

RESEARCH

Open Access



# MiREx: mRNA levels prediction from gene sequence and miRNA target knowledge

Elena Pianfetti<sup>1</sup>, Marta Lovino<sup>1\*</sup>, Elisa Ficarra<sup>1</sup> and Loredana Martignetti<sup>2,3,4,5\*</sup>

\*Correspondence:  
marta.lovino@unimore.it;  
loredana.martignetti@curie.fr

<sup>1</sup> Department of Engineering,  
University of Modena and Reggio  
Emilia, Via Vivarelli 10/1,  
Modena 41225, Italy

<sup>2</sup> Institut Curie, Rue d'Ulm 26,  
Paris 75005, France

<sup>3</sup> Inserm U900, Paris, France

<sup>4</sup> CBIO-Centre for Computational  
Biology, Paris, France

<sup>5</sup> PSL Research University, Paris,  
France

## Abstract

Messenger RNA (mRNA) has an essential role in the protein production process. Predicting mRNA expression levels accurately is crucial for understanding gene regulation, and various models (statistical and neural network-based) have been developed for this purpose. A few models predict mRNA expression levels from the DNA sequence, exploiting the DNA sequence and gene features (e.g., number of exons/introns, gene length). Other models include information about long-range interaction molecules (i.e., enhancers/silencers) and transcriptional regulators as predictive features, such as transcription factors (TFs) and small RNAs (e.g., microRNAs - miRNAs). Recently, a convolutional neural network (CNN) model, called Xpresso, has been proposed for mRNA expression level prediction leveraging the promoter sequence and mRNAs' half-life features (gene features). To push forward the mRNA level prediction, we present miREx, a CNN-based tool that includes information about miRNA targets and expression levels in the model. Indeed, each miRNA can target specific genes, and the model exploits this information to guide the learning process. In detail, not all miRNAs are included, only a selected subset with the highest impact on the model. MiREx has been evaluated on four cancer primary sites from the genomics data commons (GDC) database: lung, kidney, breast, and corpus uteri. Results show that mRNA level prediction benefits from selected miRNA targets and expression information. Future model developments could include other transcriptional regulators or be trained with proteomics data to infer protein levels.

**Keywords:** CNN, DNA, Gene expression levels, MiRNAs, Promoter, Sequence

## Introduction

Proteins enable the proper functioning of human organisms by regulating their biological processes and enabling cells to react in response to external and internal stimuli. Proteins are produced in ribosomes during the translation process from mRNA molecules.

The quantity of circulating mRNA (mRNA expression level) is crucial in understanding transcription and translation processes, as according to Liu and Vogel [1, 2] it is responsible for 40–50% of the variability in protein levels.

Organisms store the information that regulates their biological processes in DNA, which can be sequenced and written as a sequence of four letters representing the four amino acids (ACGT, Adenine, Cytosine, Guanine, and Thymine, respectively).



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

However, the mechanism from DNA sequence to mRNA levels is not straightforward, and many processes are still unknown.

Post-transcriptional regulators are known to regulate the expression levels through covalent or structural mRNA modifications. Among post-transcriptional regulators, miRNAs are small RNA molecules, usually 22 bp long [3], which can repress the mRNA translation by binding to target mRNAs.

Depending on the extent of the mRNA-miRNA binding site, different repression mechanisms can occur:

- slicing of the mRNA,
- shortening of the poly-A cap of the mRNA, leading to the degradation of the mRNA,
- inhibition of mRNA translation, making it less efficient.

To the best of our knowledge, while some state-of-the-art (SOTA) works investigate the miRNA regulatory effect, none currently incorporate miRNA expression levels in predicting mRNA levels from the DNA sequence.

In this work we present miREx, a Convolutional Neural Network (CNN) model for predicting mRNA expression levels from gene sequence and miRNA post-transcriptional information. miREx's architecture is inspired by Xpresso [4], a SOTA model for mRNA level prediction that exploits DNA sequence and gene features (e.g., number of exons/introns, gene length).

miREx predicts mRNA levels in four primary sites from The Cancer Genome Atlas (TCGA), namely lung, kidney, breast, and corpus uteri. The results show that by including selected miRNA expression levels, the model reaches higher performances.

### Related works

The ability to predict mRNA expression levels is fundamental for understanding the transcription process and the roles different regulatory molecules play. Indeed, some models leverage gene expression levels for patient stratification [5] and for predicting mRNA expression levels using other features instead of the DNA sequence. Commonly used features are TFs [6–8], chromatin features [9, 10], histone modifications [11], or their combination [12, 13]. These models usually outperform methods that use only the DNA sequence (like [14, 15]).

Concerning models that use the DNA sequence only, the ones with better performances exploit deep learning architectures. For instance, previous studies including Xpresso [4, 16, 17] use convolutional architectures, while models including Enformer [18, 19] use Transformers.

