Machine learning for catalysing the integration of noncoding RNA in research and clinical practice

de Gonzalo-Calvo, David; Karaduzovic-Hadziabdic, Kanita; Dalgaard, Louise Torp; Dieterich, Christoph; Perez-Pons, Manel; Hatzigeorgiou, Artemis; Devaux, Yvan; Kararigas, Georgios

Published in:
eBioMedicine

DOI:
10.1016/j.ebiom.2024.105247

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
de Gonzalo-Calvo, D., Karaduzovic-Hadziabdic, K., Dalgaard, L. T., Dieterich, C., Perez-Pons, M., Hatzigeorgiou, A., Devaux, Y., & Kararigas, G. (2024). Machine learning for catalysing the integration of noncoding RNA in research and clinical practice. *eBioMedicine*, *106*, Article 105247. https://doi.org/10.1016/j.ebiom.2024.105247

Machine learning for catalysing the integration of noncoding RNA in research and clinical practice

David de Gonzalo-Calvo,[a,b,*] Kanita Karaduzovic-Hadziabdic,[c] Louise Torp Dalgaard,[d] Christoph Dieterich,[e,f] Manel Perez-Pons,[a,b] Artemis Hatzigeorgiou,[g,h] Yvan Devaux,[i] and Georgios Kararigas[j,**]

[a]Translational Research in Respiratory Medicine, University Hospital Arnau de Vilanova and Santa Maria, IRBLleida, Lleida, Spain
[b]CIBER of Respiratory Diseases (CIBERES), Institute of Health Carlos III, Madrid, Spain
[c]Faculty of Engineering and Natural Sciences, International University of Sarajevo, Sarajevo, Bosnia and Herzegovina
[d]Department of Science and Environment, Roskilde University, Roskilde, Denmark
[e]Klaus Tschira Institute for Integrative Computational Cardiology and Department of Internal Medicine III, University Hospital Heidelberg, Germany
[f]German Center for Cardiovascular Research (DZHK) - Partner Site Heidelberg/Mannheim, Germany
[g]DIANA-Lab, Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
[h]Hellenic Pasteur Institute, Athens, Greece
[i]Cardiovascular Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, Luxembourg
[j]Department of Physiology, Faculty of Medicine, University of Iceland, Reykjavik, Iceland

eBioMedicine
2024;106: 105247
Published Online 18 July 2024
https://doi.org/10.1016/j.ebiom.2024.105247

Summary

The human transcriptome predominantly consists of noncoding RNAs (ncRNAs), transcripts that do not encode proteins. The noncoding transcriptome governs a multitude of pathophysiological processes, offering a rich source of next-generation biomarkers. Toward achieving a holistic view of disease, the integration of these transcripts with clinical records and additional data from omic technologies ("multiomic" strategies) has motivated the adoption of artificial intelligence (AI) approaches. Given their intricate biological complexity, machine learning (ML) techniques are becoming a key component of ncRNA-based research. This article presents an overview of the potential and challenges associated with employing AI/ML-driven approaches to identify clinically relevant ncRNA biomarkers and to decipher ncRNA-associated pathogenetic mechanisms. Methodological and conceptual constraints are discussed, along with an exploration of ethical considerations inherent to AI applications for healthcare and research. The ultimate goal is to provide a comprehensive examination of the multifaceted landscape of this innovative field and its clinical implications.

Keywords: Artificial intelligence; Biomarker; Machine learning; Molecular pathways; Noncoding RNA; Personalised medicine

Introduction

Although approximately 80% of the genome is transcribed, only 1–2% of the human genome represents protein-coding genes.[1] The vast majority of the human transcriptome consists of transcripts that are classically defined as not being translated into functional proteins; noncoding RNAs (ncRNAs). This broad definition encompasses a diverse family of transcripts, ranging from long ncRNAs (lncRNAs, exceeding 200 nucleotides in length) to small ncRNAs (shorter than 200 nucleotides) (Fig. 1). Among these, microRNAs (miRNAs) and lncRNAs have attracted considerable attention and extensive study.

While much remains to be unravelled about the functions of ncRNAs, they are already recognised as crucial contributors to the evolution and development of organismal complexity.[2] A substantial body of evidence highlights their roles as regulators of genome organisation and gene expression, operating at various levels including epigenetic, transcriptional and post-transcriptional.[3] Over the past decades, research has identified ncRNAs as modulators of signalling pathways and mechanisms that govern a spectrum of processes, ranging from differentiation and growth to stress responses,[4] across a variety of cell types and tissues. Given their critical regulatory roles in normal cellular activities, it is not surprising that dysregulation of ncRNAs leads to diseases. Nonetheless, a major challenge in current research is to elucidate the mechanisms of action and functions of ncRNAs, which is essential for defining

*Corresponding author. Translational Research in Respiratory Medicine, University Hospital Arnau de Vilanova and Santa Maria, IRBLleida, Avda. Alcalde Rovira Roure 80, 25198, Lleida, Spain.
**Corresponding author. Department of Physiology, Faculty of Medicine, University of Iceland, Vatnsmyrarvegur 16, 101 Reykjavik, Iceland.
E-mail addresses: dgonzalo@irblleida.cat (D. de Gonzalo-Calvo), georgekararigas@gmail.com (G. Kararigas).

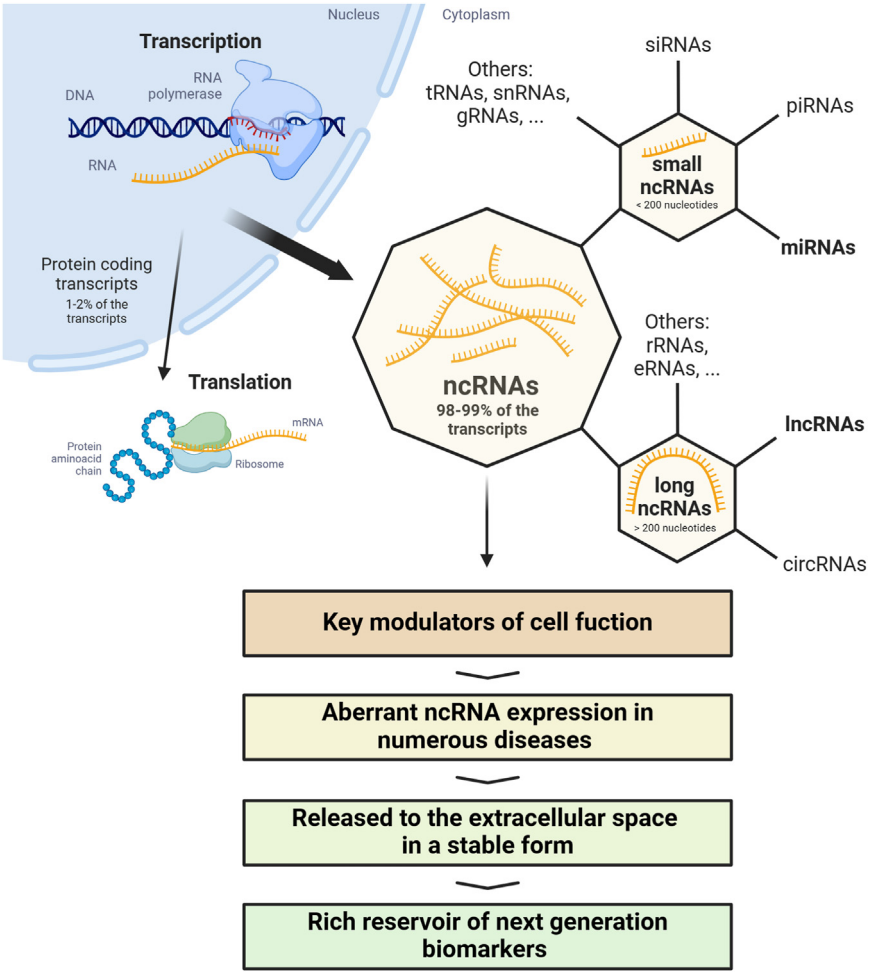

Fig. 1: Noncoding transcriptome and key potential points for molecular phenotyping and biomarker development. Abbreviations: circRNAs, circular RNAs; eRNAs, enhancer RNAs; gRNAs, guide RNAs; lncRNAs, long noncoding RNAs; miRNAs, microRNAs; piRNAs, piwi-interacting RNAs; rRNAs, ribosomal RNAs; siRNAs, small interfering RNAs; snRNA, small nuclear RNA; tRNAs, transfer RNAs.

their clinical relevance and exploiting their potential as therapeutic targets.[5] Advancements in functional genomics have markedly enhanced our comprehension of ncRNA biology,[3] thereby expanding our understanding of the roles of these transcripts in the pathogenesis and progression of human diseases.[6] Alterations in ncRNA expression profiles directly affect the mechanisms they regulate, suggesting that ncRNAs can be causative elements in disease development. The first evidence describing the role of ncRNAs in this context appeared at the beginning of the 21st century.[7] Further compelling evidence has shown that dysregulated ncRNA expression profiles are intrinsic to diseases, including cancer, cardiovascular conditions, metabolic diseases and neurological disorders.[8–10] This knowledge has provided the basis for developing therapeutic approaches.[11] Results from recent clinical trials (https://www.clinicaltrials.gov/study/NCT04045405) suggest the potential translation of ncRNA-based therapeutics, particularly specific antisense oligonucleotides.[12]

ncRNAs represent a rich resource of next-generation biomarkers. In contrast to the DNA sequence, which is constant in an individual, ncRNA expression is highly dynamic and rapidly altered in response to a number of factors, including stressors. The ncRNA profile can be informative of the molecular configuration of the patient, providing an innovative approach for the development of mechanism-based clinical biomarkers as demonstrated by different studies.[13,14] Importantly, ncRNAs are not confined within cells; they can be released into the bloodstream and other bodily fluids in a stable form.[15] Extracellular ncRNAs hold immense translational potential, as they can be obtained through noninvasive biopsies and quantified cost-effectively using techniques readily available in clinical laboratories.[16] The utility of

circulating ncRNAs as biomarkers has been demonstrated in various conditions.[17,18]

It is worth noting that a commercial ncRNA-based test is available to aid in clinical decision-making and guide healthcare. In 2012, the FDA approved the prostate cancer antigen 3 assay (the Progensa® PCA3 assay) to aid in determining the need for repeated prostate biopsies in men who have had a previous negative biopsy. The assay consists of an *in vitro* nucleic amplification test, which measures the RNA concentration of prostate-specific antigen (PSA) and a lncRNA, prostate cancer antigen 3 (PCA3). The market for miRNA-based laboratory development tests (LDTs) has been active in the last decade. Different companies have developed different *in vitro* diagnostic tests, e.g. ThyraMIR®, RosettaGX Reveal™, miRView™, NIS4®, HepatomiR®, OsteomiR® or ThrombomiR®, among others. Furthermore, numerous clinical trials (approximately 300 active trials, https://clinicaltrials.gov/, accessed November 2023) are currently underway, promising to provide valuable insights into the clinical utility of ncRNA-based biomarkers in the short and medium term.

