Computational and Structural Biotechnology Journal 18 (2020) 548–557

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Review

Prediction of the miRNA interactome – Established methods and upcoming perspectives

Moritz Schäfer [a,b], Constance Ciaudo [a,*]

[a] Swiss Federal Institute of Technology Zurich, Department of Biology, Institute of Molecular Health Sciences, CH-8093 Zurich, Switzerland
[b] Life Science Zurich Graduate School, Systems Biology Program, University of Zurich, CH-8047 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 16 October 2019
Received in revised form 21 February 2020
Accepted 27 February 2020
Available online 5 March 2020

Keywords:
Machine Learning
Deep Learning
microRNA target prediction

ABSTRACT

MicroRNAs (miRNAs) are well-studied small noncoding RNAs involved in post-transcriptional gene regulation in a wide range of organisms, including mammals. Their function is mediated by base pairing with their target RNAs. Although many features required for miRNA-mediated repression have been described, the identification of functional interactions is still challenging. In the last two decades, numerous Machine Learning (ML) models have been developed to predict their putative targets. In this review, we summarize the biological knowledge and the experimental data used to develop these ML models. Recently, Deep Neural Network-based models have also emerged in miRNA interaction modeling. We thus outline established and emerging models to give a perspective on the future developments needed to improve the identification of genes directly regulated by miRNAs.

Contents

* Corresponding author.
E-mail address: cciaudo@ethz.ch (C. Ciaudo).

https://doi.org/10.1016/j.csbj.2020.02.019
2001-0370/

1. Introduction

1.1. Regulation of gene expression by microRNAs

Mammalian canonical microRNAs (miRNAs) are key components of cellular networks, regulating gene expression in many cell types (for review [1]). Their biogenesis starts with the transcription of primary miRNAs (pri-miRNAs) by RNA polymerase II and is followed by their processing in the nucleus by the microprocessor complex, composed of two DGCR8 and one DROSHA protein, into hairpin precursor miRNAs (pre-miRNAs) of ~70 nucleotides (nts) length. After export into the cytoplasm, pre-miRNAs are cleaved by the DICER endonuclease III enzyme into mature miRNA duplexes of ~22 nts length (for review [2]). Mature miRNAs are subsequently loaded in one of the Argonaute proteins forming the RISC (RNA-induced silencing complex) to target 3′ Untranslated Region (3′UTR) of mRNAs, mostly via Watson-Crick base pairing of their seed region (comprising nucleotides 2–8 (Fig. 1a)) [3]. This interaction leads to post-transcriptional repression of mRNAs via at least one of two modes: *Translational inhibition* and *mRNA decay* [4]. Interestingly, it was previously shown, using human cells as a model, that 84% of translational inhibition by miRNAs might be attributed to mRNA degradation [5]. Nevertheless, the potential to functionally repress mRNAs varies across individual miRNA-target pairs and a number of specific factors have been associated with regulatory efficiency (see Section 2.2.1). Interestingly, a given 3′UTR can be targeted by several miRNAs and an individual miRNA might also have many targets [3], implying millions of possible interactions, which may or may not be functional. As a consequence, there is a need for computational methods to identify functional miRNA-mRNA interactions and the key features leading to post-transcriptional repression.

To determine such potential interactions, the first computational miRNA interaction prediction (MIP) methods analyzed the 3′UTRs of genes for conserved seed matches [6]. While this approach was effective to identify many interactions, it became apparent that these predicted matches were not always functional and that several other features contribute to the repression mediated by a given miRNA-mRNA interaction pair [7,8].

Experimentally, over-expression of individual miRNAs followed by the analysis of targeted mRNA expression using microarrays, produced data sets with high numbers of potential direct and indirect interactions [9]. These data sets were used as the first basis for the prediction of functional interactions using Machine Learning techniques in mammals [7].

1.2. Machine Learning

The term *Machine Learning* (ML) dates back to 1959 [10] and has gained tremendous popularity in the last decades [11]. ML methods commonly analyze collected data in order to make predictions from novel observations [12]. In this review, we focus on Machine Learning techniques that are applied in the context of MIP.

Most of these models rely on *Supervised Learning*. Here, ML algorithms are provided with (*input, output*) pairs for training of a function:

$$f(\mathit{input}) \approx \mathrm{output}$$

During training, the function, or model, f is optimized to estimate the *output*, or *label*, based on the provided *input*, or *observation*, preferably using a large number of (*input, output*) pairs. After training, the model can be used to predict the unknown output, for a provided input [11]. In MIP, the input consists of features, describing a given potential miRNA-mRNA interaction (e.g. the degree of conservation at the target site, see Section 2.2.1), while the output indicates the repressive potential of the interaction.

Unsupervised Learning is a form of learning that allows the identification of patterns in data sets without the use of labels. It is especially useful for working with unlabeled data sets and for learning more descriptive feature representations in cases where supervised learning methods struggle to capture relevant information from the *raw features* of an observation (i.e. unprocessed features from observations). Learned feature representations are computable from raw features and can be used in subsequent supervised learning models [13].

The repressive potential of a miRNA-mRNA interaction can be described either categorically or continuously. In the former case, interactions are typically classified in two classes *functional* and *nonfunctional*, reflecting whether a considered interaction mediates gene repression or not. This kind of prediction is referred to as *classification*. In the latter case, the repressive potential of an interaction is estimated as a continuous variable, which is referred to as *regression*.

Another notable property of ML methods is whether they are capable of modeling nonlinear relationships. ML models can be separated into *linear* and *nonlinear* models. Nonlinearities in the training data cannot be modeled accurately by a linear model, while nonlinear models may perform well on linear and nonlinear data [12]. To provide an example, the repression of the target mRNA is not proportional to the amount of miRNAs present in the system. It has been observed that repression usually plateaus, or changes, only after the miRNA expression reached a specific threshold. This implies a nonlinear relationship between miRNA expression and target repression [3] and a nonlinear model might thus lead to more accurate predictions than a linear one.

1.2.1. ML methods

A set of Machine Learning methods have been widely employed in biology and previously reviewed [14]. Here, we focus our introduction on methods being broadly applied in MIP.

- Linear methods
  The **Linear Regression** is the simplest imaginable regression model. It fits a line of the form $y = ax + b$ (in the one-dimensional case, with x being the observation, y the prediction and a and b the fitted model parameters) to the observed data such that deviations of the observations from that line are minimized. The possibility to directly interpret the model parameters as well as to fit the model to very small data sets are the strengths of this model. **Logistic Regression** is a method for the parameter estimation of a logistic function. The function models the probability of an observation to belong to one of two *classes* and can be used to classify an observation using a threshold. Linear methods can be based on more complex formulas with larger numbers of trainable parameters, which for example enables them to model data using polynomials. These models usually have the disadvantage of *overfitting*, i.e. they mimic the training data extremely well, but fail to predict new observations accurately. In order to keep complex models simple and to therefore avoid overfitting, model parameters are often incentivized to stay small, a process called *regularization*. These concepts and linear methods are comprehensively described in [12].
- Kernel methods
  Kernels are functions that compute similarity measures between two observations. Kernel methods make use of these measures to train models based on pairwise similarities. This enables them to model nonlinear relationships between features without the need to explicitly convert observations to a high-dimensional feature space as described in [15].