Indeed, Enformer is a state-of-the-art (SOTA) model with a transformer architecture. One of its inherent limitations is the challenge of incorporating information beyond the sequence it was designed to process. This limitation stems from the sequential nature of transformers, which excel at capturing long-range dependencies within sequences but are less accommodating for integrating additional contextual information. In contrast, models with convolutional architectures are more flexible in

this regard. CNNs are well-suited for handling multi-channel inputs and can incorporate external information to enhance their performance. Indeed, recently, CNNs have been exploited in many medical and biological tasks [20–22].

The CNN-based SOTA model for mRNA level prediction from DNA sequence is Xpresso, that receives two inputs: the DNA sequence surrounding the Transcription Start Site (TSS) and the mRNA half-life features associated with each gene. The mRNA degradation rate impacts the steady-state mRNA levels [23], and the half-lives of mRNA molecules are one way to consider this process in the model.

It was shown in [24, 25] that features linked to mRNA half-life encompass specific aspects of gene structure. These include the length and CG content of particular gene regions, such as the 5' UnTranslated Region (UTR), Open Reading Frame (ORF), and 3' UTR, intron length and exon junction. These eight half-life features are considered in the Xpresso and miREx models.

## Materials and methods

In the following sections, we provide a comprehensive description of our methodology, encompassing a description of the data, the processing, and the architecture used.

### Data description

In this section, we provide an overview of the different data used by the model, along with references to sources for data acquisition. miREx predicts mRNA expression levels by exploiting a portion of the gene DNA sequence, mRNA half-life features, and miRNA targets.

### Sequences and half-life features

Sequences and mRNA half-life features were obtained from [4]. A critical role in the transcription process is played by the promoter, a region of the DNA sequence, usually located upstream of the TSS [26] and spanning a few thousand base pairs. In addition, other regulatory sequences around the promoters have an essential role in transcription. Therefore, Xpresso and miREx models use a sequence centered on the TSS, thus including the promoter and parts of other regulatory elements. The sequence used has a length of 10,500 bases, 3000 preceding the TSS and 7500 following it.

### miRNA targeting

TargetScan [27] is a state-of-the-art database for miRNA annotation and target description. Cumulative Weighted Context++ Scores (CWCS) from TargetScan were downloaded to embed miRNA target-specific information in the model. This score is a metric to assess the likelihood of a given miRNA binding to a target mRNA. Consequently, for each miRNA-gene pair, we have a CWCS. Therefore, our model also considers the effectiveness of miRNAs in the repression of gene expression.

### mRNA and miRNA expression values

mRNA and miRNA data are obtained from the Genomic Data Commons (GDC) portal [28] for four primary sites: lung, kidney, breast, and corpus uteri. Each primary site can contain data from different cancer subtypes.

**Lung** Data from LUng ADenocarcinoma (LUAD) and LUng Squamous Cell carcinoma (LUSC) was downloaded. Mesothelioma (MESO) data was discarded since miRNA expression levels were unavailable. The dataset is composed of 982 samples (507 for LUAD and 475 for LUSC).

**Kidney** Data from KIDney CHromophobe (KICH), KIDney Renal Papillary cell carcinoma (KIRP), and KIDney Renal Clear cell carcinoma (KIRC) was downloaded. Sarcoma (SARC) was excluded due to the low number of samples. The dataset is composed of 871 samples (66 for KICH, 290 for KIRP, and 515 for KIRC).

**Breast** Data from Breast Invasive Carcinoma (BRCA) was downloaded. Diffuse Large B-Cell Lymphoma (DLBC) was excluded due to the low number of samples. The dataset is composed of 1076 BRCA samples.

**Corpus Uteri** Data from Uterine Corpus Endometrial Carcinoma (UCEC) was downloaded. Sarcoma (SARC) was excluded due to the low number of samples. The dataset is composed of 536 UCEC samples.

### Data processing

In each of the four datasets acquired, we preserved only those samples that included mRNA and miRNA expression values.

To make a fair comparison with Xpresso [4], histone and Y chromosome genes were discarded, and only protein-coding genes were kept. Ultimately, 18,377 genes were used by Xpresso. Out of these, 18,347 genes also had corresponding expression values in the GDC dataset and were subsequently employed by miREx. The miRNA expression values data downloaded from GDC comprises 1881 miRNAs. Of those, only 243 had known gene targets in the TargetScan database and were used in our model. The CWCS was extracted for each remaining miRNA-target pair from TargetScan.

