\mathbf{A}^{r,m}, \mathbf{g}^m, \mathbf{g}^r) \tag{B8}$$

Now, the entire problem is reduced to a vector dot product and the term $\mathbf{a}_i^{m,c}$ is a binary vector. The key for a transformation here is the observation that swapping the $\mathbf{a}_i^{m,c}$ and $s(\mathbf{A}^{r,m}, \mathbf{g}^m, \mathbf{g}^r)$ terms in interpretation, this equation is actually the same as the mRNA part of the p-value computation in [60]. Consequently, to compute a p-value, we can use circGPA-implemented probability-generating polynomials as in Eq. B9.

$$pvalue_{\|\mathbf{a}_i^{m,c}\|_1}(s(c^i, \mathbf{a}_i^{m,c})) = \left(\text{genpoly}_{[\mathbf{g}^m + (\mathbf{A}^{r,m})^T \mathbf{g}^r]}(x, y) \mid y^{\|\mathbf{a}_i^{m,c}\|_1}\right) \tag{B9}$$

By repeating the p-value computation for every d^j and c^i, we can compute a p-value probability of every disease-circRNA pair given the predicted network $\mathbf{a}^{m,c}$. The remaining part is to compare the p-value to the validated ground-truth 0/1 value $\mathbf{g}_i^c(d_j)$.

Appendix C evaluation of eta hyperparameter impact on model performance

Understanding the role and behavior of the mixing parameter η is critical to interpreting the performance of the Pi-GMIFS framework. To investigate this, we performed a series of experiments assessing how η influences the model's structure inference capabilities across various prior types and sample sizes. The results are presented in the form of sensitivity (recall), precision and prior usage analysis plots, shown in Figs. 7, 8 and 9.

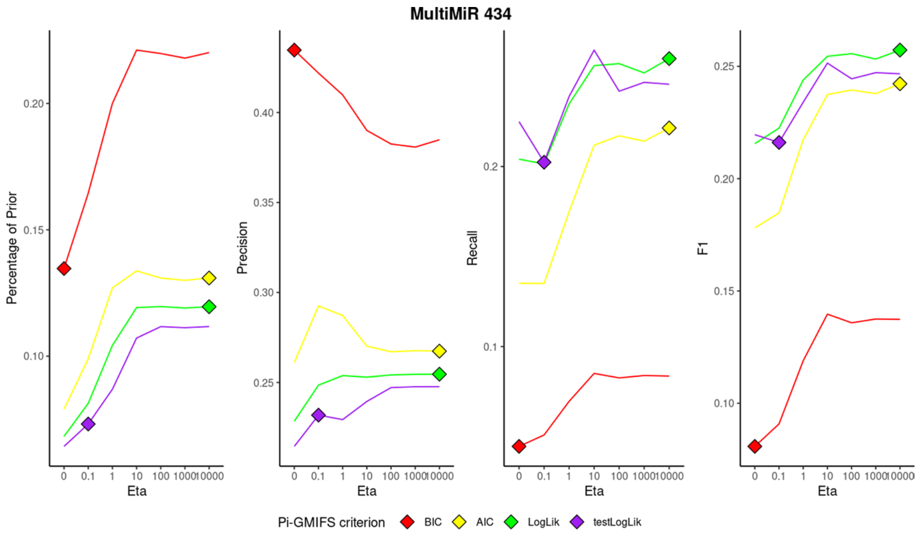


Fig. 7 Sensitivity (recall), precision and prior analysis of the mixing parameter η in the MultiMiR prior experiment with 434 RNA-seq samples. Each panel illustrates the effect of varying η (x-axis) on key performance metrics: percentage of prior interactions used in non-zero coefficients (left), precision (middle-left), recall (middle-right), and F1-score (right). Colored lines represent metric values computed for each globally fixed η value. Diamond markers indicate the η selected by global optimization under different Pi-GMIFS selection criteria: BIC (red), AIC (yellow), training log-likelihood (green), and test log-likelihood (purple). This analysis demonstrates the robustness and interpretability of η across performance and model selection objectives