The **Support Vector Machine** (SVM) is a classification method that commonly employs kernel functions. It works by spanning a hyperplane in the feature space to linearly separate training observations of different classes. Using kernel functions, this feature space can be implicitly transformed, enabling the nonlinear separation of observations in the original feature space. Notably, in order to improve classification of *unseen* samples (i.e. samples that have not been observed during training), the hyperplane is optimized to maximize its margin to the training samples of two classes. SVMs are able to efficiently process high-dimensional and large data sets and reach high prediction accuracies [16]. They have been previously reviewed in the context of biology in [15].

- Tree-based methods

  Decision Tree models are able to classify observations by traversing through the nodes of a *tree-like* structure. Each node represents a test of one or more features of the observation and determines, which of the subsequent nodes to traverse next, ultimately leading to a predicted class when reaching a leaf node (i.e. a node without subsequent child nodes) [17]. Using *ensemble learning*, multiple decision trees can be trained and combined into a *Random Forest* model, which usually leads to improved predictive performance. Random Forests are able to efficiently model complex and nonlinear data types, while still retaining the ability to interpret the relevance of individual features [18,19]. They have been described in the context of biology in [17,18].
- **Neural networks** (NNs)

  NNs, in their ordinary form, work by transforming input features towards more abstract *representations*. These networks are built up of *layers*, each responsible for a linear feature transformation followed by a nonlinear activation function. The output of the NN is usually retrieved from the final layer. NNs having a larger number of layers, are also referred to as Deep Neural Networks (DNNs) [20]. As an example, an image-based object recognition NN model, receiving raw pixel values as input, usually learns common concepts comparable to lines and curves in its first layers and combines these features to shapes, like circles, corners and more complex patterns in subsequent layers, to finally predict object classes such as handwritten characters or cell types in the last layer. Given a sufficient amount of training data, the back-propagation algorithm enables the training of the transformations directly from raw features [21]. This enables NNs to automatically extract complex and useful features from the provided input. NN methods have been extended to effectively model different kinds of data. High-dimensional data with homogenous features, as in images or nucleotide sequences, can be modeled efficiently using **Convolutional Neural Networks** (CNNs). In CNNs, trainable filters are applied with convolutional operations onto input data, to compute the output of a layer in the network. This leads to fewer model parameters needed to be trained, as compared to NNs with fully connected layers [20].

  In many cases, input representations used to train ML models are substantially larger than necessary for carrying the contained information. This hampers training due to the increased number of model parameters needed to be trained. The **Autoencoder** architecture, a specific type of NNs, uses unsupervised learning to encode the input data into a form of fewer dimensions, which can be used for further training. A NN Autoencoder is designed and trained such that it reproduces the input data in its output layer, while its inner *encoded* representations have fewer dimensions than the input data [22].

  Another type of NNs are **Recurrent Neural Networks** (RNNs). They are capable of *storing* previous events for the adapted processing of new observations, which is suitable to model sequential data including nucleotide sequences [23–25].

As of yet, DNN models have paved the way for groundbreaking performance in many fields including speech recognition, natural language processing and visual object recognition (for review [13,26]). In biology, DNNs have outperformed other ML methods in a wide range of applications and have been comprehensively reviewed in [20].

1.2.2. Model evaluation

A major challenge in ML is to design models that gain the ability to *generalize* their predictive power onto unseen data. Observations, used for prediction of a trained model, usually deviate from the training observations. Nevertheless, it is desirable that a model predicts the label of new observations as accurate as for observations it has already observed during training. In order to verify this property of a model, the model is evaluated on a *test data set*, which must be different from the *training data set*, previously used for the training. Especially in computational biology, researchers often use and recommend labeled test sets that are completely independent of the training data (e.g. from a different organism or experimental method) [27]. Furthermore, during training and model optimization, instead of using the test set, a part of the training set is excluded from training and used as separate *validation set*. This step is necessary to avoid systematic optimization of the model towards the test set. As observations in a given data set may share common properties, it is important to carefully split data sets in a way that observations with the same property never appear at the same time in the training set and the validation set [28]. As an example, this can be an issue for miRNA interaction data sets, which mostly contain data on a limited number of miRNAs (see Section 2.1). Some miRNAs may be generally more potent at repressing their targets than others. This notion could be learned by a model and, in case of an improperly split validation set, be exploited to report better model performances than reachable in a realistic scenario, where mostly interactions with previously unobserved miRNAs are being predicted.

In order to assess, or score, the performance of a model for comparison and optimization, a number of metrics exist. For regression models, the *root-mean-square error* (RMSE) and the *coefficient of determination* denoted as R^2 are common scoring metrics.

Given a test set of length n with known *outputs* y and predicted *outputs* $\hat{y}$ it is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i^n (y_i - \hat{y}_i)^2}{n}}$$

Due to low reliability of the RMSE [29], the R^2 is used more often for model scoring. It is a measure for the proportion of variance in a data set's *output*, explained by a given regression model and defined as

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}$$

with $\bar{y} = \frac{\sum_i^n y_i}{n}$ being the mean of the *outputs* of the data set [30].

Binary classification is the most common form of classification in MIP as interactions are usually separated in one of two classes positive/functional or negative/nonfunctional. Metrics for classification models have been discussed in [31,32] and are summarized below.

When evaluating a binary classification model on a test set, the number of correctly predicted positives (true positives (TP)) and incorrectly predicted positives (false positives (FP)), as well as true negatives (TN) and false negatives (FN) can be counted and used to derive a number of metrics, including

- precision: $\frac{TP}{TP+FP}$, i.e. the fraction of predicted positives that are actually true positives.
- true-positive rate or recall: $\frac{TP}{TP+FN}$, i.e. the fraction of positives that have been identified as such.
- false-positive rate or (1-specificity): $\frac{FP}{FP+TN}$, i.e. the fraction of negatives that have been wrongly identified as positives.
- accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$, i.e. the fraction of predictions that have been predicted correctly.

It might seem intuitive at first to use the accuracy for evaluating a model's performance. However, none of these metrics are robust against an *imbalanced data set*, where the number of positive and negative observations is largely different. For example, a (hypothetical) prediction model that predicts every observation as negative, could yield high accuracy scores, given a data set with a large number of negatives and few positives. To counteract such issues, it is advised to use these metrics in combination as does for example the F_1 score, which is more robust against imbalanced data sets and defined as:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Furthermore, as binary classifiers mostly use thresholds to determine the output class based on a predicted score, it is more informative to evaluate the performance of a model for a range of thresholds. Here, receiver operating characteristic (ROC) graphs are often used to plot the false-positive rate (i.e. 1 – specificity) on the x axis against the true-positive rate on the y axis, covering the full range of possible threshold values. This graph can be reduced to a single value metric by calculating the *area under the ROC curve* (AUROC or AUC) with 1 being the best and 0 being the worst value [33].

For a concise overview of recommended ML practices beyond model evaluation, the reader is referred to [34].

2. Methods

Despite several years of effort to better understand the underlying factors and contexts involved in efficient translational inhibition of mammalian mRNAs by miRNAs, it is still challenging to identify functionally relevant interactions [35]. To tackle this issue, researchers have employed Machine Learning techniques to build models that predict functional interactions based on experimental observations.