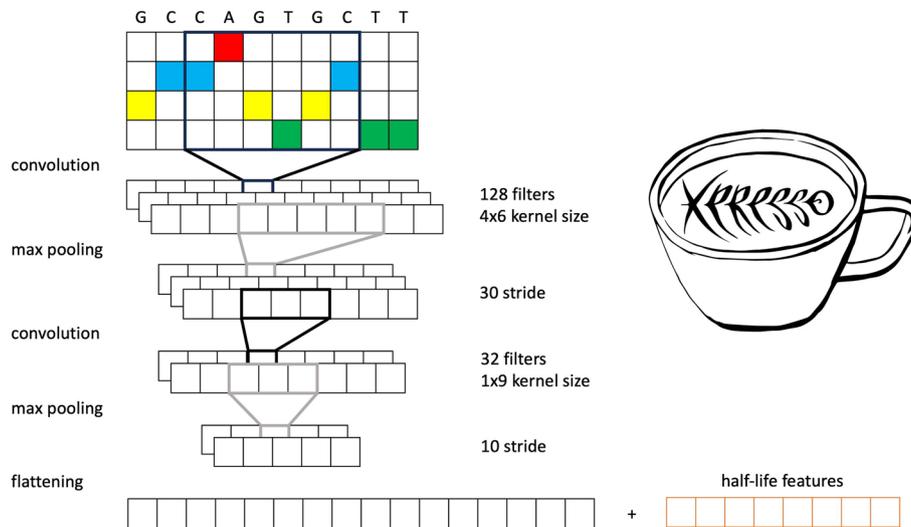
We computed the mean expression levels of each gene and miRNA across all cancer primary sites. In cases where multiple subtypes were present within a primary site, we also calculated the overall mean expression across all subtypes for that site. The raw counts of mRNA and miRNA were log-normalized  $x \leftarrow \log_{10}(x + 0.1)$  to reduce the right skew of the data.

### Method

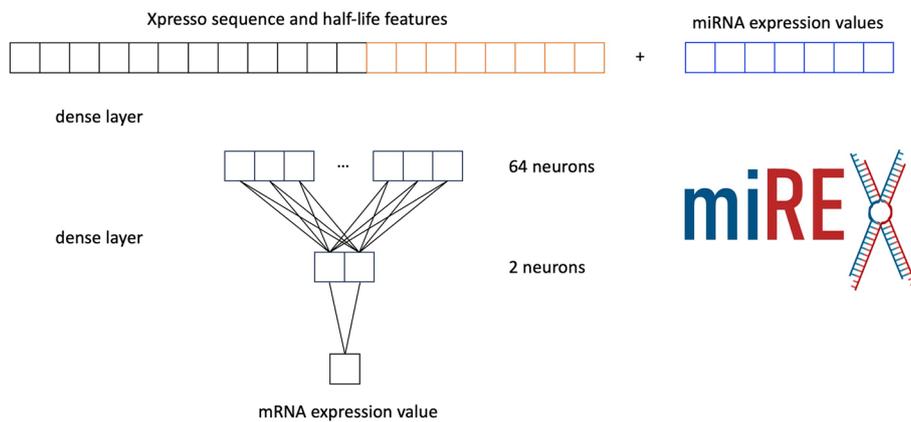
MiREx exploits the Xpresso CNN architecture as a backbone. It consists of convolutional and max-pooling layers applied on the one-hot encoded DNA sequence. The max pooling output is flattened and concatenated to the previously described eight half-life features. A complete figure describing the Xpresso backbone is reported in Fig. 1. In our model, miRNA expression levels are also concatenated to the DNA sequence and half-life features, as shown in Fig. 2. Finally, two densely connected layers output the results.

Post-transcriptional information is exploited, providing gene-specific miRNA expression levels after the last max pooling layer as a vector. Each element of the vector  $x$  encodes miRNA  $i$ .  $x$  is specific for each gene, according to which miRNAs target it. For each gene,  $x_i$  is the expression level of miRNA  $i$  if miRNA  $i$  targets the gene, 0 otherwise. Table 1 reports an example of  $x$  vectors for some genes.

The dataset was partitioned into training, validation, and test sets. For a fair comparison, we employed the same split used by Xpresso (16,348 genes in training, 1000



**Fig. 1** The encoded input sequences are processed by the Xpresso backbone architecture and concatenated to the half-life features [4]



**Fig. 2** MiREx incorporates miRNA expression and targeting information by concatenating it with sequence and gene features

**Table 1** Example of matrix obtained by miRNA targeting

	<i>miRNA</i> <sub>1</sub>	<i>miRNA</i> <sub>2</sub>	...	<i>miRNA</i> <sub>N</sub>
<i>Gene</i> <sub>1</sub>	0	1.34		0
<i>Gene</i> <sub>2</sub>	0	0		1.26
...				
<i>Gene</i> <sub>N</sub>	1.1	0		0

in validation, and 999 in test). We used an early stopping procedure to avoid overfitting. Indeed, the training was stopped if the loss on the validation set did not improve for 20 epochs. The rectified linear unit (ReLU) was used as the activation function for convolutional and dense layers. Stochastic Gradient Descent (SGD) served as the

optimization algorithm. The Mean Square Error (MSE) is minimized during the network training.