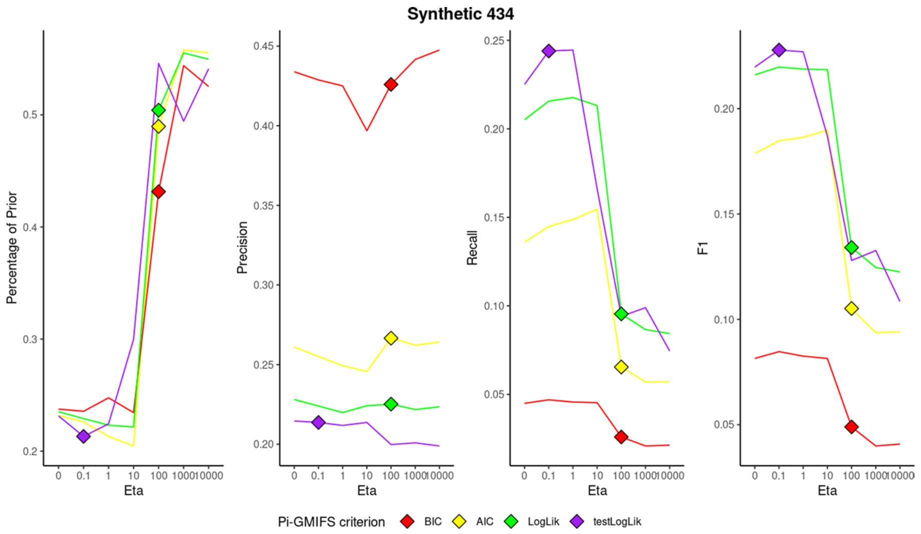


Fig. 8 Sensitivity (recall), precision and prior analysis of the mixing parameter η in the Synthetic-TargetScan prior experiment with 434 RNA-seq samples

To simplify interpretation, we applied a global η selection strategy, where a single constant η value was used uniformly across all mRNAs. This contrasts with the primary formulation of the algorithm in the preceding Sect. 4, which used a local selection approach, optimizing η individually for each mRNA. While local selection consistently achieves slightly higher F1-scores due to its flexibility, the globally fixed η yields performance that

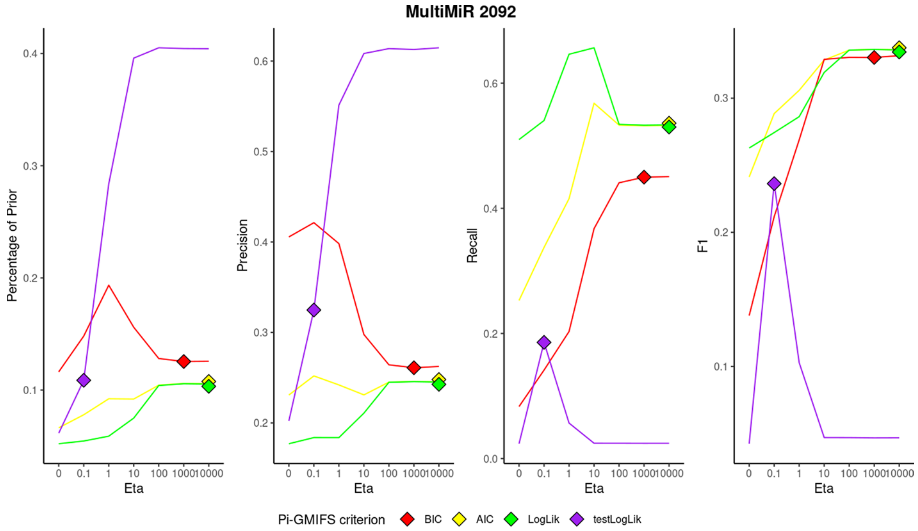


Fig. 9 Sensitivity (recall), precision and prior analysis of the mixing parameter η in the MultiMiR prior experiment with 2092 RNA-seq samples

is only marginally lower. Given the observed variability of optimal η values across different mRNAs, the strong performance of the global setting indicates a degree of robustness in η selection. That is, while per-mRNA optimization can enhance results, a single, globally chosen η is sufficient to deliver competitive performance at scale.