The aim of this article is to review the potential and challenges of using artificial intelligence (AI)/machine learning (ML)-based approaches to identify clinically applicable ncRNA biomarkers for personalised healthcare and the elucidation of ncRNA-related molecular mechanisms underlying disease pathogenesis. In this regard, we discuss the methodological and conceptual limitations as well as potential future steps. As applications based on AI may carry several risks, ethical aspects are also discussed.

Machine learning to address the biological complexity of the noncoding transcriptome

Despite substantial investments aimed at developing ncRNA-based biomarkers, the translation of these findings into clinical practice is still limited.[19] The reasons for this gap between preclinical research and clinical adoption are multiple. In addition to technical limitations, one of the important barriers lies in data analysis.[20] A notable portion of studies focused on ncRNA-based biomarkers have traditionally employed approaches that overlook valuable information. These methods often rely on simplistic criteria, such as the fold change between two sample groups or univariate and linear associations between ncRNAs and clinical outcomes.

Routine clinical practice provides a large amount of information on patient data, including baseline characteristics, disease-related information and pharmacological variables. An effective analysis should encompass the complex interplay between ncRNAs and patient outcomes, considering demographic factors, clinical data and pharmacological variables. Furthermore, it is crucial to account for biological sex, as mounting evidence highlights its impact on disease development, progression and response to treatment.[21–29] Ultimately, the integration of ncRNA data into electronic health records and the existing biomarker landscape promises to yield invaluable insights for their practical implementation in clinical settings. In line with this strategic approach, previous research underscores the potential of ncRNA signatures, such as miRNA profiles, to serve as adjuncts to current benchmark biomarkers, facilitating the transition from laboratory research to clinical application.[30,31]

The inherent biological complexity of the ncRNA family requires particular attention. Coordinated alterations in multiple ncRNAs play a pivotal role in orchestrating a diverse array of biological processes.[32] Individual ncRNAs exert only modest effects on the multifaceted and intricate biological mechanisms underlying diseases. Rather than understanding the implications of isolated biological alterations in disease, it becomes imperative to decipher the interconnections between these alterations and other layers of information, e.g. genomics, proteomics and metabolomics, to gain insight into their impact on pathology. Therefore, the ideal framework to study ncRNAs is based on the concept of "several mediators-one disease" as opposed to the traditional "one mediator disease" notion. Nevertheless, given the vast number of ncRNAs involved in regulating biological processes, identifying those specifically relevant to a particular disease remains a formidable challenge.

The integration of innovative strategies is paramount to make substantial progress in incorporating ncRNA-based biomarkers into clinical practice.[33] ML methodologies learn from vast datasets to discern patterns for diagnostic, prognostic or predictive purposes. Through ML analyses, intricate relationships between molecular components and various variables, including sociodemographic and clinical factors, could be revealed. In contrast to traditional statistical methods that consider limited interactions, ML methods identify complex interactions between features and clinical outcomes. ML algorithms are also well-suited for dissecting complex conditions and biological mechanisms, providing valuable insights into the biological programmes underlying diseases. As such, ML-based approaches have emerged as an indispensable tool for addressing the intricate nature of ncRNAs. ML approaches may capture the multifaceted interplay inherent in ncRNA networks facilitating the identification of relationships between ncRNAs and their interactions with other molecular components within the cellular environment. Given these advantages, ML methods represent the logical progression in the field of ncRNAs. In the following section, we discuss in further detail the potential of ML-based applications in ncRNA research and biomarker discovery.

Application of machine learning in noncoding transcriptome research

Techniques, such as qPCR, microarray and next-generation sequencing, have generated an abundance of transcriptomic data, facilitating the search for meaningful features within extensive datasets.[34] In recent years, there has been a rapid proliferation of ML methods in the analysis of high-dimensional ncRNA datasets. Table 1 summarises the most common supervised ML methods in biomarker discovery using ncRNAs, providing an overview of advantages, disadvantages and performance of each approach. Given that different ML models have distinct strengths and weaknesses, choosing the appropriate algorithm typically involves trade-offs. This decision is not solely data-dependent but also relies on factors, such as the research question, available computational resources and the need for data interpretability. Therefore, it is essential to evaluate multiple ML methods and compare their performance before deciding which one to use for the specific research question (as discussed below). In the following paragraphs, we include examples that suggest the prominent role of the application of ML in the analysis of ncRNAs (Table 2).

Numerous studies have suggested that the integration of supervised ML techniques represents a valuable resource for the identification of potential ncRNA-based biomarkers, the development of classifiers for data-informed clinical decision-making and the elucidation of the molecular underpinnings of diseases. One remarkable example is provided by Errington et al.[35] The authors used four separate ML methods (LASSO, random forest, regression partition tree and XGBoost) to identify miRNA biomarkers associated with pulmonary arterial hypertension (PAH). The consensus ML approach defined miR-187–5p and miR-636 as putative biological markers to predict PAH. The performance of each method was variable [area under the ROC curve (AUC) from 0.72 to 0.84]. Of note, the addition of the random forest model to the PAH-established biomarker NT-proBNP provided a classification model with higher accuracy compared to the single NT-proBNP model (AUC = 0.84 vs. AUC = 0.97 for the random forest model). The miRNA expression profiles also provided novel insights into disease pathogenesis. The integration with their associated target genes using public human lung and whole blood RNA datasets revealed novel mediators of PAH and drug targets.

More recently, an ML-based integrative procedure using 101 algorithm combinations in 2509 colorectal cancer patients from 17 independent public datasets and a clinical in-house cohort was implemented to construct an immune-related lncRNA signature (IRLS).[36] IRLS showed consistent and successful performance in multiple cohorts (5-year AUC up to 0.79) and was proven to be an independent risk factor for overall survival. The signature had superior accuracy to other clinical characteristics and molecular alterations used to assess the prognosis of colorectal cancer in clinical practice. The combination of IRLS with a conventional tool to evaluate the risk and treatment demand, the American Joint Committee on Cancer (AJCC) classification, was significantly better than that of each score alone in multiple datasets. Interestingly, responders to fluorouracil-based adjuvant chemotherapy presented a significantly higher IRLS score than nonresponders (AUC up to 0.84). IRLS also defined a low-risk group of patients, which benefited more from bevacizumab and tended to be resistant to fluorouracil-based adjuvant chemotherapy (AUC up to 0.78). Therefore, the signature might be useful to optimise personalised treatment since it allows to predict the response to therapy. The same group identified a consensus ML-derived lncRNA signature (CMDLncS) for predicting recurrence risk in stage II/III colorectal cancer.[37] To do that, ten ML algorithms were used in 76 combinations and tested in 1640 stage II/III colorectal cancer patients from 15 independent datasets and a clinical in-house cohort. The CMDLncS model had a stable and robust performance at different follow-up timepoints in multiple independent cohorts (C-index up to 0.85). The signature showed a better performance than common clinical and molecular features previously associated with the disease. Patients with high CMDLncS also showed a higher sensitivity to fluorouracil-based adjuvant chemotherapy. Similar results were observed when validated using RT–qPCR assays. Overall, lncRNA-based signatures constructed using ML constitute a clinically translatable tool to improve clinical outcomes in patients with colorectal cancer. Such signatures also contribute to the refinement of personalised therapeutic strategies.