2.1. Experimental methods and data used for ML-based miRNA target prediction

As highlighted previously, supervised ML relies on labeled data sets, both for training models and for subsequently testing them. For the generation of ML data sets, most MIP methods rely on finding seed matches in the 3′UTR to extract potential interactions. The regulatory efficiency of the extracted interactions is assigned based on data sets from experimental methods described in this section. For regression models, the regulatory effect observed for a given interaction is used as continuous variable [36]. For classification models, a threshold usually decides whether an interaction is *positive* or *negative* [37]. Although data sets generated in this way do contain errors (e.g. experimental false positives), they are the basis for ML training and hence are seen as *ground truth* for the ML model.

A direct functional miRNA-mRNA interaction involves the binding of an AGO2-bound miRNA to an mRNA, leading to its translational inhibition and generally the subsequent downregulation of that mRNA [5]. In practice, RNA levels are easier to measure with high-throughput methods than total protein levels. Consequently, mRNA expression is widely used as a proxy for miRNA-mediated gene regulation and used as ground truth.

Several groups have generated data sets based on miRNA perturbation and subsequent mRNA expression measurements, in order to build ML models. In 2005, Lee et al. were the first to perform a miRNA transfection in HeLa cells with high-throughput measurement of induced mRNA repression 12–24 h post transfection using microarrays [9]. Similarly, knockdown of miRNAs, followed by transcript measurement has been performed [38]. In 2015, Agarwal et al. have collected and compiled a data set containing 74 microarray data sets of HeLa cells transfected with individual miRNAs [36]. More recently, Liu et al. also overexpressed 25 miRNAs individually in HeLa cells and performed RNA-sequencing (RNA-seq) to identify affected transcripts [39]. Such approaches have some obvious issues: 1. The overexpression of miRNAs might lead to interactions that cannot be observed in physiological conditions where miRNA levels are lower. 2. It cannot be guaranteed that an observed repression was provoked by a transfected miRNA or by some secondary effect, for example through a repressed transcription factor.

In order to observe direct functional mRNA-miRNA interactions, researchers commonly use Luciferase reporter assays [40]. Van Peer et al. performed luciferase assays for 470 miRNAs and 3′UTRs of 17 human genes in parallel, thus observing 7990 potential interactions [37]. Since a protein-based readout is taken here, this kind of measurement is very direct but experimentally time consuming due to the necessary cloning and transfection steps. Mutation of putative miRNA binding sites in the 3′UTR of interest can furthermore confirm a direct interaction of the studied miRNA. This data set contains a significant number of tested miRNAs (but rather a small number of targets), which might be favorable for modeling functional interactions, since binding affinities can deviate greatly between distinct miRNAs [41]. However, the number of observed genes and potential interactions is low, compared to RNA expression-based approaches described previously.

Other data sets emerged with the development of Cross-Linking Immunoprecipitation with High-Throughput Sequencing methods (CLIP-seq, for review [42]), which allow the identification of RNA-binding protein target sites up to single nucleotide level precision. This approach has been used to reveal binding sites of miRNAs by immunoprecipitation of Argonaute protein *in vitro* [43] and *in vivo* [44]. Nevertheless, the identification of miRNA target sites did, in many cases, not reveal the associated miRNA. MiR-CLIP [45] and C*ross-linking Ligation and Sequencing of Hybrids* (CLASH) [46] methods were then developed to reveal this additional information. However, miR-CLIP is only capable of revealing interactions of a single miRNA and CLASH only identifies miRNA-mRNA interactions at a very low efficacy [46,47]. Finally, despite their high sensitivity, it has been shown that a substantial number of CLIP-seq- and CLASH-seq-identified sites, especially in the coding region of a gene (CDS), do not mediate target repression [36].

In vitro approaches have also been undertaken recently in order to identify functional miRNA-mRNA interactions. Two studies carried out high-throughput binding affinity experiments of the RISC for several individual miRNAs [41,48]. They generated libraries of potential miRNA binding sites, containing mismatches and non-canonical seed sites, and measured the amount of RISC bound to their library in cell-free extracts by high-throughput sequencing. Both groups identified large differences in the binding patterns of individual miRNAs but also confirmed previous observations made *in vivo*, e.g. regarding seed binding patterns.

Finally, curated databases of published experimental miRNA regulation experiments are also available: MiRTarBase [49], DIANA-TarBase [50] and miRecords [51]. Experimentally tested miRNA-target pairs are regrouped to reflect the employed experimental method and whether a repression has been observed or not. Using these validated miRNA-mRNA interactions for modeling brings the advantage of introducing little bias as these data were generated from different laboratories, cell types and contexts. This however comes at the cost of introducing a high variance, since it is challenging to integrate data obtained from wide-ranging experimental conditions. As an example, depending on the laboratory, different stringencies may have been utilized in the cut-offs used to report functional interactions, and this information is not transparent to the ML method. Furthermore, experimental data stored in these databases often stem from indirect readout methods described above (e.g. RNA-seq) and are indeed correlative and not causative.

To conclude, experimental data sets for miRNA-mRNA interactions come from a variety of experiments, each having their unique *pros and cons*. If possible, it is recommended to use a combination of data sources for MIP such that shortcomings of different types of data sets can be factored out. The variation of miRNAs, mRNAs and cell types used in different experiments however, makes it a challenging task to combine them.

2.2. Computational miRNA target prediction methods

A large number of prediction models have been published, based on specific features of miRNA-mRNA interactions, extracted from the previously described data sets (for review [52–57]). Here, we outline the main features used by current MIP ML methods and review the potential of modern neural network-based methods to improve prediction accuracy. An overview of these methods can be found in Table 1.

2.2.1. Development of biological knowledge shapes ML features

Most ML methods used in MIP rely on manually engineered features for effective prediction. Instead of providing *raw* descriptors of miRNA-mRNA interactions (e.g. the nucleotide sequences of the miRNA and the mRNA target site), features that have been shown to correlate with regulation efficiency, are being computed. These are typically represented either by a binary value (e.g. 0 or 1) or by a continuous value (e.g. any value between 0 and 1). The former case is usually used to describe categories like the seed type or a nucleotide identity (e.g. miRNA nucleotide at position 1 is an Adenine). Continuous features are in many cases represented by existing metrics (e.g. binding free energy, conservation). In other cases, simple computations can be used (e.g. the ratio of Adenine and Uracil vs Cytosine and Guanine in a given sequence). The most relevant biological findings and their integration into prediction models via feature engineering are visualized in Fig. 1a and