Moreover, even under the global η setting, it was possible to approximate the behavior of the algorithm's automatic η selection mechanism across individual mRNAs. Specifically, for each fixed global η value, we computed the aggregate values of the four selection criteria – AIC, BIC, training log-likelihood, and test log-likelihood-by summing the respective scores across all mRNAs. The η corresponding to the optimal aggregate score for each criterion was then identified and marked as a representative point on the sensitivity plots. This approach provides insight into how different selection metrics influence η preference on a global scale, complementing the localized per-mRNA optimization used previously.

A second key observation is that the mixing parameter behaves as theoretically expected: the proportion of prior interactions included in the final model generally increases with higher η values, as seen in left plot with prior edge percentage metric. Although the relationship between η and performance metrics – such as precision, recall, and F1-score – is not strictly monotonic, a consistent pattern emerges across experiments. Specifically, lower η values tend to be favored when the prior knowledge is less precise (e.g., the synthetic TargetScan-derived prior), whereas higher η values are selected when working with more accurate priors, such as MultiMiR. This supports the intended interpretation of η as a soft control over the influence of prior information, proving the stated effectiveness and robustness of the pLASSO's [46] robust prior incorporation technique.

Interestingly, increasing the RNA-seq sample size in the MultiMiR experiments did not lead to a reduction in the selected η values, contrary to initial expectations. One possible explanation is that the added samples contribute more to the identification of new interactions rather than diminishing the relative importance of existing prior knowledge, as seen in Fig. 9 in contrast to Fig. 7.

Finally, the analysis reveals that in some cases, the selected η may not be optimal, suggesting that more refined or adaptive tuning strategies could further improve the algorithm's performance. These findings open avenues for future exploration, including dynamic η estimation procedures and the integration of prior quality assessment directly into the selection mechanism.

Acknowledgements

The experimental work was partially supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/184/OHK3/3 T/13. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The results published here are in part based upon data generated by the TCGA Research Network:https://www.cancer.gov/tcga.

Author contributions

AA participated in drafting and designing the method, implementation, evaluation, and text of the paper. JK participated in drafting and designing the method, evaluation, and paper text.

Funding

The experimental work was partially supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/184/OHK3/3 T/13. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Data availability

The source code and data collection used are publicly available at: https://gitlab.fel.cvut.cz/anuarali/pikobnet. The RNA-Seq data for Breast Invasive Carcinoma (TCGA-BRCA) used in this study were obtained from the publicly available National Cancer Institute Genomic Data Commons (GDC) Data Portal https://portal.gdc.cancer.gov/projects/TCGA-BRCA, and data for TCGA Colon Adenocarcinoma (TCGA-COAD) were obtained from https://portal.gdc.cancer.gov/projects/TCGA-COAD. The other RNA-Seq datasets used in this study are publicly available and deposited in the NCBI Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo) under the accession numbers GSE236933, GSE236932 and GSE138734, and in the NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) under PRJNA896500 and PRJNA679200.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no Competing interests.

Received: 25 February 2025 Accepted: 7 May 2025

Published online: 27 May 2025

References

1. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586–97. https://doi.org/10.1016/j.molcel.2015.05.004.
2. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051.
3. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. Front Genet. 2019. https://doi.org/10.3389/fgene.2019.00535.
4. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Front Cell Dev Biol. 2014. https://doi.org/10.3389/fcell.2014.00038.
5. Kristensen LS, Andersen MS, Stagsted LVW, Ebbesen KK, Hansen TB, Kjems J. The biogenesis, biology and characterization of circular RNAs. Nat Rev Genet. 2019;20(11):675–91. https://doi.org/10.1038/s41576-019-0158-7.
6. Patop IL, Wüst S, Kadener S. Past, present, and future of circ RNA s. EMBO J. 2019. https://doi.org/10.15252/embj.2018100836.