The use of single supervised ML methods has also yielded prominent results. For example, the support vector machine (SVM) algorithm was used to construct a diagnostic classifier based on mRNAs, lncRNAs and circular RNAs (circRNAs) detected in extracellular vesicles from human plasma.[38] The study population included 159 healthy individuals, 150 patients with cancer (five cancer types) and 43 patients with other diseases. The classifier showed an optimal sensitivity and specificity (between 96% and 98%) to first distinguish patients with cancer from healthy individuals and then diagnose hepatocellular cancer (between 84% and 94%). The classifier also displayed a higher accuracy than α-fetoprotein (AFP), a traditional biomarker for risk assessment in hepatocellular cancer (AUC = 0.95 vs. AUC = 0.83). In the same way, the SVM algorithm was combined with exosomal miRNA profiling to develop diagnostic models for pulmonary tuberculosis and tuberculous meningitis, a complex clinical condition, in which the diagnosis is challenging when based solely on

Method	Brief method description	Advantages	Disadvantages	Performance and computational efficiency
K-nearest neighbours (K-NN)	To classify samples, K-NN method assigns samples to the class label of the most similar training samples. Similarity is determined by a distance metric, e.g. Euclidean distance (a most common distance metric). The method first computes the distance between a new data sample to all the samples in the training set. The class label of the new data sample is determined by using the class label of *k* nearest neighbours, for example by taking majority vote. For regression tasks, the value of new data sample is computed by using the mean of *k* nearest neighbours' values.	• There is no training required. • Works well with datasets when there is a clear distinction among classes. • Makes no assumptions about data distribution.	• Selection of *k* is required for optimal results. • Sensitive to outliers. • Sensitive to high-dimensional data containing irrelevant features, more than most other ML methods.	• Performance: good when there is a clear distinction among classes. Note the accuracy can vary depending on the selected *k* value, and the dataset (e.g. presence of irrelevant features in the dataset). • Computational efficiency: fast during training, as there is no training required. Training data is only loaded in memory. The method may be slow during testing, especially for large datasets, and/or high number of features, or large values of k.
Naïve Bayes	Naïve Bayes method is based on Bayes' Theorem which specifies how dependent events are related. It determines conditional probability (probability of an event given the occurrence of another event). Naïve Bayes method on the other hand, considers the naïve assumption of conditional feature independence. That is, all features are equally important and the value of one feature is independent of the value of any other feature for a given class variable.	• In general, the method performs better than other methods with small data sets if the assumption of conditional feature independence holds. • Very fast to execute as probabilities can be directly computed.	• Assumption of conditional independence of features, which does not always hold. However, despite this, the method often achieves good performance.	• Performance: high if assumption of conditional independence of features holds and lower if this assumption is violated. • Computational efficiency: Fast.
Decision Trees (CART, C4.5, ID3, CHAID)	Decision trees are rule-based methods that apply divide and conquer approach to create models with logical decisions organized in a tree-like structure. The model divides/splits the data into increasingly smaller groups of related classes using the decisions based on the data's features/predictors. Various decision tree algorithms exist, e.g. CART (classification and regression trees), C4.5, ID3, CHAID (chi-square automatic interaction detection). Mostly vary depending on the method used to find the most important feature to choose from in order to split the data in the dataset. To find the most important feature CHAID uses chi-square tests, ID3 uses information gain, C4.5 uses gain ration and CART uses GINI index.	• Can effectively handle data nonlinearity. • Can handle missing data. • Small trees yield interpretable model, making decision trees suitable for tasks when interpretability is crucial. • Can be used on data with both relatively small and large sample size.	• Have high variance, i.e. small changes in data can result in large change in the model's predictions. • Prone to overfitting (i.e. high performance on the training data, but poor performance on test/unseen data), especially when deep trees are built. They are also prone to underfitting (i.e. poor performance both on both training and test/unseen data), especially with small trees. • Due to above disadvantages decision trees are more suitable during exploratory analysis of ncRNA research. • Large trees may be difficult to interpret.	• Performance: even though performance for training data can be high, decision trees can have lower performance in test data as they are prone to overfitting. In this case, pruning of decision trees is recommended. • Computational efficiency: in general, fast. However, large trees take longer to train and make predictions.
Random Forest (RF)	An ensemble-based method that performs predictions using a group of decision trees. For each decision tree, RF applies bootstrap resampling on the input dataset. In addition, each decision tree is trained on a random feature subset. For classification, the method uses majority voting to combine decision trees' predictions. For regression, the final decision is made by averaging decision trees' predictions.	• Well-known for its high accuracy. • Can effectively handle data nonlinearity. • Can handle noisy or missing data. • When compared to single decision tree, RF reduces overfitting due to ensemble learning, bootstrap resampling and the selection of random feature subsets within each decision tree. • Can handle large dataset with high dimensionality.	• Challenging to interpret, especially for larger complex models.	• Performance: In general, high, however hyperparameter tuning is often needed to achieve optimum results (e.g. selection of number of trees, and number of features to use during splitting of data). • Computational efficiency: depends on the number and the depth of the trees. Thus, for large number of trees that are deep, training of RF can be computationally intensive.

(Table 1 continues on next page)

Method	Brief method description	Advantages	Disadvantages	Performance and computational efficiency
(Continued from previous page)				
Extreme Gradient Boosting (XGBoost)	Gradient Boosting is an ensemble method that combines several models, usually decision trees, to make a final decision. In order to minimize the loss function, the method applies gradient descent. XGBoost is a grading boosting that computes second-order gradients of the loss function to minimize loss. The method also uses regularization to penalize complex models, using L1 an L2 regularization, which helps to reduce over fitting, a common problem in machine learning. See LASSO and Ridge regression method descriptions for more detail on L1 and L2.	• Well-known for its high accuracy. • Reduces overfitting by including regularization. • Can handle missing data effectively.	• Difficult to interpret due to the complexity of the model. • Parameter tuning may be needed to tune the model in order to achieve optimum results.	• Performance: in general, high, if tuning of hyperparameters is performed. • Computational efficiency: fast and efficient.
Artificial Neural Networks (ANN)	ANN methods aim to resemble the behaviour of human neural architecture. ANN is made up of interconnected layers of neurons consisting of input layer (independent variables), one or more hidden layers and an output layer (dependent variable). Each neuron receives weighted outputs from neurons in a previous layer, which are then added together and transformed nonlinearly at neuron's output. The main aim is to iteratively modify the weights in order to minimize the prediction error.	• Well-known for its high accuracy. • Able to capture intricate non-linear relationships in the data.	• Prone to overfitting and underfitting. • Parameter tuning necessary to find the optimal model, especially for complex tasks. • In general, requires large amounts of data to train the model. • Due to its black box nature ANNs are difficult to interpret.	• Performance: high with sufficient data and with well-tuned hyperparameters. • Computational efficiency: in general, slow during the training phase, especially for complex network architectures and faster during testing.
Support Vector Machine (SVMs)	Support vector machines use a kernel to translate data input into a multidimensional space. To separate the data, the algorithm creates a hyperplane by maximizing the margin and minimizing the classification error. SVM kernels that are frequently employed include sigmoid, radial basis function, polynomial, linear, and nonlinear kernels.	• Well-known for high accuracy. • Can effectively handle data nonlinearity. Compared to ANNs, SVMs are more suitable when dealing with small to medium-sized datasets. Furthermore, SVMs are effective when the number of features is large relative to the number of samples (as is usually the case for ncRNA data). • Are not very sensitive to noisy data. • In general, SVMs are resistant to overfitting. • Effective in minimizing outliers.	• Testing several kernels and model parameter combinations is necessary to find the optimal model. • Difficult to interpret.	• Performance: High, especially with well selected kernel and optimized hyperparameters. • Computational efficiency: slow during training, especially when: non-linear kernels are used, for large datasets and high-dimensional data (such as ncRNA data). SVMs are in general fast during testing.
Least absolute shrinkage and selection operator (LASSO) and Ridge regression	A regression method that uses L1 regularization, i.e. it adds a penalty that equals to the absolute value of the magnitude of the coefficients. As a result, some coefficients may become zero and get eliminated from the model, reducing model complexity. The method thus automatically performs feature selection by eliminating less important features. Ridge regression uses L2 regularization, i.e. the penalty is the sum of the squares of the magnitude of the coefficients. Similar to LASSO, Ridge regression can also shrink some coefficients, but never sets them to zero. Unlike LASSO, Ridge regression does not perform variable selection.	• Well-suited for problems that have large number of features. • Reduces model complexity and over-fitting. • LASSO: automatically performs feature selection. • Ridge regression: effectively handles multicollinearity.	• LASSO: If features are correlated, the method selects one feature arbitrarily, which may result in removing features that may be as (or more) important than the selected feature. The method should thus be avoided if the data has many correlated features.	• Performance: generally good when hyperparameters of L1 (for LASSO) and L2 (for ridge regression) regularization are fine-tuned. In the presence of multicollinearity, LASSO may result in reduced performance. • Computational efficiency: LASSO can be more computationally intensive than Ridge regression due to the feature selection.

(Table 1 continues on next page)

Method	Brief method description	Advantages	Disadvantages	Performance and computational efficiency
(Continued from previous page)				
Elastic network (Enet) regression	A linear regression algorithm that uses regularization to penalize complex models. The method combines the penalties of Lasso and Ridge regression i.e. it applies both L1 and L2 penalties to the standard least-squares objective function. This helps to reduce overfitting and handles multicollinearity. The method can also be applied for classification problems.	• Combines the advantages of feature selection from LASSO and effectively handles multicollinearity from Ridge regression. • Has shown to outperform other linear regression methods. • Reduces overfitting.	• Not suitable for datasets with very large number of features compared to the number of samples.	• Performance: generally high when hyperparameter tuning is performed. • Computational efficiency: more computationally expensive when compared to Lasso and Ridge regression.

