Heliyon 11 (2025) e41488

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

RNA sequence analysis landscape: A comprehensive review of task types, databases, datasets, word embedding methods, and language models

Muhammad Nabeel Asim [a,*], Muhammad Ali Ibrahim [a,b], Tayyaba Asif [b], Andreas Dengel [a,b]

[a] German Research Center for Artificial Intelligence GmbH, Kaiserslautern, 67663, Germany
[b] Department of Computer Science, Rhineland-Palatinate Technical University of Kaiserslautern-Landau, Kaiserslautern, 67663, Germany

ARTICLE INFO

Keywords:
RNA sequence analysis
Deep learning
Machine learning
Artificial intelligence
AI applications in genomics
Multi-omics
RNA-sequencing
Gene expression regulation
Gene analysis
Gene network analysis
RNA modifications
RNA functional analysis
Disease analysis

ABSTRACT

Deciphering information of RNA sequences reveals their diverse roles in living organisms, including gene regulation and protein synthesis. Aberrations in RNA sequence such as dysregulation and mutations can drive a diverse spectrum of diseases including cancers, genetic disorders, and neurodegenerative conditions. Furthermore, researchers are harnessing RNA's therapeutic potential for transforming traditional treatment paradigms into personalized therapies through the development of RNA-based drugs and gene therapies. To gain insights of biological functions and to detect diseases at early stages and develop potent therapeutics, researchers are performing diverse types RNA sequence analysis tasks. RNA sequence analysis through conventional wet-lab methods is expensive, time-consuming and error prone. To enable large-scale RNA sequence analysis, empowerment of wet-lab experimental methods with Artificial Intelligence (AI) applications necessitates scientists to have a comprehensive knowledge of both DNA and AI fields. While molecular biologists encounter challenges in understanding AI methods, computer scientists often lack basic foundations of RNA sequence analysis tasks. Considering the absence of a comprehensive literature that bridges this research gap and promotes the development of AI-driven RNA sequence analysis applications, the contributions of this manuscript are manifold: It equips AI researchers with biological foundations of 47 distinct RNA sequence analysis tasks. It sets a stage for development of benchmark datasets related to 47 distinct RNA sequence analysis tasks by facilitating cruxes of 64 different biological databases. It presents word embeddings and language models applications across 47 distinct RNA sequence analysis tasks. It streamlines the development of new predictors by providing a comprehensive survey of 58 word embeddings and 70 language models based predictive pipelines performance values as well as top performing traditional sequence encoding based predictors and their performances across 47 RNA sequence analysis tasks.

[*] Corresponding author.
E-mail address: Muhammad_Nabeel.Asim@dfki.de (M.N. Asim).

https://doi.org/10.1016/j.heliyon.2024.e41488
Received 28 September 2024; Received in revised form 23 December 2024; Accepted 24 December 2024

Available online 6 January 2025

1. Introduction

Cutting-edge sequencing technologies, such as next-generation sequencing and the innovative third-generation sequencing, have revolutionized the exploration of genetic sequences in a cost-efficient manner [1]. These methods have generated vast amounts of DNA, RNA, and protein sequence data [1]. In particular, RNA sequence data is being utilized to uncover hidden information such as distinct roles of RNAs in living organisms (e.g. protein synthesis and gene regulation) and their associations with various diseases, including cancers, genetic disorders, and neurodegenerative conditions [2]. To gain deep insights of RNA sequences information, researchers are utilizing the potential of wet-lab experimental approaches for performing diverse types of RNA sequence analysis tasks [3]. However, wet-lab experiments based RNA sequence analysis is time consuming, and expensive. Following inherent limitations of conventional wet-lab experimental approaches, researchers are harnessing the potential of Artificial Intelligence (AI) methods to develop AI-driven RNA sequence analysis applications [3].

Most of the AI-driven RNA sequence analysis applications fall under the hood of regression, clustering, and classification paradigms. Clustering paradigm objective is to make groups of RNA sequences with similar characteristics [3,1]. Regression paradigm focuses on the prediction of continuous numerical values based on RNA-seq data [3,1]. For instance, researchers might utilize regression to predict how a specific gene's expression level might change under varying environmental conditions [3,1]. Classification paradigm involves assigning RNA sequences to pre-defined categories [3,1]. A unified workflow for all three distinct types paradigms based AI-driven RNA sequence analysis predictive pipelines are illustrated in Fig. 1. A closer look on Fig. 1 reveals that AI-driven RNA sequence analysis predictive pipelines working paradigm can be segregated into four different stages.

The first stage focuses on the collection and curation of high-quality benchmark datasets. This stage either employs datasets developed by existing studies or creates new datasets. The creation of new datasets involves obtaining RNA sequences and their associated information from public databases or acquiring data through wet-lab experiments. The second stage is known as representation learning, it employs diverse methods to capture the informative distribution of nucleotides in RNA sequences and encode this information into statistical vectors. This transformation is essential because AI methods inherently rely on statistical vectors. The third stage utilizes statistical vectors of RNA sequences alongside machine learning or deep learning algorithms to make predictions. The objective of the fourth stage is to evaluate the performance of predictive pipelines that utilize representation learning and machine/deep learning methods. Among all 4 stages, representation learning stage is the most critical as quality statistical vectors allow even simple machine learning algorithms to perform well, while poor vectors hinder the performance of sophisticated algorithms. There is a marathon to develop potent sequence encoders for generating informative statistical vectors [4]. Up to date, hundreds of representation learning methods have been developed that can be categorized into three groups: domain-specific methods, neural word embedding methods, and language models [4]. Domain-specific methods utilize pre-computed physical and chemical values of nucleotides or occurrence frequencies of nucleotides to generate statistical vectors of RNA sequences [5,4,6]. Although, these methods manage to capture intrinsic characteristics of biological sequences like nucleotides compositional or distributional information. However, these methods fail to fully capture complex nucleotides relationships and semantic similarities between nucleotides [5,4].