7. Sakshi S, Jayasuriya R, Ganesan K, Xu B, Ramkumar KM. Role of circRNA-miRNA-mRNA interaction network in diabetes and its associated complications. Mol Ther- Nucl Acids. 2021;26:1291–302. https://doi.org/10.1016/j.omtn.2021.11.007.
8. Ala U. Competing endogenous RNAs, non-Coding RNAs and diseases: an intertwined story. Cells. 2020;9(7):1574. https://doi.org/10.3390/cells9071574.
9. Demirci YM, Sacar Demirci MD. Circular RNA-MicroRNA-MRNA interaction predictions in SARS-CoV-2 infection. J Integr Bioinform. 2021;18(1):45–50. https://doi.org/10.1515/jib-2020-0047.
10. Ayaz H, Aslam N, Awan FM, Basri R, Rauff B, Alzahrani B, Arif M, Ikram A, Obaid A, Naz A, Khan SN, Yang BB, Nazir A. Mapping CircRNA-miRNA-mRNA regulatory axis identifies hsa_circ_0080942 and hsa_circ_0080135 as a potential theranostic agents for SARS-CoV-2 infection. PLOS ONE. 2023;18(4):0283589. https://doi.org/10.1371/journal.pone.0283589.
11. Kariuki D, Asam K, Aouizerat BE, Lewis KA, Florez JC, Flowers E. Review of databases for experimentally validated human microRNA-mRNA interactions. Database. 2023. https://doi.org/10.1093/database/baad014.
12. Skoufos G, Kakoulidis P, Tastsoglou S, Zacharopoulou E, Kotsira V, Miliotis M, Mavromati G, Grigoriadis D, Zioga M, Velli A, Koutou I, Karagkouni D, Stavropoulos S, Kardaras FS, Lifousi A, Vavalou E, Ovsepian A, Skoulakis A, Tasoulis SK, Georgakopoulos SV, Plagianakos VP, Hatzigeorgiou AG. TarBase-v9.0 extends experimentally supported miRNA-gene interactions to cell-types and virally encoded miRNAs. Nucleic Acids Res. 2023;52:304–10. https://doi.org/10.1093/nar/gkad1071.
13. Riolo G, Cantara S, Marzocchi C, Ricci C. miRNA targets: from prediction tools to experimental validation. Methods Protoc. 2020;4(1):1. https://doi.org/10.3390/mps4010001.
14. Riolo G, Cantara S, Marzocchi C, Ricci C. miRNA targets: from prediction tools to experimental validation. Methods Protoc. 2020;4(1):1. https://doi.org/10.3390/mps4010001.
15. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015;4:05005.
16. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in drosophila. Gen Biol. 2003. https://doi.org/10.1186/gb-2003-5-1-r1.
17. Sun Z-Y, Yang C-L, Huang L-J, Mo Z-C, Zhang K-N, Fan W-H, Wang K-Y, Wu F, Wang J-G, Meng F-L, Zhao Z, Jiang T. circRNADisease v2.0 an updated resource for high-quality experimentally supported circRNA-disease associations. Nucl Acids Res. 2023;52:1193–200. https://doi.org/10.1093/nar/gkad949.
18. Borella M, Martello G, Risso D, Romualdi C. PsiNorm: a scalable normalization for single-cell RNA-seq data. Bioinformatics. 2021;38(1):164–72. https://doi.org/10.1093/bioinformatics/btab641.
19. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, Noble F, Rajewsky N. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495(7441):333–8. https://doi.org/10.1038/nature11928.
20. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics. 2009;25(23):3049–55. https://doi.org/10.1093/bioinformatics/btp565 (https://academic.oup.com/bioinformatics/article-pdf/25/23/3049/16891016/btp565.pdf).
21. Talukder A, Zhang W, Li X, Hu H. A deep learning method for miRNA/isomiR target detection. Sci Rep. 2022. https://doi.org/10.1038/s41598-022-14890-8.
22. Fridrich A, Hazan Y, Moran Y. Too many false targets for microRNAs: challenges and pitfalls in prediction of miRNA targets and their gene ontology in model and non-model organisms. BioEssays. 2019. https://doi.org/10.1002/bies.201800169.
23. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63. https://doi.org/10.1038/nrg2484.
24. Madhumita M, Paul S. A review on methods for predicting miRNA-mRNA regulatory modules. J Integr Bioinf. 2022. https://doi.org/10.1515/jib-2020-0048.