Table 1: **Most prevalent supervised machine learning methods in biomarker discovery utilizing noncoding RNAs, based on current publications.**

the electronic health record due to, for example, overlapping symptoms and radiological features, technical limitations, cost, accessibility to tests, among other factors.[39] The prospective multistage study including 370 individuals facilitated the development of an ML framework combining patient data from electronic health records and exosomal miRNAs, with superior performance in differentiating pulmonary tuberculosis and tuberculous meningitis from highly suspected cases (sensitivity and specificity over 89%). Comparable approaches obtained similar results in other conditions. Neural networks based on miRNA features were reported to improve the diagnosis of acute coronary syndrome.[40] Herein, 34 previously published miRNAs associated with myocardial infarction were validated for their predictive value as early diagnostic markers of acute coronary syndrome. A neural network model including ten miRNAs provided the highest accuracy (0.96) in the diagnosis of the conditions while maintaining the same specificity as the clinical gold standard troponin T (0.96). Devaux et al.[41] have developed a ML model designed to predict in-hospital mortality following SARS-CoV-2 infection. The study analysed a panel of 2906 cardiac-enriched lncRNAs, previously established,[42] alongside clinical data from 1286 COVID-19 patients across four distinct cohorts (PrediCOVID from Luxembourg, NAPKON from Germany, ISARIC4C from United Kingdom and BQC19 from Canada). The three European cohorts totalling 804 patients were merged and used as a discovery cohort for feature selection and model optimization. The fourth cohort, BQC19, consisting of 482 patients, was used for validation purposes. The study identified age and LEF1-AS1, a lncRNA, as predictive features, achieving an AUC of 0.83 through utilization of a multilayer feed-forward neural network classifier. Random forest has also demonstrated its utility in identifying circulating miRNA profiles associated with pulmonary function and radiologic features in survivors of SARS-CoV-2-induced ARDS. This approach has provided novel insights into the potential molecular pathways underlying the pathogenesis of pulmonary sequelae.[43]

According to these findings, ML algorithms have the capability to aggregate and assess information from various dimensions, resulting in models that incorporate ncRNAs with diagnostic and/or predictive potential. These approaches hold particular importance in heterogeneous populations and multifactorial diseases, where the delineation of patient subgroups using novel predictors, in conjunction with traditional clinical variables, offers considerable value. Ultimately, this approach has the potential to stratify relatively homogeneous patient groups, thus facilitating more precise medical management and substantially influencing healthcare procedures for personalised patient benefit.[44] In a recent study by Katipally et al.,[45] molecular subtypes were identified among patients who underwent hepatic resection for limited colorectal liver metastases. Using a 31-feature set comprising 24 mRNAs and 7 miRNAs, the researchers developed a neural network classifier to predict molecular subtypes in the discovery cohort, which was then applied to the validation cohort. The study revealed three distinct molecular subtypes: immune, canonical and stromal, each associated with varying rates of 5-year progression-free survival and 5-year overall survival. Notably, integrating molecular subtypes into a clinical risk score led to improved predictive accuracy. Two circulating miRNAs, let-7g–5p and miR-143–3p, have been proposed as novel tools to define subpopulations of patients with suspected stable coronary artery disease referred for coronary computed tomography angiography.[46] A panel of ten miRNAs previously associated with the disease showed low discriminating value in the whole population (AUC = 0.54–0.64) and there was no incremental benefit when combining it with a clinical model based on traditional cardiovascular risk factors with respect to the discriminatory capacity for coronary atherosclerosis burden using a classical statistical analysis. However, both miRNAs facilitated the classification of patients into distinct subpopulations with specific clinical profiles based on the presence, extent and severity of coronary atherosclerosis. This classification was achieved using a classification tree algorithm, specifically the chi-

	Machine learning (ML) approach	Noncoding RNA family	Disease/ Condition	Sample matrix	Study question	Population	Main findings	Reference
ncRNA-based biomarker discovery	Supervised. LASSO, RF, Regression Partition Tree and XGBoost.	miRNAs	Pulmonary arterial hypertension (PAH)	Plasma	To identify patients at risk of PAH earlier and provide new insights into disease pathogenesis	64 treatment naïve patients with PAH and 43 disease and healthy controls	miR-187–5p and miR-636 as potential biological markers to predict PAH, and reveal novel disease mechanisms and highlight future putative drug targets	(Errington et al., EBioMedicine, 2021)
	Supervised. RSF, Enet, Lasso, Ridge, stepwise Cox, CoxBoost, plsRcox, SuperPC, GBM, and survival-SVM.	lncRNAs	Colorectal cancer (CRC)	Multiple sample matrices	To apply immune-related lncRNA signature (IRLS) to develop and validate a risk stratification signature in CRC	2509 CRC patients from 17 independent public datasets and a clinical in-house cohort	IRLS as a powerful signature for assessing the prognosis, recurrence, and benefits of fluorouracil-based ACT, bevacizumab, and pembrolizumab treatments in CRC	(Z. Liu et al., Nat Commun, 2022)
				Multiple sample matrices	To explore the clinical significance of lncRNAs in stage II/III CRC and systematically identify a consensus machine learning-derived lncRNA signature (CMDLncS)	1640 stage II/III CRC patients were enrolled from 15 independent datasets and a clinical in-house cohort	CMDLncS for identifying patient at high risk of recurrence that could optimize precision treatment to improve the clinical outcomes in stage II/III CRC	(Z. Liu et al., EBioMedicine, 2022)
	Supervised. SVM	mRNAs, lncRNAs and circRNAs	Cancer (five cancer types: hepatocellular, gastric, colorectal, breast, kidney cancer)	Plasma	To investigate the potential of extracellular vesicle long RNA for cancer diagnosis	159 healthy individuals, 150 patients with cancer and 43 patients with other disease	exLR as specific markers potentially useful for cancer diagnosis.	(Y. Li et al., Clin Chem, 2019)
	Supervised. SVM	miRNAs	Background Tuberculosis (TB): Pulmonary tuberculosis (PTB) and tuberculous meningitis (TBM)	Plasma	To investigate the potential of exosomal miRNAs and electronic health records in TB diagnosis	370 individuals, including PTB, TBM, non-TB disease controls and healthy controls	Patients' data from electronic health records combined with exosomal miRNAs (miR-20a, miR-20b, miR-26a, miR-106a, miR-191, miR-486) achieved superior performance in differentiating PTB and TBM from those highly suspected cases	(Hu et al., EBioMedicine, 2019)
	Supervised. NN	miRNAs	Acute coronary syndrome (ACS)	Whole blood and serum	To determine the diagnostic value of miRNA profiling in ACS	66 patients with ACS and 68 controls	NN model including ten miRNAs provided the highest accuracy (0.96) in diagnosis of the conditions while maintaining the same specificity as the clinical gold standard troponin T (0.96)	(Kayvanpour et al., J Mol Cell Cardiol, 2021)
	Unsupervised. K-means clustering	miRNAs	Subclinical lung injury	Plasma	To evaluate associations of plasma EV-miRNAs with lung function	656 participants	Specific miRNA expression profile identified a cluster of patients with an increased risk of declining lung function over time	(Eckhardt et al., Am J Respir Crit Care Med, 2023)

(Table 2 continues on next page)

square automatic interaction detector (CHAID). Strikingly, circulating miRNAs added higher discriminative value to the trees compared with other biomarkers, e.g. hs-cTnT or hs-CRP. In the same manner, the classification and regression tree (CART) algorithm was used to construct regression tree models in 810 patients with end-stage renal disease on hemodialysis.[47] Again, two miRNAs, miR-186–5p and miR-632, complemented risk factors to identify patient subpopulations with specific cardiovascular risk patterns, particularly a subgroup with high risk, which may benefit most from intensive monitoring. While their discriminative value was diluted when analysing the whole population, the regression tree selected miRNAs as biomarkers particularly relevant for four subpopulations of patients. The inclusion of both miRNAs allowed for better discrimination during the first two years of the follow-up (integrated AUC = 0.71) compared with regression tree models without miRNAs (integrated AUC = 0.68). In the same line, the combined use of circulating miR-133a–3p

	Machine learning (ML) approach	Noncoding RNA family	Disease/ Condition	Sample matrix	Study question	Population	Main findings	Reference
(Continued from previous page)								
ncRNA-based classification of disease subtypes	Supervised. NN	miRNAs	Oligometastatic colorectal liver metastases	Formalin-fixed paraffin-embedded specimens	To independently validate previously defined molecular subtypes in the phase 3 New EPOC randomized clinical trial	240 patients who underwent hepatic resection for limited colorectal liver metastases	Molecular subtypes of oligometastatic colorectal liver metastases and integrated risk stratification are prognostic and warrant further study as a possible predictive biomarker to personalize therapies	(Katipally JAMA Oncol, 2023)
	Supervised. CHAID	miRNAs	Coronary artery disease (CAD)	Plasma	To explore the diagnostic performance of circulating miRNAs as biomarkers in patients with suspected stable CAD	200 patients with suspected stable CAD	Circulating miRNAs emerge as an interesting tool to classify subpopulations of patients with suspected stable CAD according to the presence, extension and severity of coronary atherosclerosis	(de Gonzalo-Calvo et al., J Intern Med, 2019)
	Supervised. CART	miRNAs	Patients on haemodialysis (HD)	Plasma	To test whether miRNAs, and nonstandard predictive models, such as decision tree learning, provide useful information for medical decision-making in patients on HD	810 patients with end-stage renal disease who had been treated with regular HD	miR-186–5p and miR-632, complemented risk factors to identify patient subpopulations with a higher cardiovascular risk	(de Gonzalo-Calvo et al., Theranostics, 2020)
	Supervised. Decision Trees, Naïve Bayes, K-nearest neighbors, LogitBoost, Logistic Model Tree, Simple Logistic, RF and Sequential Minimal Optimization	miRNAs	Endocrine hypertension (EHT)	Plasma	To train ML algorithms for diagnosing endocrine hypertension subtypes using multi-omics (MOmics) data. It also aims to provide an understanding of discriminating features and their importance to different disease combinations	354 hypertensive subjects and 133 normotensive volunteers	Potential of ML-based approaches in the combination of Momics data, including ncRNAs, to construct innovative tools with a high impact on patient management	(Reel et al., EbioMedicine, 2022)
	Supervised. Stepwise Cox, CoxBoost, ridge regression, RSF, GBM, Survival-SVM, LASSO, SuperPC, plsRcox, and Enet	lncRNAs and miRNAs	Muscle-invasive urothelial cancer (MUC)	Multiple Bladder Cancer (BLCA) and MUC tissue samples	To combine mRNA, lncRNA, miRNA expression profiles, genomic mutations, and epigenomic DNA methylation data to develop an integrated consensus subtype of MUC	BLCA cohort and cohorts from external datasets	Comprehensive analysis of multiomic data can offer important insights and further refine the molecular classification of MUC. Identification of robust consensus machine learning-driven signature represents a valuable tool for early prediction of patient prognosis and for screening potential candidates likely to benefit from immunotherapy.	(Chu et al. Mol Ther Nucleic Acids, 2023)

Abbreviations: CART, Classification and Regression Tree; CHAID, Chi-square Automatic Interaction Detector; circRNA, circular RNA; Enet, Elastic network; GBM, Generalized boosted regression modelling; lncRNA, long noncoding RNA; mRNA, messenger RNA; miRNA, microRNA; NN, Neural network; plsRcox, Partial least squares regression for Cox; RF, Random Forest; RSF; Random survival forest; SuperPC, Supervised principal components; SVM, Support vector machine.