Compared to domain-specific methods, neural word embedding techniques offer multiple advantages. These methods capture and encode distribution and semantic relationships of nucleotides or groups of nucleotides (k-mers) as dense vectors in a continuous vector space [7,8]. They also support transfer learning, as word embeddings are generated in an unsupervised manner. Transfer learning is a technique where a deep learning model first learns to solve one task really well (identification of disease genes) and then applies that knowledge to solve a different but related task (like identification of disease pathways) more efficiently. The model “transfers” its existing understanding of important features, like identifying abnormal genetic patterns, to the new task. Transfer learning strategy empowers machine and deep learning algorithms to perform better even on small datasets. On the other hand, language models learn representations of nucleotides or k-mers by predicting masked nucleotides based on their surrounding context [9–11]. Unlike word embedding methods which generate static k-mer vectors [7,8], language models consider different contexts of nucleotides or k-mers by capturing complex relationships through masked word prediction [9–11]. Similar to word embeddings, language models enhance performance of machine and deep learning-based RNA sequence analysis pipelines with the strength of transfer learning.

Despite the numerous benefits of word embedding approaches and language models, most AI-driven RNA sequence analysis applications still rely on domain-specific methods that transform raw sequences into statistical vectors. Moreover, development of AI-driven RNA-sequence analysis applications requires expertise in both RNA biology and artificial intelligence. Unfortunately, a significant knowledge gap often exists between AI researchers and biologists. AI researchers usually lack in deep understanding of biological applications, while biologists lack in fundamental AI concepts. A significant gap between both fields hinder development of powerful AI-driven sequence analysis applications. For example, Natural Language Processing field has witnessed development of powerful applications which have integrated multi-task learning strategies, but such advancements have not been mirrored in the realm of RNA sequence analysis. This is partly because AI experts often lack a comprehensive understanding of various RNA analysis tasks necessary for developing effective multitask learning strategies based applications.

To address this need and to accelerate the development of robust predictive pipelines for RNA sequence analysis tasks, several review articles have been published. However, these reviews typically concentrate on individual tasks rather than providing a comprehensive overview. Recognizing the importance of bridging the gap between biologists and AI experts, this review paper offers several key contributions:

- Biologists can utilize this review article to gain insights of AI potential for RNA sequence analysis tasks, while AI researchers can gain a deeper understanding of specific challenges and opportunities within RNA sequence analysis field.

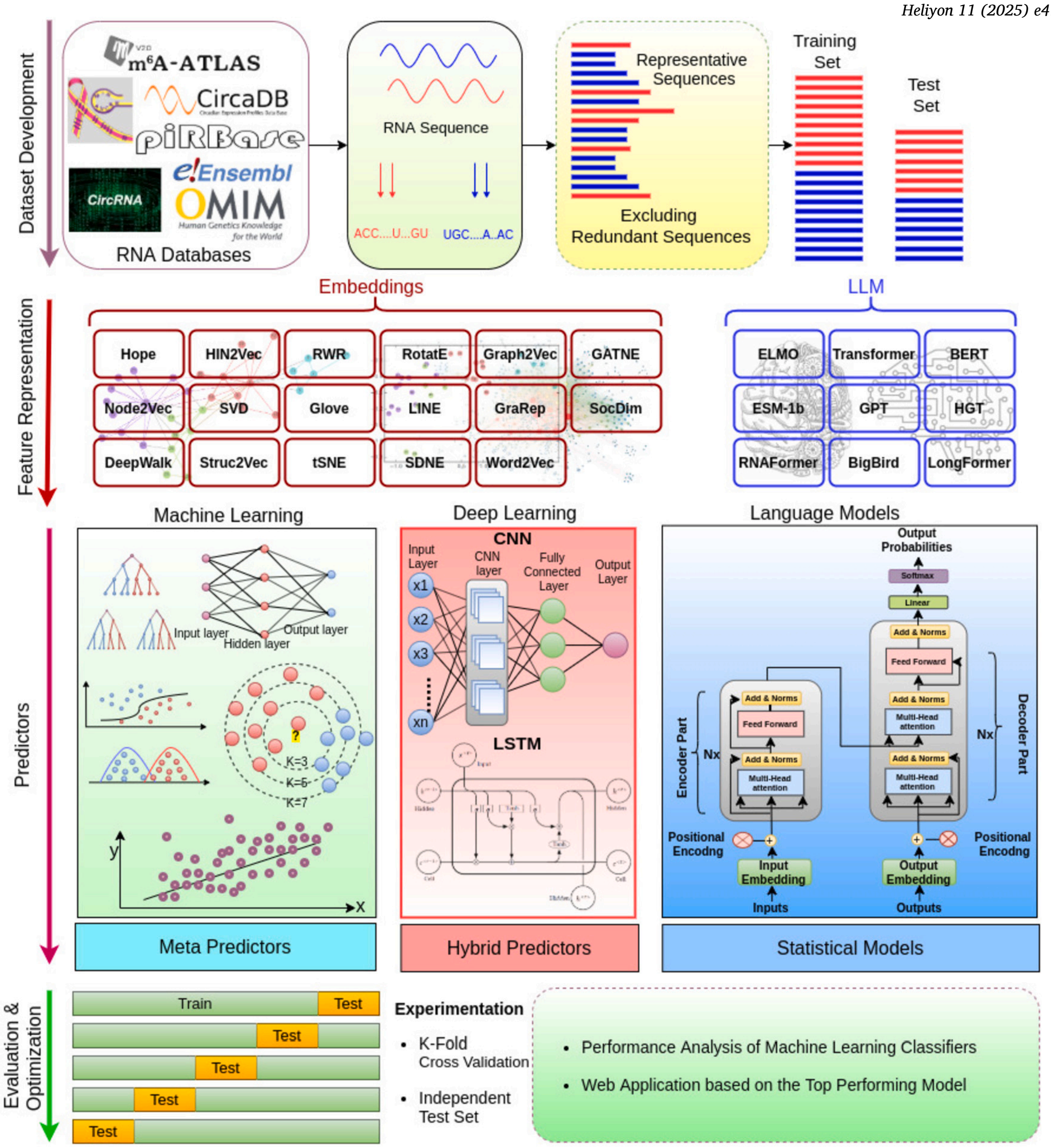

Fig. 1. Predictive pipeline of RNA Sequence Analysis Tasks.

- It empowers AI researchers by imparting biological insights of 47 distinct RNA sequence analysis tasks and aligns these tasks with 3 distinct AI paradigms namely classification, regression and clustering.
- It lays the foundation for the development of new datasets by offering a comprehensive overview of 64 RNA sequence analysis databases.
- To ensure a fair performance comparison between existing and new AI predictors, it provides details of 310 benchmark datasets related to 47 unique RNA sequence analysis tasks.
- Within AI predictive pipelines, it elucidates the application of 16 different word embedding methods and 8 language models across 47 RNA sequence analysis tasks.
- It streamlines novel predictors development by facilitating a detailed summary of current state-of-the-art predictors, their performances across 47 unique RNA sequence analysis tasks, and their availability to scientific community. This detailed summary sets a valuable stage for researchers aiming to develop and evaluate new predictors for distinct types of RNA sequence analysis tasks.