Performances were evaluated with the coefficient of determination ( $R^2$ ), which measures how well the model predicts the ground truth. Given the vector with ground truth expression levels  $y$ , their mean value  $\bar{y}$ , and the vector with the predicted expression levels  $\hat{y}$ , the Residual Sum of Squares and the Total Sum of Squares are defined as follows:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Finally, the value of  $R^2$  is computed as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

The parameters used for the training of the model are shown in Table 2.

**Xpresso** does not consider the effect of regulatory molecules. Therefore, upregulated genes (e.g., genes regulated by enhancers that increase the transcription rate) should have positive residuals, and downregulated genes (e.g., genes targeted by miRNAs that repress transcription) should have negative residuals. Where residuals are computed as *mRNA ground truth value – mRNA predicted value*.

We train a model using all miRNAs for which we have both expression levels and miRNA targeting information. In the [results](#) section, we call this model **AllMirna**. This method considers targeting information but not the CWCS information from TargetScan.

A gene targeted by a miRNA can have negative residuals that depend on the miRNA's effectiveness at repressing transcription. TargetScan provides the CWCS, a measure of the likelihood of miRNA binding for each target gene. The lower the CWCS, the greater the probability that the miRNA will bind to the mRNA, repressing its transcription. Consequently, there should be a positive correlation between CWCS and residuals (the more negative the residual, the more negative the CWCS). Therefore, for each miRNA, we computed the Spearman correlation between the residuals and the CWCS, and the ten miRNAs with the highest correlation were chosen. Our model, **miREx**, uses target information from those ten miRNAs.

Although miRNAs can directly impact translation, lowering mRNA expression, indirect regulations could lead to the opposite effect. We also considered the correlation

**Table 2** Parameter configuration values

Parameter name	Parameter value
Batch size	32
Learning rate	0.0005
Number of epochs	100
Early stopping patience	20

between the absolute values of the residuals and the CWCS. The ten miRNAs with the highest correlation were selected to train another model. In the [results](#) section, we call this model **AbsCorr**.

In the end, four different models were trained:

- proposed model: **miREx**,
- sota competitor: **Xpresso**,
- additional configuration: **AllMirna**,
- additional configuration: **AbsCorr**.

The models were trained for each primary site and cancer subtype. Multiple runs were executed, and the best ten were considered for the results.

## Results

In this section, we present the results obtained from the four datasets.

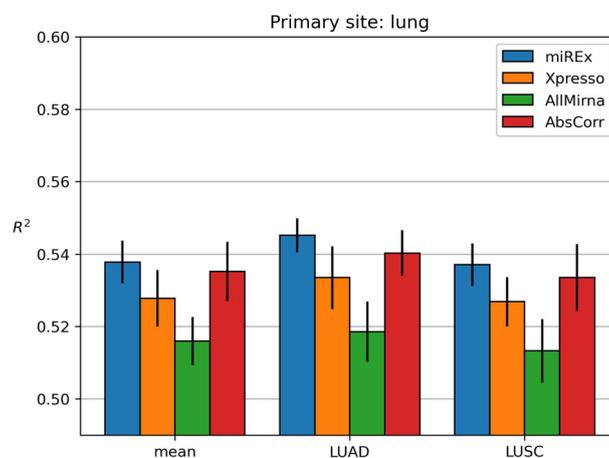
As previously outlined in Sect. [Materials and methods](#), our approach involves training four distinct models. miREx, our proposed model, leverages multiple aspects of miRNAs (expression levels, targeting, and CWCS) as well as sequence and gene features. In contrast, the state-of-the-art model (Xpresso) only uses sequence and gene features. AllMirna includes expression levels and targeting of miRNA but does not take into account the CWCS. Finally, AbsCorr, similarly to miREx, considers all available information for miRNAs; however, AbsCorr examines indirect regulation.

In all the plots in the [results](#) section, we display the mean of the  $R^2$  values along with their corresponding 95% confidence intervals.

### Lung

Figure 3 shows the results obtained on the lung dataset. MiREx and the other models were trained with three targets: LUAD, LUSC, and their mean.

Firstly, we compared the Xpresso results with those obtained with the AllMirna model. The  $R^2$  decreased with the inclusion of miRNA expression levels in all three



**Fig. 3** Results obtained on lung dataset

cases (LUAD, LUSC, and mean). The differences are statistically significant, with p-values < 0.05. MiREx and AbsCorr were trained using only a subset of the miRNAs, and while both approaches lead to higher  $R^2$  in comparison to Xpresso, statistical tests show that only miREx’s results are statistically significant (p-values < 0.05).

Table 3b contains the subsets of ten miRNAs selected by the two methods in the three categories mean, LUAD and LUSC.

There is a significant overlap in the miRNAs chosen by the models in the three categories. Table 3 shows that, with miREx, seven miRNAs are common in the three categories, and two are common between mean and LUSC. Table 3a shows the miRNAs chosen by AbsCorr. Four miRNAs are common, and five are in two categories: mean and LUAD.