Table 2: **State of the art of machine learning methods in noncoding RNA transcriptome research.**

and clinical data has emerged as a promising biomarker for delineating a low-risk subphenotype in patients suffering from heart failure and central sleep apnea.[48] The utilisation of ncRNAs in decision tree models may not be universally applicable to all transcripts since the potential to define subpopulations of other families, such as circRNAs, is still not fully understood.[49]

Along this line, the combination of multidimensional omic analysis and ML to classify subtypes of disease has provided promising results. This approach was used in the context of arterial hypertension, a major cardiovascular factor with high heterogeneity, in which the definition of disease subtypes is fundamental to avoid underdiagnosis.[50] In a large cohort of hypertensive patients (n = 487) from 11 reference centres, 409 features (plasma small metabolites, plasma miRNAs, urinary steroid metabolites, plasma steroids and plasma catechol O-methylated metabolites) were analysed. Using eight classifiers (decision trees, naïve bayes, K-nearest neighbours, logitBoost, logistic model tee,

simple logistic, random forest and sequential minimal optimisation), an ML pipeline was developed based on multiomic features that allowed the classification of five disease combinations (AUC = 0.95, Specificity = 96%). miR-15a–5p and two plasma small metabolites were present in all disease combinations. Although the findings should be further validated in external cohorts, this work demonstrates the potential of ML-based approaches in the combination of omic data, including ncRNAs, to construct innovative tools with a high impact on patient management. Indeed, Chu et al.[51] employed an integrative approach, combining mRNA, lncRNA and miRNA expression profiles, along with genomic mutations and epigenomic DNA methylation data, to develop a consensus subtype of muscle-invasive urothelial cancer. The authors explored ten different multiomic integration strategies to achieve this. Subsequently, they identified stable prognosis-related genes by analysing differential expression across subtypes and utilised ten ML algorithms to construct a robust consensus signature. This signature not only demonstrated significant prognostic value but also exhibited strong performance in predicting responses to both immunotherapy and drug-based therapies in the training and validation cohorts.

Unsupervised ML approaches have also been demonstrated to be useful in the development of ncRNA-based biomarkers. K-means clustering was used to find previously unknown patterns of plasma extracellular vesicle-enriched miRNAs associated with subclinical lung injury in a large longitudinal cohort study [The U.S. Veterans Affairs Normative Aging Study (NAS)].[52] A cluster of participants with a characteristic miRNA expression profile that showed an increased risk of declining lung function over time [relative risk (RR) 1.19, 95% CI 1.05–1.35] was defined. The 11 miRNAs differentially expressed in the group at higher risk were related to specific biological pathways implicated in cellular immunity, inflammatory response and airway structural integrity that could emerge as potentially treatable targets. The results not only have a direct impact on biomarker development but also on other aspects of disease.

Challenges and perspectives in the application of AI/ML-based methods to implement noncoding RNA in research and clinical practice

The use of AI/ML-based methods in the field of ncRNA research faces several obstacles. Many of these obstacles are similar to those experienced in other areas of biomedical research.

The common experimental design is based on "many ncRNAs in few patient samples" resulting in relatively small datasets with high-dimensional feature spaces, a phenomenon known as "curse of dimensionality." This scenario leads to a significant risk of overfitting, i.e. the ncRNA-based ML models might fit too closely to the training data, limiting the generalisation to broader patient populations. The validation through multicentric studies with adequately sized samples is therefore fundamental to evaluate model performance and, by extension, its clinical applicability. In addition, the effective management of dimensionality is crucial. Dimensionality reduction techniques such as principal component analysis (PCA), mitigate collinearity issues and generate a concise feature set that effectively preserves a substantial portion of variability within the data. Other strategies for reducing the number of ncRNA candidates can be found in the literature. A clustering-based method was proposed to reduce the number of candidate miRNA combinations, thereby avoiding the exponential number of combinations.[53] The combinations of representative cluster members could then be entered into ML-based analyses and provide miRNA-based biomarkers with high accuracy. The selection of discriminative features using random forest or analogous methods, in conjunction with a preliminary filtration step to eliminate redundant information by removing highly correlated ncRNAs, have also been employed.[51,54]

A complex ncRNA-based classifier with an excess of features may have limited utility, as the findings are likely to be challenging to interpret from a biological perspective. Indeed, the interpretability of the ML models is often complicated due to their complex architectures, making it difficult to understand how predictions are made. Feature importance analysis and techniques such as SHAP (SHapley Additive exPlanations) increase the interpretability. Additionally, incorporating biological expertise into the model development process can help ensure biologically meaningful decisions. Nevertheless, most ncRNAs have poorly understood functions, making it challenging to interpret their biological involvement and, consequently, their associations with disease. The absence of functional annotation may impede the biological insights derived from ncRNA expression data.

ML methods require, as do most analyses, reliable and trustworthy data. ncRNA datasets often exhibits a significant amount of missing data due to technical errors. The handling of undetectable values because ncRNAs may be simply nonexpressed also deserves special consideration.[55] Such incomplete data can present difficulties for ML methods, potentially leading to unreliable outcomes. Addressing this limitation requires careful data preprocessing, including for example imputation methods.[56,57] In addition, imbalanced class distributions and biases in the ncRNA data can impact ML model performance, particularly in the context of underrepresented ncRNA candidates.

Another limitation is the paucity of studies that have compared different ML methods and their effectiveness in analysing ncRNA data,[58] as ML-based algorithms do

not perform consistently across all datasets. Therefore, the most appropriate method to use for analysis is often determined empirically.[59] Conducting comparative studies to evaluate various ML algorithms could aid in identifying the most suitable approach for specific ncRNA datasets and research objectives. The combination of multiple algorithms to address limitations arising from algorithm selection has also been documented in the literature.[51,60] Ensemble methods, which use multiple algorithms to construct a model, represents an indispensable tool.[61]

The integration of ncRNA and clinical features using ML methods in decision support system may improve patient management, similar to other biomarkers.[62] Nevertheless, a critical evaluation of ncRNA-based model performance against guideline-recommended pathways and established clinical algorithms is crucial to assess their utility in clinical practice. Rigorous validation of ML-based profiles or algorithms should precede their application to patient groups for which they were not originally designed.

The compartmentalisation of ncRNAs in biofluids, as well as the characterisation of ncRNA profiles within different cell populations and across different subcellular compartments, also warrants investigation. In this regard, integration of single-cell RNA sequencing (scRNA-seq) technology with ML holds immense potential for ncRNA research, as scRNA-seq enables the profiling of ncRNA expression at the single-cell level, providing insights into cellular heterogeneity. ML algorithms designed for scRNA-seq data analysis can assist in identifying ncRNA biomarkers associated with specific cell populations or disease states.

The proliferation of high-throughput technologies has facilitated the generation of omic data with different but complementary information, i.e. genomics, transcriptomics, proteomics or metabolomics, among others. Multiomic data integration, combining ncRNA expression data with other omic data could provide a more comprehensive understanding of the biological processes driving disease pathogenesis. In this context, the use of ML models has emerged as a useful approach to integrate information from different omic layers.[63] Additionally, many ML-based methods have been developed for predicting ncRNA-gen/protein interactions.[64] Integrating ML approaches with underlying biological networks, such as gene regulatory networks and protein–protein interaction networks, represents an asset for elucidating the molecular mechanisms of ncRNAs in biological processes and disease pathology. This is particularly relevant since the combination of experimentally confirmed ncRNA with protein–protein interactions enhance the capacity to identify disease modules and predict comorbidity patterns between diseases and could ultimately facilitate the identification of novel drug–targets and a better understanding of disease progression.[65]

Emerging methods such as deep learning and reinforcement learning offer exciting opportunities to address current challenges in ncRNA research. Deep learning models, with their capacity to discern informative patterns from large datasets, can identify nonlinear relationships between ncRNAs and biological phenotypes. Various deep learning models that have successfully been used in ncRNA research with contributions in the areas of ncRNA identification, prediction of interactions and biological mechanisms and disease classifications, among others.[66] Reinforcement learning, although less explored, could be used for the prediction of potential ncRNA-based biomarkers and ncRNA-disease association prediction.[67,68] A more comprehensive examination of these techniques will provide valuable insights in the short to medium term.

ML techniques are progressively being employed for the study of ncRNA biology, especially in the case of the prediction of miRNA targets.[69] However, the performance of ML heavily relies on user-defined variables chosen for model training. Considerations for classifier-based ncRNA target prediction methods include the challenges of class imbalance and dataset reliability mentioned above. The landscape of mRNA regulation is characterised by the intricate interplay between miRNAs and their target mRNAs, where each mRNA can be modulated by multiple miRNAs, and conversely, each miRNA may have thousands of potential binding sites across the transcriptome. The skewed distribution between these classes impacts the performance of prediction algorithms, impeding their ability to accurately discern true miRNA-mRNA interactions. The absence of a predicted target may also stem from various biological factors beyond the scope of computational prediction, adding further complexity to the analysis. Addressing these challenges requires a detailed understanding of both the computational methodologies employed and the underlying biological intricacies governing ncRNA regulatory mechanisms. Moreover, evolutionary disparities among biological species, for instance in the case of lncRNAs, complicates data integration and generalisation of findings.