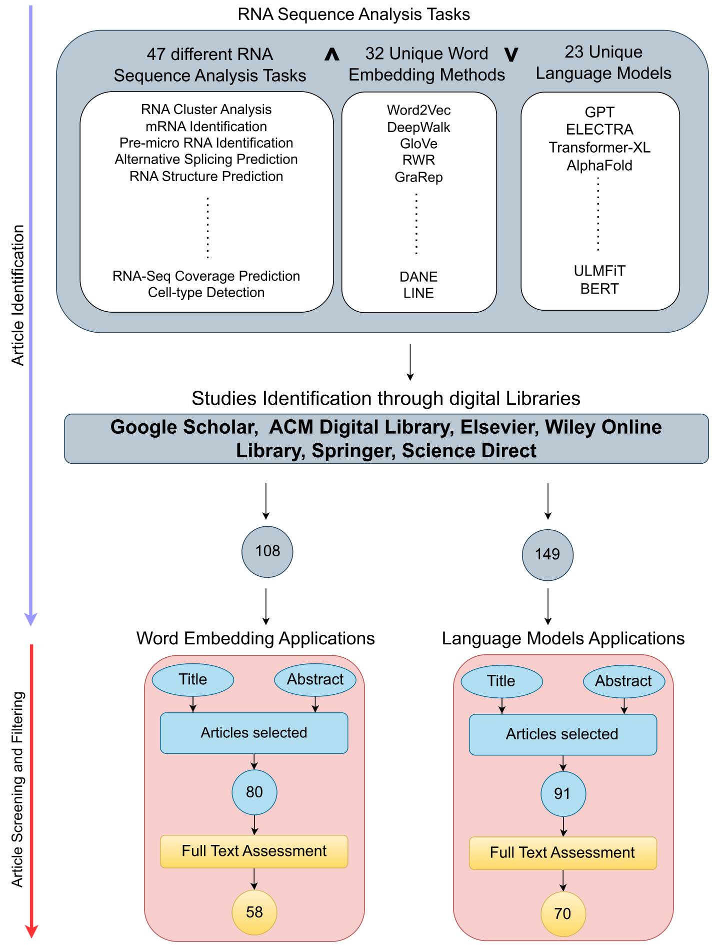
RNA Sequence Analysis Tasks
47 different RNA Sequence Analysis Tasks
∧
32 Unique Word Embedding Methods
∨
23 Unique Language Models
RNA Cluster Analysis
mRNA Identification
Pre-micro RNA Identification
Alternative Splicing Prediction
RNA Structure Prediction
RNA-Seq Coverage Prediction
Cell-type Detection
Word2Vec
DeepWalk
GloVe
RWR
GraRep
DANE
LINE
GPT
ELECTRA
Transformer-XL
AlphaFold
ULMFiT
BERT
Article Identification
Studies Identification through digital Libraries
Google Scholar, ACM Digital Library, Elsevier, Wiley Online Library, Springer, Science Direct
108
149
Word Embedding Applications
Language Models Applications
Article Screening and Filtering
Title
Abstract
Articles selected
80
Full Text Assessment
58
Title
Abstract
Articles selected
91
Full Text Assessment
70

Fig. 2. Research Methodology.

2. Research methodology

This section provides high level overview of research methodology that is used to find word embeddings and language models based articles for RNA sequence analysis applications. To ensure thoroughness and reliability of selected articles, this methodology follows two stage process: 1) Article identification, 2) Article screening and filtering.

2.1. *Article searching*

In Fig. 2, article identification module contains three cells for different kinds of keywords namely RNA sequence analysis tasks, word embedding methods, and Language models. To formulate search queries, keywords within same cell are combined using OR $\vee$ operator while keywords from different cells are combined using AND $\wedge$ operator. For instance few sample queries include; mRNA identification using FastText word embedding, enhancer RNA identification using BERT language model, etc. These search queries are executed on academic search engines such as Google Scholar,[1] ACM Digital Library,[2] IEEEXplore,[3] Elsevier,[4] Wiley Online Library,[5] Springer[6] and ScienceDirect.[7] Furthermore, snowballing is employed to identify more research articles by examining reference list of extracted papers.

[1] https://scholar.google.com/.

[2] https://dl.acm.org/.

[3] https://ieeexplore.ieee.org/.

[4] https://www.elsevier.com/.

[5] https://www.wiley.com/en-us.

[6] https://www.springer.com/gp.

[7] https://www.sciencedirect.com/.