There are almost no miRNAs in common between the two methods. mir-23a and mir-23b are the only two in common.

### Kidney

Figure 4 shows the results obtained on the kidney dataset. The models were trained with four targets: KICH, KIRC, KIRP, and the mean of all three classes.

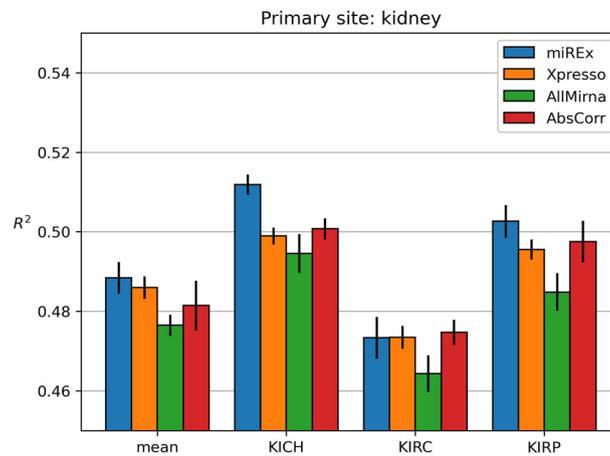
We can see that in all four cases, the AllMirna model leads to worse results. In two classes (KICH and KIRP), we can see that using a selected subset of miRNAs leads to better predictions when compared to Xpresso. In particular, with miREx, the results are statistically significant (p-value << 0.05). The third class, KIRC, has a different trend. In this case, our model does not seem to bring an improvement in the results. The results on the mean of all three subtypes show an improvement with our model. However, it is not statistically significant.

Table 4b displays the miRNAs selected by MiREx and AbsCorr. Again, there is a high overlap between miRNAs chosen by the same method in the different classes, while there is a low overlap between the two techniques.

**Table 3** Lists of miRNAs chosen for each cancer subtype of the lung site by miREx and AbsCorr methods

(a) miREx			(b) AbsCorr		
Mean	LUAD	LUSC	Mean	LUAD	LUSC
<b>mir-23a</b>	<b>mir-23a</b>	<b>mir-23a</b>	<b>mir-15a</b>	<b>mir-15a</b>	<b>mir-15a</b>
<b>mir-23b</b>	<b>mir-23b</b>	<b>mir-23b</b>	<b>mir-15b</b>	<b>mir-15b</b>	<b>mir-15b</b>
<b>mir-101-2</b>	<b>mir-101-2</b>	<b>mir-101-2</b>	<b>mir-200c</b>	<b>mir-200c</b>	<b>mir-200c</b>
<b>mir-145</b>	<b>mir-145</b>	<b>mir-145</b>	<b>mir-340</b>	<b>mir-340</b>	<b>mir-340</b>
<b>mir-199b</b>	<b>mir-199b</b>	<b>mir-199b</b>	<b>mir-23b</b>	<b>mir-23b</b>	mir-130a
<b>mir-206</b>	<b>mir-206</b>	<b>mir-206</b>	<b>mir-195</b>	<b>mir-195</b>	mir-130b
<b>mir-655</b>	<b>mir-655</b>	<b>mir-655</b>	<b>mir-424</b>	<b>mir-424</b>	mir-301b
<b>mir-19a</b>	mir-20b	<b>mir-19a</b>	<b>mir-429</b>	<b>mir-429</b>	mir-301a
<b>mir-506</b>	mir-93	<b>mir-506</b>	<b>mir-497</b>	<b>mir-497</b>	mir-454
mir-140	mir-101-1	mir-381	mir-200b	mir-23a	mir-506