Beyond the specific obstacles of incorporating AI/ML in ncRNA research, the field of ncRNA-based biomarkers and therapeutics must also address more general challenges. These limitations are multifaceted, involving methodological, technical and experimental considerations.[70] The use of artificial case–control designs with selected patients and healthy controls, which, while useful for evaluating molecular pathways and pathological mechanisms, tend to overestimate the value of a given biomarker. To achieve more accurate and clinically relevant results, biomarker analyses should be conducted in real clinical settings, e.g. with patients suspected of having a disease. Additionally, cost-effectiveness is rarely explored, yet this is crucial for evaluating their potential for routine clinical practice.

The use of biospecimens that are easy to collect and commonly used in clinical laboratories, such as urine or saliva, should also be considered, especially for specific populations, e.g. children.

Another significant challenge is the labour-intensive nature of current ncRNA quantification assays, which necessitates simplification, miniaturisation and automation. Pre- and post-analytical variables significantly impact results and are often the primary causes of inconsistencies among findings published by different groups.[19] For instance, pre-analytical variables, such as the impact of circadian rhythm, are frequently overlooked but warrant careful consideration during study design. Adopting best practice guidelines and standardising protocols are urgently needed to enhance the reliability and comparability of measurements.

The decreasing costs of microarrays and sequencing techniques present an opportunity to exploit the transcriptome for biomarker development. However, differences in sensitivity, specificity and biases between platforms, e.g. microarrays, RNA-seq or scRNA-seq could lead to inconsistent findings. Most scRNA-seq approaches employ poly-A dependent methods, which restrict the ability to quantify ncRNA species, such as mature miRNAs or circRNAs. To overcome these challenges, it is necessary to address the variability inherent to different platforms and implement cross-platform validation techniques.

In the past decade, the field of therapeutic tools based on ncRNA has witnessed considerable growth, particularly miRNA-based products, e.g. anti-microRNAs (antimiRs) and miRNA mimics, with the technology now regarded as a promising component in the therapeutic market.[71] Numerous ncRNA-based therapeutics are presently undergoing clinical investigation for various conditions, such as diverse forms of cancer, cardiovascular disease, genetic disorders and viral infections.[72] These therapies offer several advantages, including high specificity, precise targeting of disease-related genes or proteins, cost-effectiveness and a relatively straightforward manufacturing process.[73] Nevertheless, the safety of mimic therapy is still a matter of debate.[74] The successful translation of RNA therapies into widespread clinical use still depends on further interdisciplinary research on delivery, stability and potential off-target effects.[73] The promising advancements arising from preclinical research will ultimately overcome the issues currently faced in the field of ncRNA therapeutics.

The implementation of open science practices becomes essential to ensure the advancement of knowledge and the acquisition of generalizable data. Particularly, sharing data, code and research protocols, permit facilitating transparency and reproducibility in ncRNA-ML research. Given the dynamic and heterogeneous nature of the field, the ability to reproduce results is fundamental for validating research findings. Sharing scientific data additionally enables researchers to combine data types to strengthen analyses, facilitates to reuse data that are difficult to generate or from limited sources. Beyond its role in the generalization of the scientific results, data sharing also fosters collaboration and accelerates scientific progress by enabling researchers to address more complex biological questions.

It is of paramount importance to ensure the use of high-quality data to guarantee both the accuracy and the usefulness of applications and projects that are based on shared data. For this purpose, various publicly available repositories have been specifically designed for the harmonisation of ncRNA research outputs. These include RNAcentral (https://rnacentral.org) and miRBase (https://www.mirbase.org) databases, which collect published ncRNA and miRNA sequences and annotations, respectively. Furthermore, platforms such as RNALocate v2.0 (http://www.rna-society.org/rnalocate/) have been developed to explore the associations between ncRNAs at the subcellular level.[75] General-purpose data sharing platforms have also been launched, e.g. NCBI Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo) or EMBL-EBI (https://www.ebi.ac.uk/), permitting researchers query, download experiments and curated gene expression profiles.[76] Additionally, platforms providing secure spaces for creating and sharing methods and protocols across different scientific areas have been designed to address the reproducibility of the procedures, e.g. protocols.io (https://www.protocols.io).[77] Consistently, software development services have also been created for code sharing with the aim to replicate generated code in other investigations, e.g. GitHub (https://github.com), the world's leading platform for software development promoting developers' collaboration in a secure space.[78] Open-source code platforms provide a venue to support for ML research, namely Kaggle (https://www.kaggle.com), a comprehensive repository of community-published models, data and code, further allows data scientists participate in ML and data challenges.[79] In these trials, created models can be trained on different datasets, what ultimately exposes the model to a broader range of variations, enabling it to learn more comprehensive and discriminative feature representation.

Ethical aspects

The potential of AI/ML-based methods and tools in research and clinical routine is high, and the ability of these to be a positive force in medicine creates optimism among patients and other end-users. However, the complexity and unknowns of such systems may hinder their appropriate use. Additionally, due to the so-called black boxes of AI/ML, their adoption into research and clinical routine may be further impeded, as they are not readily trusted.[80,81] There are also issues of trust among patients for the use of AI/ML-based systems

that need attention before widespread adoption of such tools in clinical decision-making.[82] All these factors lead to the quest for ethical and trustworthy AI, and this has become a central issue for governance and technology impact when implementing AI/ML-based systems.[81]

Ethical concerns include bias in general. For example, an algorithm may benefit some specific groups more than others, or there could be differing performance of an AI/ML-based system for different subpopulations.[83,84] Further bias may arise due to lack of data from underreporting (based on stigma and/or silence), but it can also be due to unreliable or poorly performing biomarkers. Bias in the build and composition of the AI/ML-based models also constitutes an important reason for concern and could be a result of there being a substantial amount of missing data. This is also an issue regarding the use of ncRNA biomarkers with health record data. Algorithms are highly dependent on the objective measures included in them, meaning that unreliable or absent biomarkers and subjective measures will have a very negative impact on AI model performance. In addition, large-scale omic data employment for AI would suffer from both technology-specific biases and batch effects, and differences in data treatment processes or data preprocessing pipelines may confound data analysed by AI, impairing model performance.[85]

Lack of transparency in AI models may also result in inappropriate use, low explainability or low trust.[86] Therefore, a concern for the trustworthiness of AI and ML-based results is that these methods are sensitive to errors in data integrity and health record data. There is also the possibility that AI models may reinforce existing biases in healthcare because of using inherently biased learning data or due to developers unintentionally causing the AI model to incorporate their own, often unconscious, biases.[82] Algorithmic bias is common and is accentuated by health disparities,[87] as exemplified in a recent article that found consistent underreporting of female and black patients regarding the management of acute chest pain in US emergency departments.[88] Such biases and inequalities would also be preserved in potential AI/ML-based models built on these data.

Consequently, the clinical safety of AI/ML-based tools is a major issue, and to ensure their safe use, it has been suggested that independent assessment, regulatory protection and governmental oversight procedures need to be installed to protect against potential harm and ensure validity, precision and accuracy.[82] Accordingly, inherent ethical, technical, domain-specific and legal consequences of the use of specific AI/ML-based models need to be evaluated, considering the different phases of AI, i.e. design, development, deployment and monitoring.[80,85]

There are also concerns about the cost–benefits of adding advanced molecular analysis, such as that of ncRNAs, in combination with AI algorithms, as the development, testing and validation, as well as running costs should be justified by the extra gain in health outcomes. Furthermore, in insurance-covered healthcare systems, patients may also risk that insurance companies employ AI to discover otherwise unknown medical information that could increase costs for individual patients (by either denying coverage or increasing premiums). Thus, there are also important considerations regarding the protection of patient rights.

Overall, it is therefore not surprising that first AI international regulations have already been proposed. For instance, on December 2023 the European Union Parliament and Council reached a political deal on a bill to ensure AI in Europe is safe, respects fundamental rights and democracy. Ultimately, the aim is to establish obligations for AI based on its potential risks and level of impact.[89] Nevertheless, as gaps in AI regulation remain, it is important to assess ethical aspects based on soft ethics guidelines that go beyond hard legal requirements. Broad and interdisciplinary expertise will be needed to make the required assessment of trustworthiness of AI/ML-based tools that will eventually support their implementation in research and clinical routine.

Conclusions and future steps

The successful integration of ncRNA biomarkers into clinical practice demands substantial efforts, encompassing the establishment of experimental standards, rigorous evaluation of the biomarker value and comprehensive functional investigations aimed at understanding their role in disease progression and therapeutic potential. In this context, the more widespread utilisation of ML for biomarker development holds the potential to expedite the identification of disease-relevant ncRNAs, thereby reducing both the cost and time needed for these discoveries to reach clinicians and patients. Furthermore, ML methodologies offer a more unbiased exploration of features, allowing for the identification of unexpected or previously overlooked molecules that might remain undiscovered if solely guided by existing medical knowledge.

These formidable tasks can only be accomplished through collaborative partnerships between academia and industry. Such partnerships are crucial in advancing the development of clinically applicable and cost-effective molecular tests. These collaborations are pivotal in ensuring the reliability, acceptance, use and sustainability of novel molecular tests based on ncRNAs and ML.

In summary, the fusion of cutting-edge technology, interdisciplinary collaboration and a commitment to rigorous standards holds the promise of unlocking the full potential of ncRNAs as transformative tools in clinical practice and personalised healthcare.

Search strategy and selection criteria

Two databases, PubMed and Scopus, were employed for the literature search. Selected full-text studies underwent individual examination to determine their eligibility for inclusion in the review. Each document was meticulously evaluated to ensure a comprehensive overview of the topic, encompassing all relevant perspectives.

Outstanding questions

Many AI/ML methods are currently employed in ncRNA research. However, it is essential for researchers to understand their practical application, including their benefits, obstacles and ethical dimensions. Additional efforts at different levels are necessary to integrate ML into ncRNA research for both biomarker development and molecular phenotyping.

Contributors

Conception and design: DdGC and GK. Data acquisition: All authors. Manuscript drafting: DdGC, MPP and GK. Writing: All authors Review & editing: All authors. Final approval of the submitted version: All authors.