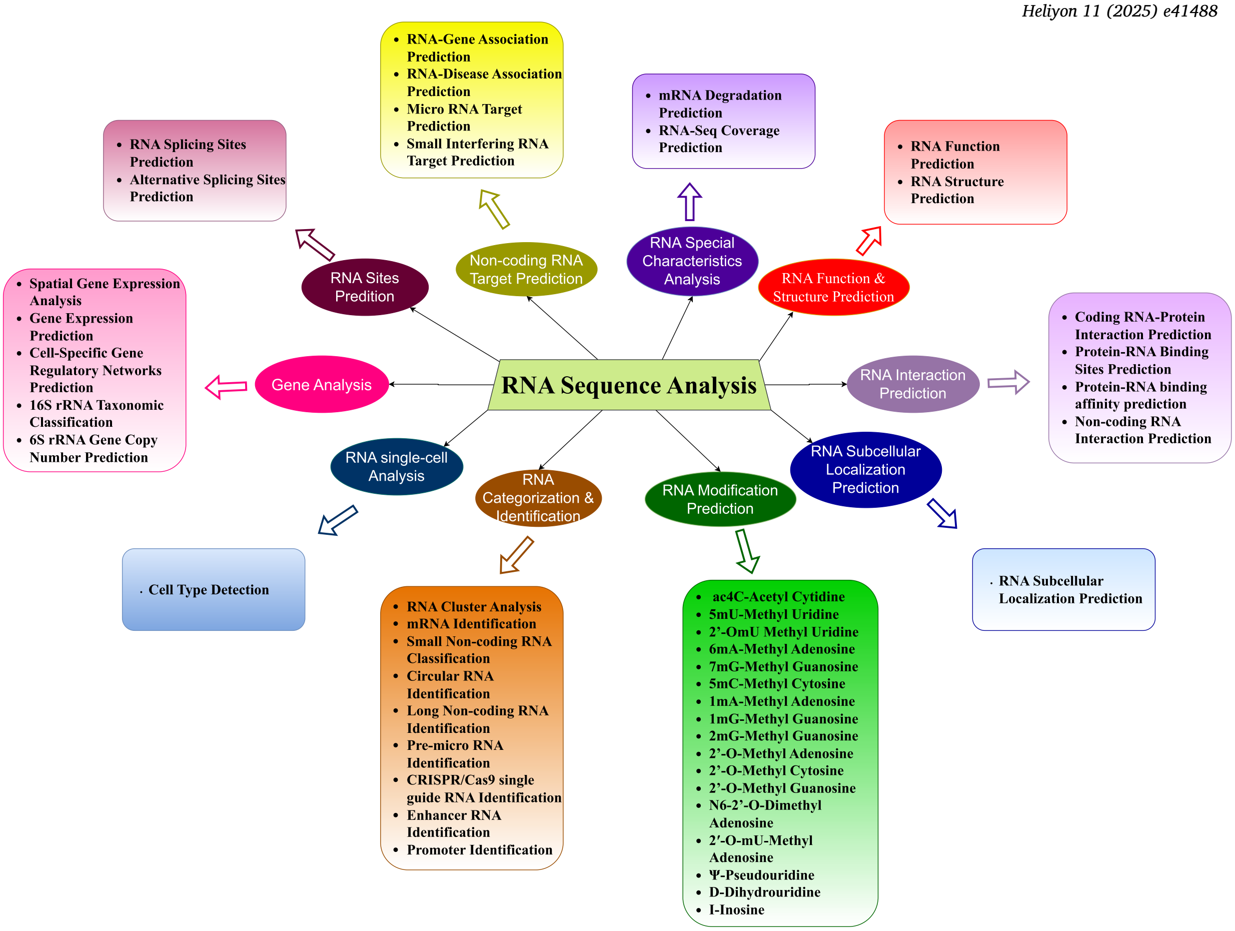
RNA Sequence Analysis
RNA Sites Predition
• RNA Splicing Sites Prediction
• Alternative Splicing Sites Prediction
Non-coding RNA Target Prediction
• RNA-Gene Association Prediction
• RNA-Disease Association Prediction
• Micro RNA Target Prediction
• Small Interfering RNA Target Prediction
RNA Special Characteristics Analysis
• mRNA Degradation Prediction
• RNA-Seq Coverage Prediction
RNA Function & Structure Prediction
• RNA Function Prediction
• RNA Structure Prediction
Gene Analysis
• Spatial Gene Expression Analysis
• Gene Expression Prediction
• Cell-Specific Gene Regulatory Networks Prediction
• 16S rRNA Taxonomic Classification
• 6S rRNA Gene Copy Number Prediction
RNA Interaction Prediction
• Coding RNA-Protein Interaction Prediction
• Protein-RNA Binding Sites Prediction
• Protein-RNA binding affinity prediction
• Non-coding RNA Interaction Prediction
RNA single-cell Analysis
• Cell Type Detection
RNA Categorization & Identification
• RNA Cluster Analysis
• mRNA Identification
• Small Non-coding RNA Classification
• Circular RNA Identification
• Long Non-coding RNA Identification
• Pre-micro RNA Identification
• CRISPR/Cas9 single guide RNA Identification
• Enhancer RNA Identification
• Promoter Identification
RNA Modification Prediction
• ac4C-Acetyl Cytidine
• 5mU-Methyl Uridine
• 2'-OmU Methyl Uridine
• 6mA-Methyl Adenosine
• 7mG-Methyl Guanosine
• 5mC-Methyl Cytosine
• 1mA-Methyl Adenosine
• 1mG-Methyl Guanosine
• 2mG-Methyl Guanosine
• 2'-O-Methyl Adenosine
• 2'-O-Methyl Cytosine
• 2'-O-Methyl Guanosine
• N6-2'-O-Dimethyl Adenosine
• 2′-O-mU-Methyl Adenosine
• Ψ-Pseudouridine
• D-Dihydrouridine
• I-Inosine
RNA Subcellular Localization Prediction
• RNA Subcellular Localization Prediction

Fig. 3. Precise Classification of Unique RNA Sequence Analysis Tasks in 10 Major Biological Goals.

2.2. *Article screening and filtering*

Second stage consists of two step process to select the most relevant articles. In the first step, titles and abstracts of 257 articles were reviewed by domain experts resulting in identification of 80 word embedding and 91 language models based relevant articles. In second step, a full-text assessment of these articles was conducted leading to selection of 58 word embedding and 70 articles language models related articles.

3. Biological foundations of RNA sequence analysis goals and tasks

This section offers a high-level overview of RNA sequence analysis world. Scientists are performing around 47 notable sequence analysis tasks to gain a deeper understanding of RNA's diverse biological roles within living organisms, their associations with various diseases, and potentials for therapeutic development. To facilitate a more organized comprehension of these tasks, we have categorized them into 10 distinct goals, presented in Fig. 3. RNA is emerging as a key player in understanding cellular functions and providing versatile targets for novel therapeutics. To gain unprecedented insights into the intricacies of RNA regulation at the molecular level, researchers need to decode RNA's complexities, characterize the composition and structure of RNA, and uncover their functions. Also, they need to decipher the complex regulatory networks that govern their activity and determine their relevance and alterations in disease. The heart of such comprehensive analysis is the goal of RNA classification which focuses on identifying different types of RNAs on the basis of their molecular characteristics and biological roles [12]. Few notable types are miRNAs, tRNAs, lncRNAs, circular RNAs, enhancer RNAs and promoter RNAs [13,14]. RNA classification landscape is advancing the discovery of new RNAs and expanding scientists understanding of RNA's regulatory potential in living organisms [15]. The immense diversity in the roles and cellular activities of unique RNAs emerges from a complex interplay of molecular characteristics. This complex interplay includes distinct localization patterns [16–19], structural characteristics, interactions patterns [20–23], and functional properties. RNA structure and subcellular localization exhibit a bidirectional relationship as structural elements can direct cellular localization through specific recognition motifs, while the local cellular environment can also influence RNA folding and stability. These structural and spatial arrangements facilitate specific interaction networks with proteins, chromatin, and other RNA molecules which dictate

RNAs functions [24]. Understanding these interconnected relationships is crucial for deciphering RNA function and developing RNA-based therapeutic strategies.