Common miRNAs are in bold text



**Fig. 4** Results obtained on kidney dataset

**Table 4** Lists of miRNAs chosen for each cancer subtype of the kidney site by miREx and AbsCorr methods

miREx				AbsCorr			
KICH	KIRP	KIRC	Mean	KICH	KIRP	KIRC	Mean
<b>mir-1</b>	<b>mir-1</b>	<b>mir-1</b>	<b>mir-1</b>	<b>mir-124-1</b>	<b>mir-124-1</b>	<b>mir-124-2</b>	<b>mir-124-1</b>
<b>mir-124-1</b>	<b>mir-124-1</b>	<b>mir-124-1</b>	<b>mir-124-1</b>	<b>mir-30a</b>	<b>mir-30a</b>	<b>mir-30a</b>	<b>mir-30a</b>
<b>mir-124-2</b>	<b>mir-124-2</b>	<b>mir-124-2</b>	<b>mir-124-2</b>	<b>mir-30b</b>	<b>mir-30b</b>	<b>mir-30b</b>	<b>mir-30b</b>
<b>mir-142-1</b>	<b>mir-142-1</b>	<b>mir-142-1</b>	<b>mir-142-1</b>	<b>mir-30c</b>	<b>mir-30c</b>	<b>mir-30c</b>	<b>mir-30c</b>
<b>mir-142-2</b>	<b>mir-142-2</b>	<b>mir-142-2</b>	<b>mir-142-2</b>	<b>mir-30d</b>	<b>mir-30d</b>	<b>mir-30d</b>	<b>mir-30d</b>
<b>mir-145</b>	<b>mir-145</b>	<b>mir-145</b>	<b>mir-145</b>	<b>mir-340</b>	<b>mir-340</b>	<b>mir-340</b>	<b>mir-340</b>
<b>mir-199a</b>	<b>mir-199a</b>	<b>mir-199a</b>	<b>mir-199a</b>	<b>mir-429</b>	<b>mir-429</b>	mir-506	<b>mir-429</b>
<b>mir-206</b>	<b>mir-206</b>	<b>mir-206</b>	<b>mir-206</b>	<b>mir-30e</b>	<b>mir-30e</b>	mir-23c	let-7i
<b>mir-506</b>	<b>mir-506</b>	<b>mir-506</b>	<b>mir-506</b>	<b>mir-200b</b>	<b>mir-200b</b>	mir-23a	let-7g
<b>mir-199b</b>	mir-325	<b>mir-199b</b>	<b>mir-199b</b>	<b>mir-200c</b>	<b>mir-200c</b>	mir-23b	let-7b

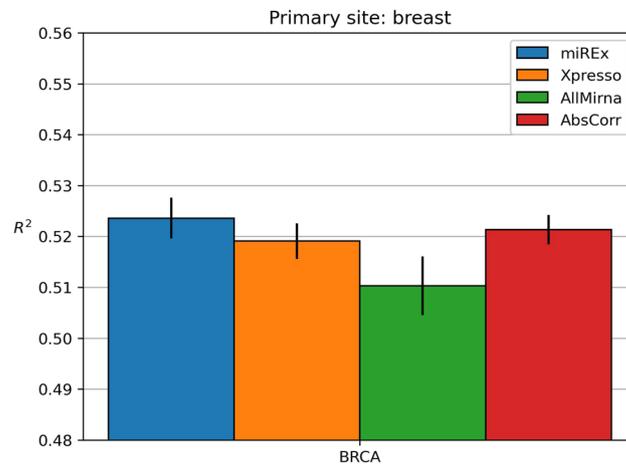
Common miRNAs are in bold text

**Breast**

Figure 5 shows the results obtained on the breast dataset. The models were trained with a single breast cancer subtype: BRCA.

The results are similar to those obtained with the other datasets: including all miRNAs leads to worse predictions than Xpresso (p-value < 0.05) while adding a subset of miRNAs leads to better results, where miREx is the best model. While it is true that the models using a subset of miRNAs have higher mean R<sup>2</sup>, in this case, both miREx and AbsCorr have p-values > 0.05, with miREx coming close (0.06) but not enough to be considered statistically significant.

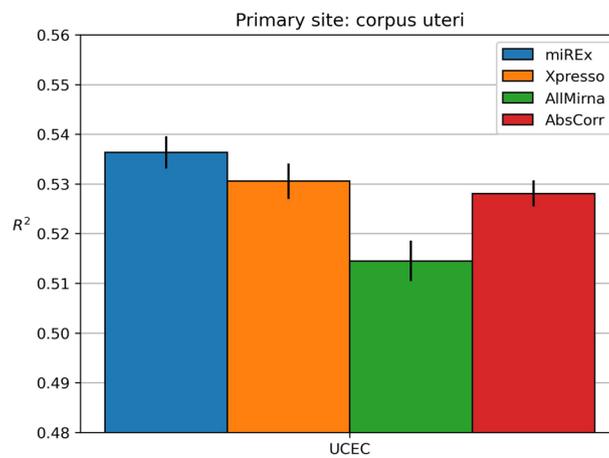
Table 5b shows the miRNA selected by miREx and AbsCorr. The two methods show a limited overlap of miRNAs, as only three are in common.



**Fig. 5** Results obtained on breast dataset

**Table 5** Lists of miRNAs chosen for the BRCA cancer subtype of the breast site by miREx and AbsCorr methods

(a) miREx BRCA	(b) AbsCorr BRCA
mir-1	mir-124-1
mir-124-1	mir-124-2
mir-124-2	mir-15a
mir-142-1	mir-15b
mir-142-2	mir-16
mir-199a	mir-195
mir-199b	mir-200b
mir-206	mir-340
mir-325	mir-497
mir-506	mir-506



**Fig. 6** Results obtained on corpus uteri dataset

### Corpus Uteri

Figure 6 shows the results obtained on the corpus uteri dataset. The models were trained with a single corpus uteri cancer subtype: UCEC.

In this dataset, AllMirna and AbsCorr lead to worse results than Xpresso. MiREx improves the prediction (p-value < 0.05).