Declaration of interests

YD holds patents and licensing agreements related to the use of RNAs for diagnostic and therapeutic purposes and is Scientific Advisory Board (SAB) member of Firalis SA. The other authors declare no competing interests.

Acknowledgements

This article is based upon work from COST Action AtheroNET, CA21153, supported by COST (European Cooperation in Science and Technology). This article is based upon work from COST Action CardioRNA, CA17129, supported by COST (European Cooperation in Science and Technology).

DdG-C has received financial support from the Instituto de Salud Carlos III (Miguel Servet 2020: CP20/00041), co-funded by the European Union. DdGC was further funded by Fundación Francisco Soria Melguizo (Madrid, Spain), Beca SEPAR – Ayuda a la investigación (1437/2023) and Beca SOCAP – Investigador emergent. CIBERES (CB07/06/2008) is an initiative of the Instituto de Salud Carlos III. MP is the recipient of a predoctoral fellowship (PFIS 2023: FI23/00022) from Instituto de Salud Carlos III and co-funded by the European Union. LTD acknowledges grants from the Novo Nordisk Foundation (NNF22OC0078203 and NNF23OC0081177) and Innovation Fund Denmark (1044-00139B, 0154-00054B). YD has received funding from the EU Horizon 2020 project COVIRNA (grant agreement # 101016072), the National Research Fund (grants #C14/BM/8225223, C17/BM/11613033 and COVID-19/2020-1/14719577/miRCOVID), the Ministry of Higher Education and Research, and the Heart Foundation-Daniel Wagner of Luxembourg. GK acknowledges lab support provided by grants from the Icelandic Research Fund (217946-051), Icelandic Cancer Society Research Fund and University of Iceland Research Fund.

The funders played no role in the design of the study, data collection, data analysis, interpretation of results or writing of the paper.

References

1 Frith MC, Pheasant M, Mattick JS. The amazing complexity of the human transcriptome. *Eur J Hum Genet*. 2005;13:894–897.
2 Liu G, Mattick JS, Taft RJ. A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle*. 2013;12:2061–2072.
3 Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*. 2021;22:96–118.
4 Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. *Cell*. 2012;148:1172–1187.
5 Nemeth K, Bayraktar R, Ferracin M, Calin GA. Non-coding RNAs in disease: from mechanisms to therapeutics. *Nat Rev Genet*. 2024;25:211–232.
6 Eichner H, Karlsson J, Loh E. The emerging role of bacterial regulatory RNAs in disease. *Trends Microbiol*. 2022;30:959–972.
7 Adrian Calin G, Dan Dumitru C, Shimizu M, et al. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A*. 2002;99:15524–15529.
8 Singh N, Ramnarine VR, Song JH, et al. The long noncoding RNA H19 regulates tumor plasticity in neuroendocrine prostate cancer. *Nat Commun*. 2021;12. https://doi.org/10.1038/S41467-021-26901-9.
9 Hunkler HJ, Groß S, Thum T, Bär C. Non-coding RNAs: key regulators of reprogramming, pluripotency, and cardiac cell specification with therapeutic perspective for heart regeneration. *Cardiovasc Res*. 2022;118:3071–3084.
10 Shirvani H, Ghanavi J, Aliabadi A, et al. MiR-211 plays a dual role in cancer development: from tumor suppressor to tumor enhancer. *Cell Signal*. 2023;101. https://doi.org/10.1016/J.CELLSIG.2022.110504.
11 Shah AM, Giacca M. Small non-coding RNA therapeutics for cardiovascular disease. *Eur Heart J*. 2022;43:4548–4561.
12 Täubel J, Hauke W, Rump S, et al. Novel antisense therapy targeting microRNA-132 in patients with heart failure: results of a first-in-human Phase 1b randomized, double-blind, placebo-controlled study. *Eur Heart J*. 2021;42:178–188.
13 Francesco Ruggiero C, Fattore L, Terrenato I, et al. Identification of a miRNA-based non-invasive predictive biomarker of response to target therapy in BRAF-mutant melanoma. *Theranostics*. 2022;12:7420–7430.
14 Dong Y, Gao Q, Chen Y, et al. Identification of CircRNA signature associated with tumor immune infiltration to predict therapeutic efficacy of immunotherapy. *Nat Commun*. 2023;14. https://doi.org/10.1038/S41467-023-38232-Y.
15 Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat Cell Biol*. 2007;9:654–659.
16 Walter E, Dellago H, Grillari J, Dimai HP, Hackl M. Cost-utility analysis of fracture risk assessment using microRNAs compared with standard tools and no monitoring in the Austrian female population. *Bone*. 2018;108:44–54.
17 Devaux Y. MicroRNAs as biomarkers in the brain-heart axis? *Eur Heart J Acute Cardiovasc Care*. 2022;11:617–619.
18 Giannella A, Castelblanco E, Zambon CF, et al. Circulating small noncoding RNA profiling as a potential biomarker of atherosclerotic plaque composition in type 1 diabetes. *Diabetes Care*. 2023;46:551–560.
19 de Gonzalo-Calvo D, Sopić M, Devaux Y. Methodological considerations for circulating long noncoding RNA quantification. *Trends Mol Med*. 2022;28:616–618.
20 de Gonzalo-Calvo D, Pérez-Boza J, Curado J, Devaux Y. Challenges of microRNA-based biomarkers in clinical application for cardiovascular diseases. *Clin Transl Med*. 2022;12. https://doi.org/10.1002/CTM2.585.
21 García-Llorca A, Kararigas G. Sex-related effects of gut microbiota in metabolic syndrome-related diabetic retinopathy. *Microorganisms*. 2023;11. https://doi.org/10.3390/MICROORGANISMS11020447.
22 Horvath C, Kararigas G. Sex-dependent mechanisms of cell death modalities in cardiovascular disease. *Can J Cardiol*. 2022;38:1844–1853.
23 Siokatas G, Papatheodorou I, Daiou A, Lazou A, Hatzistergos KE, Kararigas G. Sex-related effects on cardiac development and disease. *J Cardiovasc Dev Dis*. 2022;9. https://doi.org/10.3390/JCDD9030090.
24 Li S, Kararigas G. Role of biological sex in the cardiovascular-gut microbiome Axis. *Front Cardiovasc Med*. 2022;8. https://doi.org/10.3389/FCVM.2021.759735.
25 Kararigas G. Sex-biased mechanisms of cardiovascular complications in COVID-19. *Physiol Rev*. 2022;102:333–337.
26 Gaignebet L, Kańduła MM, Lehmann D, Knosalla C, Kreil DP, Kararigas G. Sex-specific human cardiomyocyte gene regulation in left ventricular pressure overload. *Mayo Clin Proc*. 2020;95:688–697.
27 Sanchez-Ruderisch H, Queirós AM, Fliegner D, Eschen C, Kararigas G, Regitz-Zagrosek V. Sex-specific regulation of cardiac