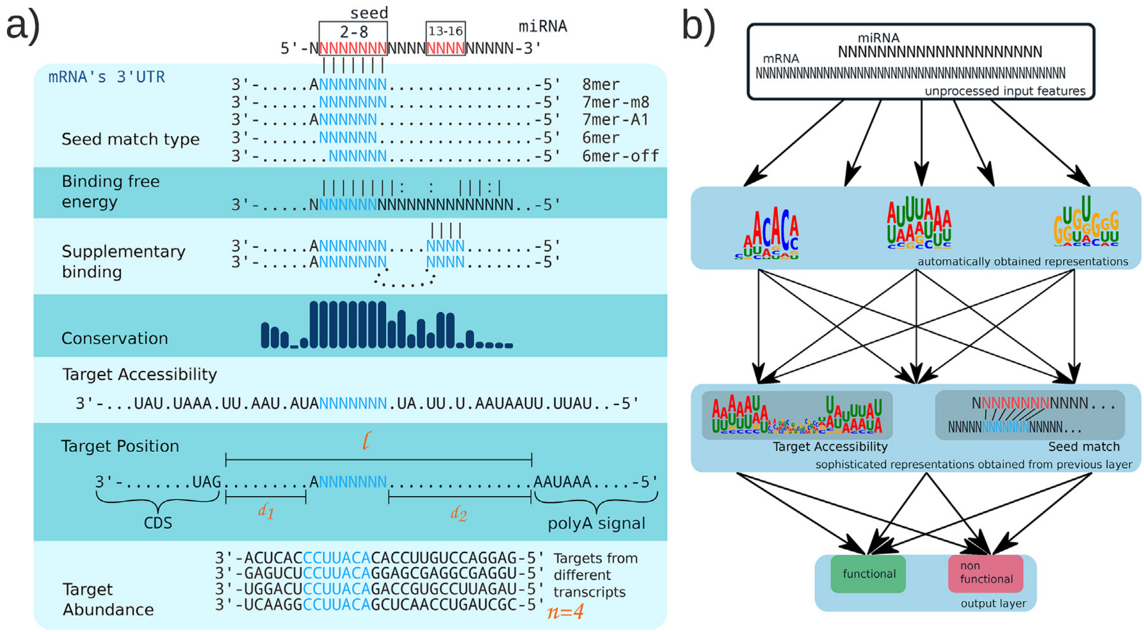
a)
seed
2-8
13-16
miRNA
5'-NNNNNNNNNNNNNNNNNNNNN-3'
mRNA's 3'UTR
Seed match type
3'-.....ANNNNNNN...............-5' 8mer
3'-......NNNNNNN...............-5' 7mer-m8
3'-.....ANNNNNN...............-5' 7mer-A1
3'-......NNNNNN...............-5' 6mer
3'-.......NNNNNN...............-5' 6mer-off
Binding free energy
3'-.....NNNNNNNNNNNNNNNNNNNN..-5'
Supplementary binding
3'-.....ANNNNNNN....NNNN.......-5'
3'-.....ANNNNNNN NNNN.......-5'
Conservation
Target Accessibility
3'-...UAU.UAAA.UU.AAU.AUANNNNNNN.UA.UU.U.AAUAAUU.UUAU..-5'
Target Position
l
3'-.......UAG.........ANNNNNNN...............AAUAAA....-5'
CDS
d1
d2
polyA signal
Target Abundance
3'-ACUCACCCUUACACACCUUGUCCAGGAG-5'
3'-GAGUCUCCUUACAGGAGCGAGGCGAGGU-5'
3'-UGGACUCCUUACAGACCGUGCCUUAGAU-5'
3'-UCAAGGCCUUACAGCUCAACCUGAUCGC-5'
Targets from different transcripts
n=4
b)
miRNA
mRNA
unprocessed input features
automatically obtained representations
Target Accessibility
Seed match
sophisticated representations obtained from previous layer
functional
non functional
output layer

Fig. 1. Predictive features for functional miRNA-target interactions. a) Manually engineered features based on biological assumptions. Seed and supplementary region of the miRNA are marked in red, corresponding complementary regions on the mRNA are marked in blue. Dots and Ns denote arbitrary nucleotides. Blue Ns and vertical bars denote Watson-Crick base pairing. From top to bottom: Nucleotides 2–8 define the **seed** of a miRNA. Extensive base matching as well as an Adenine opposite the first miRNA nucleotide generally lead to stronger repression. By simulating a heteroduplex between the miRNA and its putative binding site, the **binding free energy** of the interaction can be determined. This feature is also used for shorter interaction parts like the seed region. Nucleotides 13–16 of a miRNA are denoted as its **supplementary region** and the extent of binding to this region is extracted into a feature. Of note, the mRNA can form a bulge opposite the miRNA's central region. Functional miRNA targets are often **conserved** and features have been developed to convert conservation into a usable metric. Since mRNAs can fold and form secondary structures, some target sites are more **accessible** for RISC mediated repression than others. High AU content either near the putative site or in the whole 3′UTR has been shown to increase site accessibility and is therefore commonly used as feature. Not only the folding of mRNA can hinder efficient repression. It has been suggested that the ribosome complex can compete with the RISC for binding, which might explain why coding sequence (CDS) regions are not targeted to the same extent as 3′UTRs. The **position** (d_1, d_2) within the 3′UTR as well as the **length** (l) of it are therefore important binding site features. Individual miRNAs that target large sets of mRNAs, might distribute their repression potential, leading to a decreased repression level for individual mRNAs. Here, the **target site abundance** (n) is counted and used as features. Target nucleotides bound to the seed or the supplementary region are colored in blue and vertical dashes denote Watson-Crick base pairing. b) Implicit feature extraction by neural networks, based on the provided training data, here exemplified for a fictive MIP NN model. Nucleotide identities for the relevant input sequences are the only input data provided. In the first layer(s), simple features are extracted from the raw sequence data, which are then combined to more complex features in the later layers. Such features may resemble engineered features as described in a), but may also include unexpected, yet predictive, representations. This hierarchical structure leads to the autonomous extraction of high-level features, ultimately enabling the assessment of the repression potential of input interactions. b) inspired by [20]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

described in this section. For a dedicated and more comprehensive review on miRNA interaction prediction features, the reader is referred to [53].

- **Seed Type:** Nucleotides 2–8 define the seed of a miRNA. Extensive base matching as well as an adenine opposite to the first miRNA nucleotide generally lead to stronger repression.
- **Binding Free Energy:** By simulating a heteroduplex between the miRNA and its putative binding site, it is possible to calculate the binding free energy of an interaction [58]. ML models have made use of this feature, not only by computing the binding free energy for the complete miRNA-target duplex but also for the seed only (a feature denoted as "seed-pairing stability"). The RNA22 MIP model uses this feature and does not employ ML, giving it the opportunity to predict interactions without introducing unwanted artifacts from data sets [59].
- **Supplementary Binding:** Nucleotides 13–16 of a miRNA are denoted as its supplementary region. Having been shown to affect interaction efficacy for a number of miRNAs, the extent of pairing between the supplementary region and the target site is used as feature [7].
- **Target Site Conservation:** Functional miRNA targets are often conserved and conservation scoring metrics have been developed for use as features [6].
- **Target Site Accessibility:** Since mRNAs can fold and form secondary structures, some target sites might be more accessible for RISC mediated repression than others [60–62]. High AU content, both near the putative site and also in the whole 3′UTR, has been shown to increase site accessibility and is therefore also used as feature [8].
- **Target Site Position:** Not only can the folding of mRNA hinder efficient repression, it has furthermore been suggested that the ribosome complex competes with the RISC for binding, which might explain why CDS regions are not targeted to the same extent as 3′UTRs [63]. This competition might also explain the observation that target sites in the first and last nucleotides of a 3′UTR lead to significantly lower translational repression [64]. As such, additional features include the distances of a target site to the end of the CDS and to the polyA signal but also the length of the 3′UTR. All of these highly correlate with interaction efficiency [65,66].
- **Target Site Abundance:** A miRNA that is targeting large sets of mRNAs might distribute its repression potential, leading to a decreased repression for individual mRNAs [67]. The number of seed-complementary target sites in the 3′UTRs of the studied organism for a given miRNA, reflects this mechanistic as a simple feature [68].