25. Liu B, Li J, Tsykin A, Liu L, Gaur AB, Goodall GJ. Exploring complex miRNA-mRNA interactions with bayesian networks by splitting-averaging strategy. BMC Bioinf. 2009. https://doi.org/10.1186/1471-2105-10-408.
26. Masegosa A, Moral S. New skeleton-based approaches for bayesian structure learning of bayesian networks. Appl Soft Comput. 2013;13:1110–20. https://doi.org/10.1016/j.asoc.2012.09.029.
27. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing bayesian network structure learning algorithm. Mach Learn. 2006;65(1):31–78. https://doi.org/10.1007/s10994-006-6889-7.
28. Li Y, Ziebart BD. Distributionally robust skeleton learning of discrete bayesian networks. arXiv (2023). https://doi.org/10.48550/ARXIV.2311.06117 .
29. Aragam B, Gu J, Zhou Q. Learning large-scale bayesian networks with the sparsebn package. J Stat Softw. 2019;91(11):1.
30. Chickering DM. In: Fisher, D., Lenz, H.-J. (eds.) Learning bayesian networks is NP-Complete, pp. 121– 130. Springer, New York, NY 1996. https://doi.org/10.1007/978-1-4612-2404-4_12 .
31. Lan W, Tang Z, Liu M, Chen Q, Peng W, Chen YP, Pan Y. The large language models on biomedical data analysis: a survey. IEEE J Biomed Health Inf. 2025;25:1–13. https://doi.org/10.1109/jbhi.2025.3530794.
32. Williams S, Huckle J. Easy Problems That LLMs Get Wrong. arXiv (2024). https://doi.org/10.48550/ARXIV.2405.19616 .
33. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinf. 2013. https://doi.org/10.1186/1471-2105-14-91.
34. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):29. https://doi.org/10.1186/gb-2014-15-2-r29.
35. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9(1):284. https://doi.org/10.1038/s41467-017-02554-5.
36. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):25. https://doi.org/10.1186/s13059-014-0550-8.
38. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research. 2016;5:1438.
39. Wang L, Audenaert P, Michoel T. High-dimensional bayesian network inference from systems genetics data using genetic node ordering. Front Genet. 2019;10:1196. https://doi.org/10.3389/fgene.2019.01196.
40. Lu Y, Zhou Y, Qu W, Deng M, Zhang C. A lasso regression model for the construction of microRNA-target regulatory networks. Bioinformatics. 2011;27(17):2406–13. https://doi.org/10.1093/bioinformatics/btr410.
41. Zou H, Hastie T. Regularization and variable selection via the Elastic net. J Royal Stat Soc Ser B: Stat Methodol. 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
42. Wang Z, Ma S, Wang C. Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. Biom J. 2015;57(5):867–84. https://doi.org/10.1002/bimj.201400143.
43. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
44. Lee KH, Pedroza C, Avritscher EBC, Mosquera RA, Tyson JE. Evaluation of negative binomial and zero-inflated negative binomial models for the analysis of zero-inflated count data: application to the telemedicine for children with medical complexity trial. Trials. 2023;24(1):25. https://doi.org/10.1186/s13063-023-07648-8.
45. Lehman RR, Archer KJ. Penalized negative binomial models for modeling an overdispersed count outcome with a high-dimensional predictor space: Application predicting micronuclei frequency. PLOS ONE. 2019;14(1):0209923. https://doi.org/10.1371/journal.pone.0209923.
46. Jiang Y, He Y, Zhang H. Variable selection with prior information for generalized linear models via the prior LASSO method. J Am Stat Assoc. 2016;111(513):355–76. https://doi.org/10.1080/01621459.2015.1008363.
47. Edgar R. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10. https://doi.org/10.1093/nar/30.1.207.
48. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20. https://doi.org/10.1038/ng.2764.
49. Griffiths-Jones, S.: miRBase: The microRNA Sequence Database, pp. 129–138. Humana Press. https://doi.org/10.1385/1-59745-123-1:129 .
50. Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, Bruford EA. Genenames org the hgnc resources in 2023. Nucleic Acids Res. 2022;51(D1):1003–9.
51. Trsova I, Hrustincova A, Krejcik Z, Kundrat D, Holoubek A, Staflova K, Janstova L, Vanikova S, Szikszai K, Klema J, Rysavy P, Belickova M, Kaisrlikova M, Vesela J, Cermak J, Jonasova A, Dostal J, Fric J, Musil J, Merkerova MD. Expression of circular RNAs in myelodysplastic neoplasms and their association with mutations in the splicing factor gene SF3B1. Molecular Oncology. 2023. https://doi.org/10.1002/1878-0261.13486.
52. Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014;20(11):1666–70. https://doi.org/10.1261/rna.043687.113.
53. Dudekula DB, Panda AC, Grammatikakis I, De S, Abdelmohsen K, Gorospe M. CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs. RNA Biol. 2016;13(1):34–42. https://doi.org/10.1080/15476286.2015.1128065. (**PMID: 26669964**).
54. Merkerova MD, Klema J, Kundrat D, Szikszai K, Krejcik Z, Hrustincova A, Trsova I, Le AV, Cermak J, Jonasova A, Belickova M. Noncoding RNAs and their response predictive value in azacitidine-treated patients with myelodysplastic syndrome and acute myeloid leukemia with myelodysplasia-related changes. Cancer Genom- Proteomics. 2022;19(2):205–28.
55. Lorenzi L, Chiu H-S, Avila Cobos F, Gross S, Volders P-J, Cannoodt R, Nuytens J, Vanderheyden K, Anckaert J, Lefever S, Tay AP, Bony EJ, Trypsteen W, Gysens F, Vromman M, Goovaerts T, Hansen TB, Kuersten S, Nijs N, Taghon T, Vermaelen K, Bracke KR, Saeys Y, De Meyer T, Deshpande NP, Anande G, Chen T-W, Wilkins MR, Unnikrishnan A, De Preter K, Kjems J, Koster J, Schroth GP, Vandesompele J, Sumazin P, Mestdagh P. The RNA Atlas expands the catalog of human non-coding RNAs. Nat Biotechnol. 2021;39(11):1453–65. https://doi.org/10.1038/s41587-021-00936-1.
56. Liu S, Wang Y, Duan L, Cui D, Deng K, Dong Z, Wei S. Whole transcriptome sequencing identifies a competitive endogenous RNA network that regulates the immunity of bladder cancer. Heliyon. 2024;10(8):29344. https://doi.org/10.1016/j.heliyon.2024.e29344.
57. Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, Mahaffey S, Rossi S, Calin GA, Bemis L, Theodorescu D. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. Nucleic Acids Res. 2014;42(17):133–133. https://doi.org/10.1093/nar/gku631.
58. Leclercq M, Diallo AB, Blanchette M. Prediction of human miRNA target genes using computationally reconstructed ancestral mammalian sequences. Nucleic Acids Res. 2016;45(2):556–66. https://doi.org/10.1093/nar/gkw1085.
59. Haecker I, Renne R. HITS-CLIP and PAR-CLIP advance viral miRNA targetome analysis. Critical Rev Eukaryot Gene Exp. 2014;24(2):101–16. https://doi.org/10.1615/critreveukaryotgeneexpr.2014006367.
60. Ryšavý P, Kléma J, Merkerová MD, circGPA,. circRNA functional annotation based on probability-generating functions. BMC Bioinf. 2022. https://doi.org/10.1186/s12859-022-04957-8.
61. Wu W, Ji P, Zhao F. CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome Biol. 2020;21(1):25. https://doi.org/10.1186/s13059-020-02018-y.
62. Kern F, Krammes L, Danz K, Diener C, Kehl T, Küchler O, Fehlmann T, Kahraman M, Rheinheimer S, Aparicio-Puerta E, Wagner S, Ludwig N, Backes C, Lenhof H-P, Briesen H, Hart M, Keller A, Meese E. Validation of human microRNA target pathways enables evaluation of target prediction tools. Nucleic Acids Res. 2020;49(1):127–44. https://doi.org/10.1093/nar/gkaa1161.
63. Fan C, Lei X, Tie J, Zhang Y, Wu F.-X, Pan Y. CircR2Disease v2.0: An updated web server for experimentally validated circRNA-disease associations and its application. Genom, Proteomics, Bioinf. 2021;20(3):435–45. https://doi.org/10.1016/j.gpb.2021.10.002.