Furthermore, RNA modifications [25] represent another layer of complexity in biological systems, where chemical alterations to nucleotides significantly influence molecular stability, function, and regulatory potential. These modifications, including N6-methyladenosine (m6A), ac4C-Acetyl Cytidine, and various 2'-O-methyl modifications, work in concert with RNA special characteristics like prediction of degradation rates of mRNA molecules [26] and prediction of coverage or read counts of RNA-seq experiments [27] to create sophisticated recognition platforms for cellular factors and activate or inhibit certain cellular processes. The prediction and understanding of these modifications are essential for comprehending RNA processing, nuclear export, translation, and cellular differentiation. Furthermore, RNA target prediction [28,29] has emerged as a crucial aspect which focuses specifically on interactions between regulatory RNAs and their targets. This includes miRNA interactions with mRNAs and coding transcript sequences, siRNA interactions with genes, and non-coding RNA associations with diseases. These targeting relationships play vital roles in gene regulation, disease pathogenesis, and therapeutic development. RNA site prediction [30,31] complements this analysis as it focuses on crucial regulatory elements such as splice sites and alternative splicing patterns, which are fundamental to gain understanding of post-transcriptional regulation. Additionally, comprehensive gene analysis [32] encompasses various aspects including spatial gene expression patterns, gene regulatory networks, and taxonomic classification of microbial species based on RNA sequences. The emergence of single-cell RNA analysis [33,34] has further revolutionized scientists understanding by enabling the examination of RNA expression patterns, cellular heterogeneity, and regulatory networks at unprecedented resolution. Such analyses are decoding multi-omics data to provide insights into cellular diversity and molecular mechanisms at the individual cell level.

For detailed exploration of RNA biology fundamentals, specific aspects of each goal, and AI utility trends in RNA biology, readers are directed to comprehensive reviews [3,35–40]. The subsequent sections will delve into the nature of RNA sequence analysis tasks, and AI approaches developed for these tasks to address 10 major goals effectively.

4. Examining RNA sequence analysis tasks through the lens of computer scientists

Given the surge in biological data and the emergence of AI technologies, researchers are increasingly applying AI methodologies across various domains of molecular biology. The development of large-scale AI applications necessitates a comprehensive understanding of a wide array of sequence analysis tasks. However, a significant gap exists between the expertise of computer scientists and molecular biologists. While molecular biologists grasp the necessity, biological significance, and pharmaceutical value of diverse sequence analysis tasks, they often lack insight into selecting the most suitable machine learning or deep learning models to complement or substitute experimental work. Conversely, computer scientists are adept at identifying which AI predictive pipelines may yield optimal results with specific data types, yet they struggle to comprehend the nature of biological sequence analysis tasks. For example, RNA sequence analysis tasks such as RNA function prediction and cell-specific gene regulatory network prediction are challenging to grasp straightaway. Nevertheless, a detailed literature review which describes the basics of such tasks can significantly bridge this gap. For example, RNA function prediction seems like a multi-class classification task but it is actually a multi-label classification task. Similarly, cell specific gene regulatory network prediction seems like a clustering task but it is actually a binary classification task. With this foundation knowledge, computer scientists can more effectively design predictive pipelines tailored to binary, multi-class, multi-label classification, regression, and clustering tasks. To empower diverse AI researchers and practitioners, we have performed methodical categorization of 47 RNA sequence analysis tasks in Fig. 4 on the basis of their nature. A first look at Fig. 4 indicates that RNA sequence analysis tasks can be broadly classified into 3 primary kinds: regression, classification, and clustering, where classification can be further segregated into 3 secondary kinds: binary, classification, multi-class classification, as well as multi-label classification. Let's dive into mathematical formulation of unique types of RNA sequence analysis tasks.

For binary classification, main objective for researchers is to forecast the result of a binary variable (0 or 1). When provided with a dataset containing features $X \in \mathbb{R}^{nxd}$, binary labels $y \in 0, 1$, and a training dataset $(x_1, y_1), (x_2, y_2), \ldots,$ according to equation (1), the aim is to acquire a decision function $f : X \to Y$ that assigns inputs to binary outcomes $0, 1$ using the hypothesis function $h(x)$ derived from the training data.

$$f(x) = \begin{cases} 1 & if\, h(x) \geqslant 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In the multi-class classification, the objective for researchers is to forecast the outcome from a set of more than two classes. Specifically, when presented with a dataset containing features $X \in \mathbb{R}^{nxd}$, labels $y \in 1, 2, \ldots, K$ where K represents the total number of classes, and a training dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i \in X$ and $y_i \in Y$, according to equation (2), the aim is to develop a decision function $f : X \to Y$ that assigns inputs to one of the available classes.

$$f(x) = argmax_k h_k(x) \tag{2}$$

The hypothesis function $h_k(x)$ represents the learned hypothesis for class k derived from the training data. Conversely, in multi-label classification, each input has the potential to be associated with several classes at the same time. When provided with a dataset comprising features $X \in \mathbb{R}^{nxd}$, labels $y \in 1, 2, \ldots, K$ where K denotes the total number of classes, and a training dataset $(x_1, y_1, y_2, ..), (x_2, y_1, y_4, \ldots), \ldots, (x_n, y5, y_n, \ldots.)$ where $x_i \in X$ and $y_i \in Y$, according to equation (3), the objective is to develop a decision function $f : X \to 0, 1^K$ that simultaneously assigns inputs to multiple classes utilizing the hypothesis function $h_k(x)$ for class k obtained from the training data.