Although the previously described features have been shown to correlate with target repression, the contribution of single features to the degree of repression as well as the interplay between them is hard to determine. Here, ML methods can use miRNA mediated repression observations from biological experiments, associate them with the described features and finally determine the relationships between features and observed repression.

2.2.2. MIP methods

Since the first use of ML in MIP, models have improved significantly, predominantly driven by the discovery of additional relevant biological features as well as by new methods to produce experimental data. State-of-the-art models mostly rely on a similar set of engineered features and deviate mostly in the experimental data sets and in the ML architecture they use.

One of the oldest MIP models is **TargetScan**. It received ongoing updates, with the latest version (v7) from 2015 using a simple linear regression model. This enabled the efficient selection of the 14 (out of 26) most relevant features using *step-wise regression* to be used for training, leading to a highly interpretable model, which significantly outperformed previously published models including DIANA-microT-CDS (discussed below) on several independent test data sets [36]. Training was performed using a compiled data set of 74 microarray data sets of HeLa cells transfected with individual miRNAs. The use of such a non-complex model comes at the cost of only catching linear relationships in the data, and the *step-wise regression* technique has been suggested to oversimplify models and to bias them towards the training data [69,70]. TargetScan predicts the repressive potential for single target sites and subsequently combines these predictions to scores for miRNA-mRNA pairs using a mathematical model. Consequently, potential combinatorial effects of binding sites on the same mRNA are not modeled based on experimentally obtained training data, which might introduce inaccuracies in predictions.

The **miSTAR** method from 2016 uses a stacked model to address this major challenge [37]. The first layer uses a Random Forest classification to estimate whether a repression is provoked by a single miRNA binding site. The second layer uses these estimates on a per-transcript basis and predicts a repression for the miRNA-transcript pair using a logistic regression. miSTAR is based on 3′UTR reporter assays performed for 470 miRNAs and 3′ UTRs from 17 genes (see Section 2.1). With recent reports showing strongly deviating binding patterns across different miRNAs [41,48], this comes as an advantage to generalize for unseen miRNAs. The authors demonstrated the improved generalization capability for unseen miRNAs in comparison to unseen mRNAs by cross-validation, leaving out individual miRNAs or mRNAs as tests during training. Using their model, they outperformed popular models, including TargetScan (version 6.2), using the AUC score as evaluation metric. Although the authors used a stringent cross-validation approach to estimate the performance of their model, they only tested against the data set they had used to optimize the model. Furthermore, only a small number of interactions, from which most were negative, were used for training.

In contrast, CLIP-seq identifies large numbers of interactions covering a high number of distinct miRNAs as well as mRNAs. The **DIANA-microT-CDS** (2012) model is a popular example for the use of this type of data for model training [71]. For the generation of training samples, positive and negative interactions were extracted by identifying overlaps between CLIP-seq data and miRNA complementary regions. Two models, covering interactions in the CDS and 3′UTR separately, were trained and combined linearly to produce a final score. Testing against an independent proteomics data set, DIANA-microT-CDS outperformed all published models. As noted in Section 2.1, interactions observed in CLIP-seq data often do not mediate translational repression, which potentially introduces unwanted biases when used for ML model training.

Very recently, Liu et al. combined their newly generated RNA-seq-based data set of 25 individual miRNAs overexpressed in HeLa cells with CLIP-seq data for the training of the novel **MiRTarget** v4.0 model [39]. They combined the interaction pairs, determined by using their RNA-seq data set, with interaction pairs determined based on publicly available CLIP- and CLASH-seq data sets. They employed an SVM with radial basis function (RBF) kernel to train their model. To retrieve a metric for miRNA-mRNA pairs from their target-site predictions, they used a simple multiplication of per-binding site predictions. Testing their model on independent CLIP and microarray data sets showed that they outperform previously published methods including TargetScan (v7) and DIANA-microT.

Lately, efforts have been undertaken to also apply modern DNN methods to MIP. In contrast to the methods described above, DNNs usually do not depend on engineered features, but extract useful representations on their own, as exemplified in Fig. 1b.

Table 1
Properties of established and Deep Learning-based miRNA target prediction tools.

	ML type	Accessibility	Organisms	Last update	Features	Output	Training data set	Independent test data set	Particularities	References
Established methods										
TargetScan	Linear regression	Web, Dataset, Source code	mmu, dme, hsa, cel, re	2015	Features from Fig. 1 + ORF length + nucleotide identity features for position 1 and 8	cont.	74 individual miRNA transfections and subsequent MicroArray readout in HeLa cells	7 individual miRNA transfections and subsequent MicroArray readout in HCT116 cells; Experimentally validated interactions; CLIP-seq data set	- Strong focus on feature engineering - Stepwise regression for feature selection	First: [6] Latest: [36]
RNA22	No ML used	Web, Dataset	mmu, hsa, dme, cel, *any*	2019	folding energy, heteroduplex	cont.	N/A	N/A	- No ML employed - free binding energy as sole indicator of interaction	[59]
miSTAR	2-layer model using logistic regression and random forest	Web	hsa	2016	Features from Fig. 1 except supplementary binding	cont.	Luciferase reporter assay for 17 human mRNAs and 470 miRNA mimics	N/A	- Stacked model for ML based estimation of cooperative repression	[37]
MiRTarget	SVM	Web, Dataset	mmu, hsa, rno, clf, ggm	2019	96 features including the features from Fig. 1	cont.	25 individual miRNA transfections and subsequent RNA-seq in HeLa cells	CLIP-seq data set; concurrent knockout of 25 miRNAs and subsequent MicroArray readout	N/A	First: [86] Latest: [39]
DNN-based methods										
miRAW	"Normal" DNN	Dataset, Source code	hsa	2018	Nucleotide identities of miRNA (30 nts) and target site (40 nts)	t/n/f	Positive: CLASH and CLIP data set intersected with TarBase and mirTarBase validated interactions Negative: plausible interaction sites in validated (tarbase and mirtarbase) negative pairs	5 individual miRNA transfections and subsequent MicroArray readout	- Very broad identification of potential interaction sites - miRNA- and interaction site-unrelated features are applied *a posteriori* as filters	[73]
deepTarget	RNN with Autoencoder for unsupervised input representation learning	Source code	hsa	2016	Nucleotide identities of miRNA (30 nts) and target site (30 nts)	t/f	Positive: Experimentally validated interaction data from miRecords Negative: Computational generation of mock miRNAs with corresponding binding sites	N/A	N/A	[72]
DeepMirTar	Stacked Autoencoder	Source code	hsa	2018	750 features grouped in categories "high level", "expert-designed", "low-level" and "raw-data-level"	t/f	Positive: CLASH data set and validated interaction data from miRecords Negative: Computational generation of mock miRNAs with corresponding binding sites	PAR-CLIP based interactions	N/A	[74]
Biochemical affinity CNN[1]	CNN	Source code	hsa	2018	Nucleotide identities of miRNA (10 nts) and target site (12 nts)	cont.	RISC binding affinity data for 6 individual miRNAs (AGO2 RNA bind-n-seq)	N/A	CNN used for prediction of miRNA-target binding affinities; affinities are forwarded into a separate regressor for final miRNA interaction efficacy prediction	[48]

t/f (true/false) denotes binary classification, t/n/f (true/neutral/false) denotes ternary classification, cont. denotes a continuous regression.