Table 6b displays the miRNAs chosen by miREx and AbsCorr. The overlap of miRNAs is minimal, as only two are in common.

### Discussion

As shown in the results Sect. Data description, training the model using all the miRNAs led to lower  $R^2$  values. Most genes are not targeted by a high number of miRNAs. Because of that, the data in the vectors created for each gene was very sparse. It is possible that the model could not extract the information effectively.

To solve this problem, we propose miREx, a model that only uses a subset of miRNAs with a high impact on transcriptional regulation. Since the regulation of miRNAs can be indirect, a second model that considers indirect regulatory mechanisms was trained. Both models lead to higher  $R^2$ , but the results show that direct regulation has a greater effect. Indeed, the miREx model is the one with the best performance.

We examined the miRNAs selected for each model to see if they regulate specific genes and explain why they have a more significant impact. As expected, the two methods chose different miRNAs since they looked at miRNAs with direct or indirect effects on gene regulation. Instead, when there are multiple classes for the same cancer primary site, the miRNAs chosen by the same method in the different categories are very similar.

Many of the miRNAs selected by these methods are known to play various roles in cancer-related processes. For example, miR-23a is involved in many cancers [29], and, in particular, in non-small cell lung cancer (NSCLC) [30, 31], which is a class of lung cancers that includes LUAD and LUSC, two subtypes used in this work. At least one method chose miR-506 for all four datasets. This miRNA is known for its association with different cancers, including kidney [32] and lung [33].

**Table 6** Lists of miRNAs chosen for the UCEC cancer subtype of the corpus uteri site by miREx and AbsCorr methods

(a) miREx UCEC	(b) AbsCorr UCEC
mir-1	mir-124-1
mir-124-1	mir-15a
mir-124-2	mir-15b
mir-142-1	mir-16
mir-142-2	mir-195
mir-145	mir-200b
mir-199a	mir-200c
mir-206	mir-340
mir-325	mir-424
mir-506	mir-506

## Conclusions

This work aimed to build a model that considers the post-transcriptional regulation of miRNAs to predict mRNA levels. MiREx, the model proposed in this paper, is a CNN that takes as input the one-hot encoded sequence of a portion of a gene, mRNA half-life features, and includes information about miRNA expression levels and targeting to predict mRNA expression levels.

Future works might use these methods to predict protein expression levels. MiRNAs should have a more significant impact on protein expression levels since multiple regulation mechanisms would be considered. With mRNAs, the regulatory mechanism under consideration is the one where the miRNAs cause the degradation of the mRNA molecule so that it will not be translated. For proteins, we expect to be able to see the effects of miRNAs on translational inhibition of mRNA without causing its degradation.

## Acknowledgements

The results shown here are based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

## Author Contributions

EP: methodology, conceptualization, software, writing - original draft, writing - review and editing ML: methodology, conceptualization, writing - original draft, writing - review and editing, supervision EF: methodology, conceptualization, writing - review and editing, supervision, funding acquisition LM: methodology, conceptualization, writing - review and editing, supervision, funding acquisition

## Funding

This study was funded by the European Union's Horizon 2020 research and innovation program DECIDER under Grant Agreement 965193. This work obtained additional support from the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, "iPC-individualizedPaediatricCure" (ref. 826121).

## Code availability

The code underlying this article is available in the GitHub repository.

## Availability of data and materials

The datasets analyzed during the current study are available in the [Xpresso](#) repository [4]. The datasets generated during the current study are available in the [miREx](#) zenodo repository.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 30 June 2023 Accepted: 6 November 2023

Published online: 22 November 2023

## References

1. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. *Cell*. 2016;165(3):535–50. <https://doi.org/10.1016/j.cell.2016.03.014>.
2. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*. 2012;13(4):227–32. <https://doi.org/10.1038/nrg3185>.
3. Bartel DP. Metazoan microRNAs. *Cell*. 2018;173:20–51. <https://doi.org/10.1016/j.cell.2018.03.006>.
4. Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep*. 2020. <https://doi.org/10.1016/j.celrep.2020.107663>.
5. Cirrincione G, Ciravegna G, Barbiero P, Randazzo V, Pasero E. The gh-exin neural network for hierarchical clustering. *Neural Netw*. 2020;121:57–73.
6. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan K-K, Dong X, Djebali S, Ruan Y, Davis CA, Carninci P, Lassman T, Gingeras TR, Guigó R, Birney E, Weng Z, Snyder M, Gerstein M. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. 2012;22(9):1658–67.