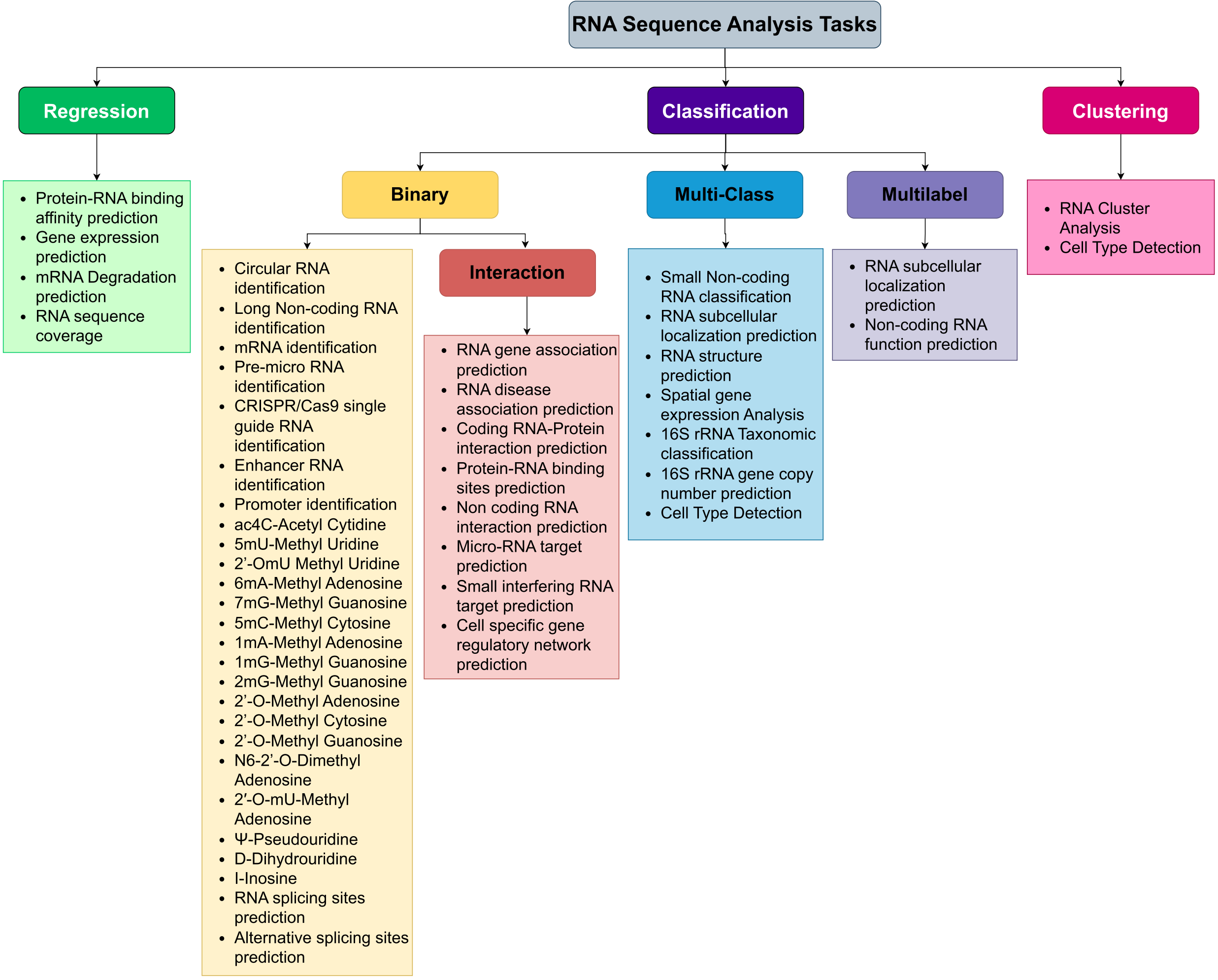
RNA Sequence Analysis Tasks
Regression
Classification
Clustering
Protein-RNA binding affinity prediction
Gene expression prediction
mRNA Degradation prediction
RNA sequence coverage
Binary
Multi-Class
Multilabel
RNA Cluster Analysis
Cell Type Detection
Circular RNA identification
Long Non-coding RNA identification
mRNA identification
Pre-micro RNA identification
CRISPR/Cas9 single guide RNA identification
Enhancer RNA identification
Promoter identification
ac4C-Acetyl Cytidine
5mU-Methyl Uridine
2'-OmU Methyl Uridine
6mA-Methyl Adenosine
7mG-Methyl Guanosine
5mC-Methyl Cytosine
1mA-Methyl Adenosine
1mG-Methyl Guanosine
2mG-Methyl Guanosine
2'-O-Methyl Adenosine
2'-O-Methyl Cytosine
2'-O-Methyl Guanosine
N6-2'-O-Dimethyl Adenosine
2'-O-mU-Methyl Adenosine
Ψ-Pseudouridine
D-Dihydrouridine
I-Inosine
RNA splicing sites prediction
Alternative splicing sites prediction
Interaction
RNA gene association prediction
RNA disease association prediction
Coding RNA-Protein interaction prediction
Protein-RNA binding sites prediction
Non coding RNA interaction prediction
Micro-RNA target prediction
Small interfering RNA target prediction
Cell specific gene regulatory network prediction
Small Non-coding RNA classification
RNA subcellular localization prediction
RNA structure prediction
Spatial gene expression Analysis
16S rRNA Taxonomic classification
16S rRNA gene copy number prediction
Cell Type Detection
RNA subcellular localization prediction
Non-coding RNA function prediction

Fig. 4. Methodical Classification of 47 RNA Sequence Analysis Tasks on the Basis of Their Nature from The Lens of Computer Scientists.

$$f(x) = (h_1(x), h_2(x), ..., h_K(x)) \tag{3}$$

Moreover, in regression, researchers aim to forecast a continuous outcome variable. When provided with a dataset containing features $X \in \mathbb{R}^{nxd}$, labels $y \in \mathbb{R}$, and a training dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i \in X$ and $y_i \in Y$, according to equation (4), the objective is to develop a function $f : X \rightarrow \mathbb{R}$ that predicts continuous outputs by utilizing the hypothesis function $h(x)$ mainly learned from training data.

$$f(x) = h(x) \tag{4}$$

In clustering, the aim is to categorize similar data points into corresponding clusters. When presented with a dataset comprising data points $X = x_1, x_2, \ldots, x_n$, where each $x_i \in \mathbb{R}^d$, the goal is to establish a partition of the data into clusters $C = C_1, C_2, \ldots, C_K$. According to equation (5), this partitioning is executed based on a distance metric $d(x, \mu_c)$ that measures the distance between a data point x and the centroid μ_c of cluster c.

$$f(x) = \text{argmin}_c d(x, \mu_c) \tag{5}$$

5. RNA sequence analysis databases

This section highlights critical role of public databases in facilitating the development of AI-driven RNA-sequence analysis applications. Biological databases house a wealth of RNA information that serves as the foundation for development of benchmark datasets. A comprehensive understanding about contents of RNA molecule related databases may enable researchers to perform large scale AI-driven RNA sequence analysis. Deep understanding of public databases can empower researchers to develop different RNA sequence analysis tasks and distinct species related benchmark datasets. Distinct species datasets of a RNA sequence analysis task is important for conducting cross-species experiments using AI pipelines. This comparative analysis is essential for gaining a broader understanding of biological processes at a more fundamental level.