[1] No name given by publication.

Abbreviations N/A (Not available, ML (Machine Learning, miRNA (microRNA, CNN (Convolutional Neural Network, SVM (Support Vector Machine, DNN (Deep Neural Network, ORF (Open reading frame, CLIP-seq (Cross-Linking ImmunoPrecipitation high-throughput sequencing, PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced CLIP-seq, CLASH (Cross-linking, Ligation and Sequencing of Hybrids, RISC (RNA induced silencing complex hsa (Homo sapiens), mmu (Mus musculus), dsa (Drosophila melanogaster), cel (Caenorhabditis elegans), dre (Danio Rerio), rno (Rattus norvegicus), clf (Canis lupus familiaris), ggm (Gallus gallus domesticus).

The **deepTarget** model (2016) uses two Autoencoders to learn compact representations for the miRNA and mRNA target site sequences separately. These representations are then processed in an RNN layer followed by an ordinary NN layer, which determines whether the observed interaction is functional or not [72]. This model runs without any engineered features and outperforms established models like miRanda and TargetScan by 26% (F1-score improvement from 0.722 and 0.7271 to 0.9105). However, no independent test data set was used for the evaluation of the model.

Pla et al. also do not use any engineered features in their **miRAW** model (2018), and argue that it reduces the bias introduced by manual feature engineering. In contrast to the deepTarget model, they test their model on two independent test sets, reporting a performance increase of 23% over the second-best model (DIANA-microT-CDS [71], F1-score 0.602 to 0.744). Their architecture also relies on a pretrained Autoencoder of 5 layers. Subsequently, three fully connected NN layers process the Autoencoder's representation to a MIP [73].

Wen et al. took a different approach for the **DeepMirTar** model (2018) [74]. They provided a very large set of different kinds of features to their NN, including seed matching-, free energy-, sequence composition-, raw nucleotide identity- and site location-, conservation- and accessibility-features. Their architecture relies on Autoencoder layers, also resulting in two outputs, which indicate whether the interaction is functional or not. By comparing their DNN method with non-NN ML methods, as well as other published methods, they demonstrate the superior performance of their model by an increase of 21% to the second-best model (TarPmiR [75], AUC measure 0.8021 to 0.9793). Of note, they compared all methods on their own data set and only provided performance data for a very limited independent test data set of 48 positive observations extracted from a CLIP-seq experiment.

McGeary et al. performed *in vitro* binding affinity experiments demonstrating a high correlation between their measured affinities and mRNA repressions, measured from miRNA-overexpression experiments. In order to be able to generalize their experimental findings from few individual miRNAs onto the whole set of annotated miRNAs, they used a CNN, which predicts binding affinities based on provided miRNA-target pair sequences. Their measurements and predictions were more accurate than previous binding affinity approximations and they were able to improve predictive performance of miRNA mediated target repression over the latest TargetScan model (v7) by 31% (r^2 score of 0.16 to 0.21). This improvement was achieved by passing the predicted binding affinity from the CNN into a regression model (used for repression prediction) and training both models simultaneously [48].

Finally, in order to improve prediction performance, some groups have developed databases and methods to combine the output of previously described existing models. Oliveira et al. took a simple but effective approach by combining the predictions of four established models. By comparing their results to a set of highly validated miRNA-target interaction pairs, they showed that the union of predictions from TargetScan [36], miRanda-mirSVR [66] and RNA22 [59] outperforms all other combinations as well as the individual models themselves [76]. Davis et al. went one step further by not only using *in silico* modeling scores from established prediction models, but by also integrating data of experimentally validated miRNA interaction pairs, gene and protein expression and CLIP-seq experiments for their **metaMIR** method [77]. The method aims at identifying individual miRNAs that co-regulate a set of provided genes. **MiRWalk** [78] and **miRGate** [79] are databases providing combined resources of integrated data and up to date predictions from established models as well as experimentally verified interactions. Both offer an accessible web interface for various species and additionally allow for programmable access either via a comprehensive downloadable database (miRWalk) or by an *application programming interface* (miRGate).

3. Discussion

Here, we have provided an overview of established MIP models, including their underlying feature engineering and experimental data and furthermore, reviewed recent DNN-based MIP attempts.

Despite a decade of evolution of MIP models, recent methods might still predict many false positives [80]. Also, recent efforts report major shortcomings of their models, as for example Agarwal et al. who attribute the 85% unexplained variability of their model to secondary effects, experimental noise and imperfections of their model [36]. A set of potential mechanisms, including the methylation of RNAs and the binding of miRNAs to long non-coding RNAs, might compete with miRNA mediated regulation and explain some of the variation [45,81,82]. Here, possible improvement might come from the identification and study of miRNA-mRNA interaction pairs that deviate from model predictions to reveal common factors explaining the model's deviations.

Another caveat, which has not yet received much attention, is the potential cell context-specific variability of miRNA regulation. One study has attributed a large part of the observed variation of miRNA regulation in different cell types, to the differential expression of 3′UTR isoforms [83]. A novel published *in vivo* CLIP method, might highlight Argonaute binding differences across different *in vivo* contexts [44].

Ground truth data sets form the basis for ML and we reviewed a number of experiments that provide relevant data. Unfortunately, most of these experiments have limitations, which potentially introduce modeling biases. For example, a part of the observed transcript repression after overexpression of a specific miRNA might be caused by secondary regulatory events [84]. Although high-throughput experiments show a direct and comprehensive picture of binding affinities for individual miRNAs, this can so far only be performed in an artificial environment outside of the cell [41,48]. Naturally, a ML model can only be as good as the *ground truth* data it is based on, so focus should lie on improving this basis. It has already been shown that the integration of existing models, and thus applied data sets, outperforms the predictive performance of each individual model [76]. Unfortunately, to date only few approaches have focused on integrating different types of data sources and modeling approaches, in order to rule out their individual limitations. For now, we recommend the reader to use multiple MIP methods and to be most confident in the repressive potential of an interaction pair, if it appears in several of the applied methods as functional.

Interestingly, recent findings improve our understanding of miRNA mediated gene regulation and are probably reflected in existing data sets. For example, a structural study further improved our understanding of miRNA binding via the supplementary region of miRNAs [85]. Although it had been suggested earlier that the targeted mRNA can form a small loop between the seed region-binding nucleotides and the supplementary region-binding nucleotides [7], their structural analysis revealed that the targeted mRNA can form an up to 15 nts large bulge between the seed complementary region and the supplementary binding region (visualized in Fig. 1 a). Especially 3′ extended isomiRs, miRNAs that are alternatively processed by the microprocessor or DICER, showed a tremendously increased potential to repress targets with supplementary binding. A recent miRNA binding affinity study brought up further evidence for this, showing that inserts of up to 7 bps in the miRNA target site between the seed region and the 3′ supplementary regions do not reduce binding affinity [41]. Indeed, most ML models need to be frequently updated to make use of such new biological findings.

NNs, with their capability to automatically obtain meaningful features from raw input data, have the potential to increase the amount of information extracted from data sets, including potential biological molecular mechanisms that have not yet been discovered. While state-of-the-art miRNA interaction models rely on a long history of feature engineering and model optimization progress, modern DNN approaches are able to achieve similar performances, ignoring most of these advancements [73].