microRNAs targeting mitochondrial proteins in pressure overload. *Biol Sex Differ*. 2019;10. https://doi.org/10.1186/S13293-019-0222-1.
28 Gaignebet L, Kararigas G. En route to precision medicine through the integration of biological sex into pharmacogenomics. *Clin Sci (Lond)*. 2017;131:329–342.
29 Kararigas G, Seeland U, De Arellano MLB, Dworatzek E, Regitz-Zagroseki V. Why the study of the effects of biological sex is important. Commentary. *Ann Ist Super Sanita*. 2016;52:149–150.
30 Schulte C, Barwari T, Joshi A, et al. Comparative analysis of circulating noncoding RNAs versus protein biomarkers in the detection of myocardial injury. *Circ Res*. 2019;125:328–340.
31 Wong LL, Zou R, Zhou L, et al. Combining circulating MicroRNA and NT-proBNP to detect and categorize heart failure subtypes. *J Am Coll Cardiol*. 2019;73:1300–1313.
32 Mattick JS, Amaral PP, Carninci P, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol*. 2023;24:430–447.
33 Jackson HR, Zandstra J, Menikou S, et al. A multi-platform approach to identify a blood-based host protein signature for distinguishing between bacterial and viral infections in febrile children (PERFORM): a multi-cohort machine learning study. *Lancet Digit Health*. 2023;5:e774–e785.
34 Karađuzović-Hadžiabdić K, Peters A. Artificial intelligence in clinical decision-making for diagnosis of cardiovascular disease using epigenetics mechanisms. *Epigenetics Cardiovasc Dis*. 2021:327–345.
35 Errington N, Iremonger J, Pickworth JA, et al. A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach. *eBioMedicine*. 2021;69:103444. https://doi.org/10.1016/J.EBIOM.2021.103444.
36 Liu Z, Liu L, Weng S, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. *Nat Commun*. 2022;13. https://doi.org/10.1038/S41467-022-28421-6.
37 Liu Z, Guo CG, Dang Q, et al. Integrative analysis from multicenter studies identities a consensus machine learning-derived lncRNA signature for stage II/III colorectal cancer. *eBioMedicine*. 2022;75. https://doi.org/10.1016/J.EBIOM.2021.103750.
38 Li Y, Zhao J, Yu S, et al. Extracellular vesicles long RNA sequencing reveals abundant mRNA, circRNA, and lncRNA in human blood as potential biomarkers for cancer diagnosis. *Clin Chem*. 2019;65:798–808.
39 Hu X, Liao S, Bai H, et al. Integrating exosomal microRNAs and electronic health data improved tuberculosis diagnosis. *EBioMedicine*. 2019;40:564–573.
40 Kayvanpour E, Gi WT, Sedaghat-Hamedani F, et al. microRNA neural networks improve diagnosis of acute coronary syndrome (ACS). *J Mol Cell Cardiol*. 2021;151:155–162.
41 Devaux Y, Zhang L, Lumley AI, et al. Development of a long noncoding RNA-based machine learning model to predict COVID-19 in-hospital mortality. *Nat Commun*. 2024;15:4259.
42 Firat H, Zhang L, Baksi S, et al. FIMICS: a panel of long noncoding RNAs for cardiovascular conditions. *Heliyon*. 2023;9. https://doi.org/10.1016/J.HELIYON.2023.E13087.
43 García-Hidalgo MC, González J, Benítez ID, et al. Identification of circulating microRNA profiles associated with pulmonary function and radiologic features in survivors of SARS-CoV-2-induced ARDS. *Emerg Microbes Infect*. 2022;11:1537–1549.
44 Goretti E, Wagner DR, Devaux Y. miRNAs as biomarkers of myocardial infarction: a step forward towards personalized medicine? *Trends Mol Med*. 2014;20:716–725.
45 Katipally RR, Martinez CA, Pugh SA, et al. Integrated clinical-molecular classification of colorectal liver metastases: a biomarker analysis of the phase 3 new EPOC randomized clinical trial. *JAMA Oncol*. 2023;9:1245–1254.
46 de Gonzalo-Calvo D, Vilades D, Martínez-Camblor P, et al. Circulating microRNAs in suspected stable coronary artery disease: a coronary computed tomography angiography study. *J Intern Med*. 2019;286:341–355.
47 de Gonzalo-Calvo D, Martínez-Camblor P, Bär C, et al. Improved cardiovascular risk prediction in patients with end-stage renal disease on hemodialysis using machine learning modeling and circulating microribonucleic acids. *Theranostics*. 2020;10:8665–8676.
48 de Gonzalo-Calvo D, Martinez-Camblor P, Belmonte T, et al. Circulating miR-133a-3p defines a low-risk subphenotype in patients with heart failure and central sleep apnea: a decision tree machine learning approach. *J Transl Med*. 2023;21. https://doi.org/10.1186/S12967-023-04558-W.
49 Vilades D, Martínez-Camblor P, Ferrero-Gregori A, et al. Plasma circular RNA hsa_circ_0001445 and coronary artery disease: performance as a biomarker. *FASEB J*. 2020;34:4403–4414.
50 Reel PS, Reel S, van Kralingen JC, et al. Machine learning for classification of hypertension subtypes using multi-omics: a multicentre, retrospective, data-driven study. *eBioMedicine*. 2022;84. https://doi.org/10.1016/J.EBIOM.2022.104276.
51 Chu G, Ji X, Wang Y, Niu H. Integrated multiomics analysis and machine learning refine molecular subtypes and prognosis for muscle-invasive urothelial cancer. *Mol Ther Nucleic Acids*. 2023;33:110–126.
52 Eckhardt CM, Gambazza S, Bloomquist TR, et al. Extracellular vesicle-encapsulated microRNAs as novel biomarkers of lung health. *Am J Respir Crit Care Med*. 2023;207:50–59.
53 Yang Y, Huang N, Hao L, Kong W. A clustering-based approach for efficient identification of microRNA combinatorial biomarkers. *BMC Genomics*. 2017;18. https://doi.org/10.1186/S12864-017-3498-8.
54 Perez-Pons M, Molinero M, Benítez ID, et al. MicroRNA-centered theranostics for pulmoprotection in critical COVID-19. *Mol Ther Nucleic Acids*. 2024;35. https://doi.org/10.1016/J.OMTN.2024.102118.
55 Lakkisto P, Dalgaard LT, Belmonte T, Pinto-Sietsma SJ, Devaux Y, de Gonzalo-Calvo D. Development of circulating microRNA-based biomarkers for medical decision-making: a friendly reminder of what should NOT be done. *Crit Rev Clin Lab Sci*. 2023;60:141–152.
56 Rios R, Miller RJH, Manral N, et al. Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: insights from REFINE SPECT registry. *Comput Biol Med*. 2022;145. https://doi.org/10.1016/J.COMPBIOMED.2022.105449.
57 Liu M, Li S, Yuan H, et al. Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques. *Artif Intell Med*. 2023;142. https://doi.org/10.1016/J.ARTMED.2023.102587.
58 Higuchi C, Tanaka T, Okada Y. Systematic comparison of machine learning methods for identification of miRNA species as disease biomarkers. *Lect Notes Comput Sci*. 2015;9044:386–394.
59 Wong WKM, Thorat V, Joglekar MV, et al. Analysis of half a billion datapoints across ten machine-learning algorithms identifies key elements associated with insulin transcription in human pancreatic islet cells. *Front Endocrinol*. 2022;13. https://doi.org/10.3389/FENDO.2022.853863.
60 García-Hidalgo MC, Benítez ID, Perez-Pons M, et al. MicroRNA-guided drug discovery for mitigating persistent pulmonary complications in critical COVID-19 survivors: a longitudinal pilot study. *Br J Pharmacol*. 2024. https://doi.org/10.1111/BPH.16330.
61 Sammut SJ, Crispin-Ortuzar M, Chin SF, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;601:623–629.
62 Doudesis D, Lee KK, Boeddinghaus J, et al. Machine learning for diagnosis of myocardial infarction using cardiac troponin concentrations. *Nat Med*. 2023;29:1201–1210.
63 Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J*. 2021;19:3735–3746.
64 Pepe G, Appierdo R, Carrino C, Ballesio F, Helmer-Citterich M, Gherardini PF. Artificial intelligence methods enhance the discovery of RNA interactions. *Front Mol Biosci*. 2022;9. https://doi.org/10.3389/FMOLB.2022.1000205.
65 Gysi DM, Barabási AL. Noncoding RNAs improve the predictive power of network medicine. *Proc Natl Acad Sci U S A*. 2023;120. https://doi.org/10.1073/PNAS.2301342120.
66 Alam T, Al-Absi HRH, Schmeier S. Deep learning in LncRNAome: contribution, challenges, and perspectives. *Noncoding RNA*. 2020;6:1–23.
67 Cui L, Lu Y, Sun J, et al. RFLMDA: a novel reinforcement learning-based computational model for human MicroRNA-disease association prediction. *Biomolecules*. 2021;11. https://doi.org/10.3390/BIOM11121835.
68 Su B, Wang W, Lin X, Liu S, Huang X. Identifying the potential miRNA biomarkers based on multi-view networks and reinforcement learning for diseases. *Brief Bioinform*. 2023;25. https://doi.org/10.1093/BIB/BBAD427.
69 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. https://doi.org/10.1186/GB-2010-11-8-R90.

70 Pinilla L, Barbé F, de Gonzalo-Calvo D. MicroRNAs to guide medical decision-making in obstructive sleep apnea: a review. *Sleep Med Rev*. 2021;59. https://doi.org/10.1016/J.SMRV.2021.101458.

71 Bonneau E, Neveu B, Kostantin E, Tsongalis GJ, De Guire V. How close are miRNAs from clinical practice? A perspective on the diagnostic and therapeutic market. *EJIFCC*. 2019;30:114.

72 Winkle M, El-Daly SM, Fabbri M, Calin GA. Noncoding RNA therapeutics - challenges and potential solutions. *Nat Rev Drug Discov*. 2021;20:629–651.

73 Ha Thi HT, Than VT. Recent applications of RNA therapeutic in clinics. *Prog Mol Biol Transl Sci*. 2024;203:115–150.

74 Hong DS, Kang YK, Borad M, et al. Phase 1 study of MRX34, a liposomal miR-34a mimic, in patients with advanced solid tumours. *Br J Cancer*. 2020;122:1630–1637.

75 Cui T, Dou Y, Tan P, et al. RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation. *Nucleic Acids Res*. 2022;50:D333–D339.

76 Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*. 2013;41. https://doi.org/10.1093/NAR/GKS1193.

77 Welcoming protocols.io. *Nat Protoc*. 2024. https://doi.org/10.1038/S41596-024-01012-Z.

78 Gilroy SP, Kaplan BA. Furthering open science in behavior analysis: an introduction and tutorial for using GitHub in research. *Perspect Behav Sci*. 2019;42:565–581.

79 Bentzien J, Muegge I, Hamner B, Thompson DC. Crowd computing: using competitive dynamics to develop and refine highly predictive models. *Drug Discov Today*. 2013;18:472–478.

80 Zicari RV, Brodersen J, Brusseau J, et al. Z-Inspection ®: a process to assess trustworthy AI. *IEEE Trans Technol Soc*. 2021;2:83–97.

81 Vetter D, Amann J, Bruneault F, et al. Lessons learned from assessing trustworthy AI in practice. *Digital Society*. 2023;2:1–25.

82 Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ Digit Med*. 2021;4. https://doi.org/10.1038/S41746-021-00509-1.

83 Zicari RV, Brusseau J, Blomberg SN, et al. On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics*. 2021;3:673104.

84 Allahabadi H, Amann J, Balot I, et al. Assessing trustworthy AI in times of COVID-19: deep learning for predicting a multiregional score conveying the degree of lung compromise in COVID-19 patients. *IEEE Trans Technol Soc*. 2022;3:272–289.

85 Uddin M, Wang Y, Woodbury-Smith M. Artificial intelligence for precision medicine in neurodevelopmental disorders. *NPJ Digit Med*. 2019;2. https://doi.org/10.1038/S41746-019-0191-0.

86 Walsh CG, Chaudhry B, Dua P, et al. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. *JAMIA Open*. 2020;3:9–15.

87 Tat E, Bhatt DL, Rabbat MG. Addressing bias: artificial intelligence in cardiovascular medicine. *Lancet Digit Health*. 2020;2:e635–e636.

88 Lee P, Le Saux M, Siegel R, et al. Racial and ethnic disparities in the management of acute pain in US emergency departments: meta-analysis and systematic review. *Am J Emerg Med*. 2019;37:1770–1777.

89 Artificial intelligence act: deal on comprehensive rules for trustworthy AI | News | European Parliament. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai. Accessed December 14, 2023.