64. Guo X-Y, He C-X, Wang Y-Q, Sun C, Li G-M, Su Q, Pan Q, Fan J-G. Circular RNA profiling and bioinformatic modeling identify its regulatory role in hepatic steatosis. BioMed Res Int. 2017;2017:1–13. https://doi.org/10.1155/2017/5936171.
65. Yao D, Zhang L, Zheng M, Sun X, Lu Y, Liu P. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Sci Rep. 2018;8(1):1108. https://doi.org/10.1038/s41598-018-29360-3.
66. Ryšavý P, Kléma J, Merkerová MD. 2024 GPACDA - circRNA-disease association prediction with generating polynomials . In: Bioinformatics and Biomedical Engineering. Lecture Notes in Computer Science, Springer, Cham;14848:33–48 . https://doi.org/10.1007/978-3-031-64629-4_3
67. Gill J, King G. What to do when your Hessian is not invertible: alternatives to model respecification in nonlinear estimation. Sociol Methods & Res. 2004;33(1):54–87. https://doi.org/10.1177/0049124103262681.
68. Lan W, Dong Y, Zhang H, Li C, Chen Q, Liu J, Wang J, Chen Y-PP. Benchmarking of computational methods for predicting circrna-disease associations. Brief Bioinf. 2023;241:25. https://doi.org/10.1093/bib/bbac613.
69. Lan W, Dong Y, Chen Q, Liu J, Wang J, Chen Y-PP, Pan S. Ignscda: Predicting circrna-disease associations based on improved graph convolutional network and negative sampling. IEEE/ACM Trans Comput Biol Bioinf. 2022;19(6):3530–8. https://doi.org/10.1109/tcbb.2021.3111607.
70. Lan W, Li C, Chen Q, Yu N, Pan Y, Zheng Y, Chen Y-PP. Lgcda: Predicting circrna-disease association based on fusion of local and global features. IEEE/ACM Trans Comput Biol Bioinf. 2024;21(5):1413–22. https://doi.org/10.1109/tcbb.2024.3387913.
71. Lan W, Wang J, Li M, Liu J, Wu F-X, Pan Y. Predicting microrna-disease associations based on improved microrna and disease similarities. IEEE/ACM Trans Comput Biol Bioinf. 2018;15(6):1774–82. https://doi.org/10.1109/tcbb.2016.2586190.
72. Lan W, Dong Y, Chen Q, Zheng R, Liu J, Pan Y, Chen Y-PP. Kgancda: predicting circrna-disease associations based on knowledge graph attention network. Brief Bioinf. 2021;23(1):25. https://doi.org/10.1093/bib/bbab494.
73. Gillis J, Pavlidis P. Guilt by Association is the exception rather than the rule in gene networks. PLoS Comput Biol. 2012;8(3):1002444. https://doi.org/10.1371/journal.pcbi.1002444.
74. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145–59. https://doi.org/10.1016/s0031-3203(96)00142-2.
75. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **2** 12(1) https://doi.org/10.1186/1471-2105-12-77
76. Dhanakshirur M, Laumann F, Park J, Barahona M. A continuous structural intervention distance to compare causal graphs. arXiv (2023). https://doi.org/10.48550/ARXIV.2307.16452 .
77. Tay JK, Narasimhan B, Hastie T. Elastic net regularization paths for all generalized linear models. J Stat Softw. 2023;106:1.
78. List M, Dehghani Amirabad A, Kostka D, Schulz MH. Large-scale inference of competing endogenous rna networks with sparse partial correlation. Bioinformatics. 2019;35(14):596–604. https://doi.org/10.1093/bioinformatics/btz314.
79. Le TD, Zhang J, Liu L, Liu H, Li J. mirlab: An r based dry lab for exploring mirna-mrna regulatory relationships. PLOS ONE. 2015;10(12):0145386. https://doi.org/10.1371/journal.pone.0145386.
80. Vo DHT, Thorne T. Shrinkage estimation of gene interaction networks in single-cell RNA sequencing data. BMC Bioinf. 2024;25(1):339. https://doi.org/10.1186/s12859-024-05946-9.
81. Thorne T. Approximate inference of gene regulatory network models from RNA-Seq time series data. BMC Bioinf. 2018;19(1):1. https://doi.org/10.1186/s12859-018-2125-2.
82. Degris, T., Javed, K., Sharifnassab, A., Liu, Y., Sutton, R.: Step-size optimization for continual learning. arXiv (2024). https://doi.org/10.48550/ARXIV.2401.17401 .

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.