However, there is still little work proving the generalization capabilities of these models, i.e. how they perform on data, which are completely independent of the model design and training process. To our best knowledge, the reviewed DNN models have not been evaluated by independent research groups. Until now, only work on human data sets has been published and there are neither web interfaces nor genome scale precomputed prediction data sets available. Work in this direction, therefore, needs to include improvements in reproducibility and accessibility. Also, since interpretation of NN models is not as straightforward as for most other types of ML models, researchers need to put increased effort into model interpretability, which is an active field of research [87,88]. Since DNNs identify predictive features automatically, they might describe novel features allowing for a better understanding of underlying biological mechanisms of miRNA-mediated regulation.

CRediT authorship contribution statement

Moritz Schäfer: Data curation, Visualization, Writing - original draft. **Constance Ciaudo:** Conceptualization, Supervision, Writing - review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgments

We would like to thank Tobias A. Beyer and the whole Ciaudo lab for discussions and very helpful comments on the manuscript. This work was supported by a core grant from ETH-Z (supported by Roche) and MS is supported by the SNF grant (31003A_173120).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.02.019.

References

[1] Ebert MS, Sharp PA. Roles for microRNAs in conferring robustness to biological processes. Cell 2012;149:515–24. https://doi.org/10.1016/j.cell.2012.04.005.

[2] Bodak M, Cirera-Salinas D, Luitz J, Ciaudo C. The role of RNA interference in stem cell biology: beyond the mutant phenotypes. J Mol Biol 2017;429:1532–43. https://doi.org/10.1016/j.jmb.2017.01.014.

[3] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009;136:215–33. https://doi.org/10.1016/j.cell.2009.01.002.

[4] Iwakawa H-o, Tomari Y. The functions of microRNAs: mRNA decay and translational repression. Trends Cell Biol 2015;25:651–65. https://doi.org/10.1016/j.tcb.2015.07.011.

[5] Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 2010;466:835–40. https://doi.org/10.1038/nature09267.

[6] Lewis BP, Shih I-H, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell 2003;115:787–98. https://doi.org/10.1016/s0092-8674(03)01018-3.

[7] Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007;27:91–105. https://doi.org/10.1016/j.molcel.2007.06.017.

[8] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007;39:1278–84. https://doi.org/10.1038/ng2135.

[9] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 2005;433:769–73. https://doi.org/10.1038/nature03315.

[10] Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 1959;3:210–29. https://doi.org/10.1147/rd.33.0210.

[11] Russell S. Artificial intelligence: a modern approach. Upper Saddle River, New Jersey: Prentice Hall; 2010.

[12] Bishop CM. Pattern recognition and machine learning. Springer; 2006.

[13] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35:1798–828. https://doi.org/10.1109/tpami.2013.50.

[14] Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Briefings Bioinf 2006;7:86–112. https://doi.org/10.1093/bib/bbk007.

[15] Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol 2008;4:. https://doi.org/10.1371/journal.pcbi.1000173e1000173.

[16] Schölkopf B, Tsuda K, Vert J-P. Support vector machine applications in computational biology. MIT press; 2004.

[17] Che D, Liu Q, Rasheed K, Tao X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. Software tools and algorithms for biological systems. Springer; 2011. p. 191–9.

[18] Qi Y. Random forest for bioinformatics. Ensemble machine learning. Springer; 2012. p. 307–23.

[19] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–17. https://doi.org/10.1093/bioinformatics/btm344.

[20] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol 2016;12:878.

[21] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6. https://doi.org/10.1038/323533a0.

[22] Hinton GE. Reducing the dimensionality of data with neural networks. Science 2006;313:504–7. https://doi.org/10.1126/science.1127647.

[23] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80.

[24] Xu R, Wunsch II D, Frank R. Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. IEEE/ACM Trans Comput Biol Bioinf 2007;4:681–92.

[25] Lee B, Lee T, Na B, Yoon S. DNA-level splice junction prediction using deep recurrent neural networks. arXiv Preprint arXiv:151205135 2015.

[26] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[27] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15:20170387. https://doi.org/10.1098/rsif.2017.0387.

[28] Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nat Methods 2012;9:1134–6. https://doi.org/10.1038/nmeth.2259.

[29] Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: empirical comparisons. Int J Forecast 1992;8:69–80.

[30] Devore JL. Probability and statistics for engineering and the sciences. Cengage Learning 2011.

[31] Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. Int J Data Mining Knowledge Manage Process 2015;5:01–11. https://doi.org/10.5121/ijdkp.2015.5201.

[32] Schrynemackers M, Küffner R, Geurts P. On protocols and measures for the validation of supervised methods for the inference of biological networks. Front Genetics 2013;4. https://doi.org/10.3389/fgene.2013.00262.

[33] Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27:861–74. https://doi.org/10.1016/j.patrec.2005.10.010.

[34] Chicco D. Ten quick tips for machine learning in computational biology. BioData Mining 2017;10. https://doi.org/10.1186/s13040-017-0155-3.

[35] Bartel DP. Metazoan microRNAs. Cell 2018;173:20–51. https://doi.org/10.1016/j.cell.2018.03.006.

[36] Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife 2015;4,. https://doi.org/10.7554/elife.05005.

[37] Van Peer G, De Paepe A, Stock M, Anckaert J, Volders P-J, Vandesompele J, et al. miSTAR: miRNA target prediction through modeling quantitative and qualitative miRNA binding site information in a stacked model structure. Nucleic Acids Res 2016;45:e51–61. https://doi.org/10.1093/nar/gkw1260.

[38] Lipchina I, Elkabetz Y, Hafner M, Sheridan R, Mihailovic A, Tuschl T, et al. Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes Dev 2011;25:2173–86.

[39] Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. Genome Biol 2019;20:18. https://doi.org/10.1186/s13059-019-1629-z.