7. Ouyang Z, Zhou Q, Wong WH. Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci*. 2009;106(51):21521–6. <https://doi.org/10.1073/pnas.0904863106>.
8. Pipoli V, Cappelli M, Palladini A, Peluso C, Lovino M, Ficarra E. Predicting gene expression levels from dna sequences and post-transcriptional information with transformers. *Comput Methods Programs Biomed*. 2022;225: 107035.
9. Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol*. 2011;12(2):15.
10. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, Birney E, Weng Z. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol*. 2012;13(9):53.
11. Karlic R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci*. 2010;107(7):2926–31. <https://doi.org/10.1073/pnas.0909344107>.
12. ...Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, Ebert P, Nordström K, Barann M, Sinha A, Fröhler S, Xiong J, Dehghani Amirabad A, Behjati Ardakani F, Hutter B, Zipprich G, Felder B, Eils J, Brors B, Chen W, Hengstler JG, Hamann A, Lengauer T, Rosenstiel P, Walter J, Schulz MH. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res*. 2016;45(1):54–66. <https://doi.org/10.1093/nar/gkw1061>.
13. McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL. Genome-wide in silico prediction of gene expression. *Bioinformatics*. 2012;28(21):2789–96. <https://doi.org/10.1093/bioinformatics/bts529>.
14. Abdalla M, Abdalla M. A general framework for predicting the transcriptomic consequences of non-coding variation and small molecules. *PLoS Comput Biol*. 2022;18(4):1010028.
15. Bessièrè C, Taha M, Petitprez F, Vandel J, Marin J-M, Bréhélin L, Lèbre S, Lecellier C-H. Probing instructions for expression regulation in gene nucleotide compositions. *PLoS Comput Biol*. 2018;14(1):1005921.
16. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*. 2018;50(8):1171–9.
17. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res*. 2018;28(5):739–50.
18. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021;18(10):1196–203.
19. Stefanini M, Lovino M, Cucchiara R, Ficarra E. Predicting gene and protein expression levels from dna and protein sequences with perceiver. *Comput Methods Programs Biomed*. 2023;234: 107504.
20. Ferretti J, Barbiero P, Randazzo V, Cirrincione G, Pasero E. Towards uncovering feature extraction from temporal signals in deep cnn: the ecg case study. In: 2020 international joint conference on neural networks (IJCNN), pp. 1–7 (2020). IEEE
21. Pollastri F, Parreño M, Maroñas J, Bolelli F, Paredes R, Ramos D, Grana C. A deep analysis on high resolution dermoscopic image classification. *IET Comput Vis*. 2021;15(7):514–26. <https://doi.org/10.1049/cvi2.12048>.
22. Cipriano M, Allegretti S, Bolelli F, Pollastri F, Grana C. Improving segmentation of the inferior alveolar nerve through deep label propagation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 21137–21146 (2022). <https://doi.org/10.1109/CVPR52688.2022.02046>
23. Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*. 2014;2:270.
24. Sharova LV, Sharov AA, Nedorezov T, Piao Y, Shaik N, Ko MSH. Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res*. 2008;16(1):45–58. <https://doi.org/10.1093/dnares/dsn030>.
25. Spies N, Burge CB, Bartel DP. 3' utr-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res*. 2013;23(12):2078–90.
26. Lin H, Liang Z-Y, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans Comput Biol Bioinform*. 2019;16(4):1316–21.
27. Agarwal V, Bell G, Nam J, Bartel D. Predicting effective microrna target sites in mammalian mRNAs. *eLife*. 2015. <https://doi.org/10.7554/ELIFE.05005>.
28. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a shared vision for cancer genomic data. *New Engl J Med*. 2016;375(12):1109–12. <https://doi.org/10.1056/nejmp1607591>.
29. Wang N, Tan H-Y, Feng Y-G, Zhang C, Chen F, Feng Y. Microrna-23a in human cancer: its roles, mechanisms and therapeutic relevance. *Cancers (Basel)*. 2018;11(1):7.
30. Cao M, Li Y, Lu H, Meng Q, Wang L, Cai L, Dong X. Mir-23a-mediated migration/invasion is rescued by its target, IRS-1, in non-small cell lung cancer cells. *J Cancer Res Clin Oncol*. 2014;140(10):1661–70.
31. Qu W-Q, Liu L, Yu Z. Clinical value of microrna-23a upregulation in non-small cell lung cancer. *Int J Clin Exp Med*. 2015;8(8):13598–603.
32. Yang F-Q, Zhang H-M, Chen S-J, Yan Y, Zheng J-H. Mir-506 is down-regulated in clear cell renal cell carcinoma and inhibits cell growth and metastasis via targeting flot1. *PLOS ONE*. 2015;10(3):0120258. <https://doi.org/10.1371/journal.pone.0120258>.
33. Yin M, Ren X, Zhang X, Luo Y, Wang G, Huang K, Feng S, Bao X, Huang K, He X, Liang P, Wang Z, Tang H, He J, Zhang B. Selective killing of lung cancer cells by mirna-506 molecule through inhibiting nf-kb p65 to evoke reactive oxygen species generation and p53 activation. *Oncogene*. 2014;34(6):691–703. <https://doi.org/10.1038/onc.2013.597>.

## Publisher's Note

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