The ever-expanding nature of public databases facilitates researchers by providing access to increasingly larger data. As new sequences are added, researchers can use expanded data to benchmark the performance of existing AI-driven RNA sequence analysis pipelines. This benchmarking process offers valuable insights into how well current predictors perform with large data and helps researchers in identifying potential areas for improvement and development of more robust AI applications. Moreover, researchers can utilize these databases to acquire large volumes of RNA sequence data. This data can then be used to train word embedding methods and large language models in an unsupervised manner. The pre-trained models can be utilized to develop diverse types of RNA sequence analysis applications. Specifically, this section provides an extensive overview of databases that have been used to create benchmark datasets for 47 distinct RNA sequence analysis tasks. A comprehensive review of 172 research articles focused on AI-driven RNA sequence analysis tasks reveals that a total of 90 distinct databases have been utilized to develop 47 different RNA sequence analysis tasks related benchmark datasets.

From 90 databases, **64** databases are publicly accessible, while the remaining **26** are either inaccessible or no longer exist. To aid research community, Table 1 provides a detailed summary of accessible databases in terms of their release year, types of inherent RNA data, species and organisms details, raw sequence statistics, and supported data formats. A thorough analysis of Table 1 reveals that out of 64 accessible databases, 6 databases encompass data related to three different types of molecules namely DNA, RNA, and Proteins. Similarly, 2 databases contain data related to Proteins and RNA molecules. Among all accessible databases, 56 databases have dedicated information related to only RNA molecule. Specifically, miRNA sequences are available in 15 different databases namely m6A-Atlas v2 [41], MNDR3.0 [42], CircBank [43], RMBase2.0 [44], miRmine [45], dbDEMC [46], miRCancer [47], Encori [48], miR2Disease [49], HMDD [50], TarBase [51], NPInter V4.0 [52], miRBase [53], ENCODE3 [54], FANTOM5 [55]. Furthermore, long non-coding RNA molecule related diverse types of information is available in 11 databases including m6A-Atlas v2 [41], MNDR3.0 [42], cantataDB 2.0 [56], LncRNADisease v2.0 [57], NONCODEV5 [58], LNCipedia [59], lncRNADisease [57], Encori [48], PLncDB 2.0, NPInter V4.0 [52], and FANTOM5 [55]. Additionally, 11 databases namely Circad [60], MNDR3.0 [42], CSCD [61], LncRNADisease v2.0 [57], CircRNADisease [62], CircBank [43], circRNADb [63], lncRNADisease [57], CircBase [64], NPInter V4.0 [52], and FANTOM5 [55] databases provide circular RNA sequences. Similarly, 6 databases (m6A-Atlas v2 [41], Encori [48], CTD [65], MNDR3.0 [42], NPInter V4.0 [52], FANTOM5 [55]) offer mRNA and snoRNA sequences. Also, 6 databases including NPInter V4.0 [52], FANTOM5 [55], MNDR3.0 [42], GtRNAdb [66], piRBase [67], and ENCODE3 [54] contain information about four distinct RNA molecules namely snRNA, tRNA, piRNA, and siRNA.

Since word embedding methods and large language models are trained on large raw sequences data in an unsupervised manner to generate better representations, these databases can be utilized to efficiently train these language models. To assist researchers and practitioners, we categorized these databases based on the volume of raw sequences into three categories: 1) low sequence facilitators, 2) medium sequence facilitators, and 3) high sequence facilitators. Specifically, 38 low sequences facilitator databases provide 100,000 RNA sequences each and these database include SPENCER [68], m6A-Atlas v2 [41], RNALocate v2.0 [69], Lnc2Cancer v3.0 [70], GENCODE Release 43 [71], circR2Cancer [72], Circad [60], EVLncRNAs 2.0 [73], PanglaoDB [74], GENCODE.v28 [75], GENCODE v18 [76], LncRNADisease v2.0 [57], CircRNADisease [62], RNALocate [69], miRmine [45], CircInteractome [77], ATtRACT [78], HMDAD [79], dbDEMC [46], circRNADb [63], NDB [80], lncRNADisease [57], miRCancer [47], Encori [48], CircBase [64], GENCODE v.17 [81], RNAcentral [82], miR2Disease [49], HMDD [50], TarBase [51], Gencode [83], NCBI [84], miRBase [53], ENCODE3 [54], ENCODE [85], ENSEMBL [86], OMIM [87]. A total of 11 public databases fall into the "medium sequence facilitators" category and each database contain approximately 1 million sequences. Medium sequences facilitator databases are MNDR3.0 [42], cantataDB 2.0 [56], EuRBPDB [88], CSCD [61], CircBank [43], NONCODEV5 [58], lncRNA2Target [89], LNCipedia [59], EPDnew [90], HGMD [91], GtRNAdb [66]. Whereas, a total of 18 high sequence facilitator databases are piRBase [67], RefSeq [92], lncRNASNP2 [93], bpRNA [94], RMBase2.0 [44], RMBase [95], DisGeNET [96], RefSeq (version 60) [97], PLncDB 2.0 [98], ClinVar [99], dbGap [100], NPInter V4.0 [52], Rfam [101], CTD [65], GEO [102], KEGG [103], EMBL-EBI [104], FANTOM5 [55], and doRiNA [105]. These databases predominantly house RNA sequences from a diverse array of species, including humans, mice, plants, bacteria, and fungi.