[40] Nicolas FE. Experimental validation of microRNA targets using a luciferase reporter system. In: MicroRNAs in development. Springer; 2011. p. 139–52.
[41] Becker WR, Ober-Reynolds B, Jouravleva K, Jolly SM, Zamore PD, Greenleaf WJ. High-throughput analysis reveals rules for target RNA binding and cleavage by ago2. Mol Cell 2019. https://doi.org/10.1016/j.molcel.2019.06.012.
[42] Lin C, Miles WO. Beyond clip: advances and opportunities to measure RBP-RNA and RNA-RNA interactions. Nucleic Acids Res 2019;47:5490–501. https://doi.org/10.1093/nar/gkz295.
[43] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010;141:129–41. https://doi.org/10.1016/j.cell.2010.03.009.
[44] Li X, Pritykin Y, Concepcion CP, Lu Y, Rocca GL, Zhang M, et al. High-resolution in vivo identification of miRNA targets by HALO-enhanced AGO2 pulldown. bioRxiv 2019. https://doi.org/10.1101/820548.
[45] Imig J, Brunschweiger A, Brümmer A, Guennewig B, Mittal N, Kishore S, et al. MiR-CLIP capture of a miRNA targetome uncovers a lincRNA h19–miR-106a interaction. Nat Chem Biol 2014;11:107–14. https://doi.org/10.1038/nchembio.1713.
[46] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 2013;153:654–65. https://doi.org/10.1016/j.cell.2013.03.043.
[47] Hausser J, Zavolan M. Identification and consequences of miRNA–target interactions—beyond repression of gene expression. Nat Rev Genet 2014;15:599–612.
[48] McGeary SE, Lin KS, Shi CY, Pham T, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science 2019;eaav1741. https://doi.org/10.1126/science.aav1741.
[49] Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res 2017;46:D296–302. https://doi.org/10.1093/nar/gkx1067.
[50] Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al. Diana-tarbase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic Acids Res 2017;46:D239–45. https://doi.org/10.1093/nar/gkx1141.
[51] Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009;37:D105–10. https://doi.org/10.1093/nar/gkn851.
[52] Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. Briefings Bioinf 2014;16:780–94. https://doi.org/10.1093/bib/bbu044.
[53] Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB. Common features of microRNA target prediction tools. Front Genetics 2014;5. https://doi.org/10.3389/fgene.2014.00023.
[54] Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD. Bioinformatic tools for microRNA dissection. Nucleic Acids Res 2015;44:24–44. https://doi.org/10.1093/nar/gkv1221.
[55] Riffo-Campos, Riquelme I, Brebi-Mieville P. Tools for sequence-based miRNA target prediction: what to choose? Int J Mol Sci 2016;17:1987. https://doi.org/10.3390/ijms17121987.
[56] Roberts JT, Borchert GM. Computational prediction of microRNA target genes, target prediction databases, and web resources. In: Bioinformatics in microRNA research. New York: Springer; 2017. p. 109–22. https://doi.org/10.1007/978-1-4939-7046-9_8.
[57] Chen L, Heikkinen L, Wang C, Yang Y, Sun H, Wong G. Trends in the development of miRNA bioinformatics tools. Briefings Bioinf 2018. https://doi.org/10.1093/bib/bby054.
[58] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte Fuer Chemie Chemical Monthly 1994;125:167–88. https://doi.org/10.1007/bf00818163.
[59] Miranda KC, Huynh T, Tay Y, Ang Y-S, Tam W-L, Thomson AM, et al. A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell 2006;126:1203–17. https://doi.org/10.1016/j.cell.2006.07.031.
[60] Ameres SL, Martinez J, Schroeder R. Molecular basis for target RNA recognition and cleavage by human RISC. Cell 2007;130:101–12. https://doi.org/10.1016/j.cell.2007.04.037.
[61] Robins H, Li Y, Padgett RW. Incorporating structure to predict microRNA targets. Proc Natl Acad Sci 2005;102:4006–9. https://doi.org/10.1073/pnas.0500775102.
[62] Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. Potent effect of target structure on microRNA function. Nat Struct Mol Biol 2007;14:287–94. https://doi.org/10.1038/nsmb1226.
[63] Bartel DP. MicroRNAs. Cell 2004;116:281–97. https://doi.org/10.1016/s0092-8674(04)00045-5.
[64] Majoros WH, Ohler U. Spatial preferences of microRNA targets in 3' untranslated regions. BMC Genomics 2007;8:152. https://doi.org/10.1186/1471-2164-8-152.
[65] Hausser J, Landthaler M, Jaskiewicz L, Gaidatzis D, Zavolan M. Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. Genome Res 2009;19:2009–20. https://doi.org/10.1101/gr.091181.109.
[66] Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11:R90. https://doi.org/10.1186/gb-2010-11-8-r90.
[67] Arvey A, Larsson E, Sander C, Leslie CS, Marks DS. Target mRNA abundance dilutes microRNA and siRNA activity. Mol Syst Biol 2010;6:363. https://doi.org/10.1038/msb.2010.24.
[68] Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 2011;18:1139–46. https://doi.org/10.1038/nsmb.2115.
[69] Chatfield C. Model uncertainty, data mining and statistical inference. J Royal Statistical Soc: Series A (Statistics in Society) 1995;158:419–44.
[70] Flom PL, Cassell DL. Stopping stepwise: why stepwise and similar selection methods are bad, and what you should use. NorthEast SAS Users Group (NESUG): Statistics and Data Analysis; 2007.
[71] Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. Functional microRNA targets in protein coding sequences. Bioinformatics 2012;28:771–6. https://doi.org/10.1093/bioinformatics/bts043.
[72] Lee B, Baek J, Park S, Yoon S. DeepTarget. Proceedings of the 7th acm international conference on bioinformatics, computational biology, and health informatics - bcb '16, 2016. https://doi.org/10.1145/2975167.2975212.
[73] Pla A, Zhong X, Rayner S. miRAW: a deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. PLoS Comput Biol 2018;14:. https://doi.org/10.1371/journal.pcbi.1006185e1006185.
[74] Wen M, Cong P, Zhang Z, Lu H, Li T. DeepMirTar: a deep-learning approach for predicting human miRNA targets. Bioinformatics 2018;34:3781–7. https://doi.org/10.1093/bioinformatics/bty424.
[75] Ding J, Li X, Hu H. TarPMir: a new approach for microRNA target site prediction. Bioinformatics 2016;32:2768–75. https://doi.org/10.1093/bioinformatics/btw318.
[76] Oliveira AC, Bovolenta LA, Nachtigall PG, Herkenhoff ME, Lemke N, Pinhal D. Combining results from distinct microRNA target prediction tools enhances the performance of analyses. Front Genetics 2017;8. https://doi.org/10.3389/fgene.2017.00059.
[77] Davis JA, Saunders SJ, Mann M, Backofen R. Combinatorial ensemble miRNA target prediction of co-regulation networks with non-prediction data. Nucleic Acids Res 2017;45:8745–57. https://doi.org/10.1093/nar/gkx605.
[78] Sticht C, De La Torre C, Parveen A, Gretz N. MiRWalk: an online resource for prediction of microRNA binding sites. PLoS ONE 2018;13:. https://doi.org/10.1371/journal.pone.0206239e0206239.
[79] Andrés-León E, Peña DG, Gómez-López G, Pisano DG. MiRGate: a curated database of human, mouse and rat miRNA-mRNA targets. Database 2015. https://doi.org/10.1093/database/bav035.
[80] Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, et al. MicroRNA target prediction programs predict many false positives. Genome Res 2016;27:234–45. https://doi.org/10.1101/gr.205146.116.
[81] Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. Nat Cell Biol 2014;16:191.
[82] Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? Cell 2011;146:353–8. https://doi.org/10.1016/j.cell.2011.07.014.
[83] Nam J-W, Rissland OS, Koppstein D, Abreu-Goodger C, Jan CH, Agarwal V, et al. Global analyses of the effect of different cellular contexts on microRNA targeting. Mol Cell 2014;53:1031–43. https://doi.org/10.1016/j.molcel.2014.02.013.
[84] Thomson DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. Nucleic Acids Res 2011;39:6845–53. https://doi.org/10.1093/nar/gkr330.
[85] Sheu-Gruttadauria J, Xiao Y, Gebert LF, MacRae IJ. Beyond the seed: structural basis for supplementary microRNA targeting by human Argonaute2. EMBO J 2019;e101153.
[86] Wong N, Wang X. MiRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res 2014;43:D146–52. https://doi.org/10.1093/nar/gku1104.
[87] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th international conference on machine learning-volume 70, JMLR.org. p. 3319–28.
[88] Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res 2020. https://doi.org/10.1101/gr.247494.118.