An extensive analysis of different databases reveals that about 9 databases, such as SPENCER [68], CSCD [61], GENCODE.v28 [75], miRmine [45], CircInteractome [77], circRNADb [63], NDB [80], LNCipedia [59], and GENCODE.v17 [81], focus on Homo sapiens RNA sequences, miR2Disease [49], and HGMD [91]. Whereas, OMIM [87] databases provide both homo sapiens and animal RNA sequences. Additionally, 6 databases namely GENCODE Release 43 [71], PanglaoDB [74], Gencode [83], NCBI [84], and ENCODE [85] facilitate Homo sapiens, animals and mus musculus RNA sequence. On the other hand, Circad [60] offers RNA sequences of Homo sapiens, Mus musculus, and Rattus rattus. Sequences from other organisms, such as eukaryotes, invertebrates, fungi, and various microorganisms, are also well-represented in this database. Databases can be categorized into three distinct groups based on the variety of species they accommodate; 1) Broad coverage databases, 2) Moderate coverage databases, 3) Limited coverage databases. A total of 33 limited coverage databases facilitate RNA sequences of upto 20 different species including SPENCER [68], GENCODE Release 43 [71], Circad [60], PanglaoDB [74], CSCD [61], GENCODE.v28 [75], LncRNADisease v2.0 [57], CircBank [43], RMBase2.0 [44], miRmine [45], CircInteractome [77], NONCODEV5 [58], circRNADb [63], DisGeNET [96], NDB [80], LNCipedia [59], doRiNA [105], lncRNADisease [57], CircBase [64], EPDnew [90], ClinVar [99], GENCODE.v17 [81], miR2Disease [49], HGMD [91], Gencode [83], NCBI [84], NPInter V4.0 [52], ENCODE3 [54], ENCODE [85], ENSEMBL [86], KEGG [103], OMIM [87], and CircRNADisease [62].

A total of 9 moderate coverage databases encompass data related to 80 species. Moderate coverage databases are GEO [102], Encori [48], TarBase [51], ATract [78], cantataDB 2.0 [56], m6A-Atlas v2 [41], piRBase [67], RMBase [95], RNALocate [69]. Whereas, a total of 22 broad coverage databases contain data of more than 80 different species. These databases are PLncDB 2.0 [98], RNALocate v2.0 [69], MNDR3.0 [42], EVLncRNAs 2.0 [73], EuRBPDB [88], GtRNAdb [66], RefSeq (version 90) [106], GENCODE.vM18 [76], lncRNASNP2 [93], bpRNA [94], HMDAD [79], dbDEMC [46], RefSeq (version 60) [97], miRCancer [47], EMBL-EBI [104], RNAcentral [82], HMDD [50], dbGap [100], miRBase [53], Rfam [101], CTD [65], and FANTOM5 [55]. For example, pirbase [67] offers RNA sequences of 44 species, EuRBPDB [88] houses sequences of 162 species, EVLncRNAs 2.0 [73] has RNA sequence data of 124 species, RNALocate [69] contains RNA sequences of 104 species, m6A-Atlas v2 [41] houses RNA sequences of 42 species, and MNDR [42] has RNA sequence data of 117 species.

Furthermore, a deep analysis of Table 1 reveals that in total 30 distinct data formats have been used to store data in 64 distinct databases. These data formats include TXT, FASTA, VCF, XLSX, BED, JSON, PDF, TSV, CSV, GTF, GFF, XML, BAM, BigWig, MySQL, KDML, DAT, FPS, BB, and IDX, etc., TXT and FASTA formats are universally accepted by almost all RNA sequence analysis programs. Each entry in these formats contains at least two lines: header includes accession number, species name, or identification details, while subsequent lines contain nucleotide sequences. CSV and TSV are text-based formats in which values in rows are separated by commas or tabs, respectively. In both formats, first row specifies headers that define column names ("Sequence ID", "Sequence Name", "Type", "Function"), and subsequent rows represent data. VCF format also specifies headers in first row and is specifically used to store genetic variation data including single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variants. Additionally, XLSX formats represent complex datasets containing information computed with various formulas across multiple columns, whereas EMBL format includes structured sections for sequence data, feature annotations, organism information, references, and other details. An extensive analysis of Table 1 reveals that most widely used data formats in RNA sequence analysis are FASTA, TXT, CSV, XLSX, and EMBL.

From 64 publicly available databases, RNA categorization and identification tasks related data is available in 13 different databases namely SPENCER [68], cantataDB 2.0 [56], piRBase [67], EVLncRNAs 2.0 [73], CSCD [61], RefSeq (version 90) [106], LNCipedia [59], RefSeq (version 60) [97], GtRNAdb [66], Rfam [101], circRNADb [63], EPDnew [90], PLncDB 2.0 [98]. Similarly, different RNA interaction and binding sites tasks including RNA-protein binding sites prediction, coding RNA–protein interaction prediction, and RNA-protein binding affinity prediction related data is available in 10 databases namely CircBank [43], ClinVar [99], GENCODE Release 43 [71], ENCODE3 [54], EuRBPDB [88], CircInteractome [77], ATtRACT [78], ENCODE [85], NDB [80], doRiNA [105]. In addition, RNA-disease association prediction task related data is available in 12 databases namely miR2Disease [49], HMDD [50], HMDAD [79], dbDEMC [46], Circad [60], MNDR3.0 [42], lncRNADisease [57], NPInter V4.0 [52], CTD [65], miRCancer [47], LncRNADisease v2.0 [57], and CircRNADisease [62]. RNA modification prediction tasks related data is available in RMBase [95], m6Atlas [41], and RMBase2.0 [44]. Furthermore, GENCODE [83] provides RNA sequences for RNA categorization, identification and interaction tasks. RNA sequences data related to sub-cellular localization prediction, gene analysis, RNA single cell analysis, RNA special characteristics analysis, RNA categorization, association and interaction tasks are available in remaining databases namely NCBI [84], dbGap [100], RNAcentral [82], OMIM [87], ENSEMBL [86], GEO [102], TarBase [51], HGMD [91], RNALocate [69], PanglaoDB [74], KEGG [103], EMBL-EBI [104], FANTOM5 [55].

6. RNA sequence analysis benchmark datasets

This section offers a comprehensive overview of public and in-house datasets employed to develop AI applications for 47 different RNA sequence analysis tasks. Publicly available datasets are accessible to broader research community and are commonly used to develop AI-based predictive pipelines. These datasets enhance accessibility, reusability, and encourages collaboration and knowledge sharing within scientific community. In contrast, in-house datasets are developed within specific labs or institutions. These datasets often contain sensitive data tailored to specific research goals. Their proprietary nature limits broader access, reproducibility, and applicability of findings.

A comprehensive review of 172 research articles reveals that a total of 310 unique datasets have been utilized in the development of AI-driven applications for 47 distinct RNA sequence analysis tasks. These datasets have either been created by the authors or sourced from existing studies. Among these 310 benchmark datasets, 236 are publicly available datasets, whereas, 74 are in-house datasets. Table 2 facilitates the distribution of these datasets and their use in the validation of AI-driven predictive models using three representation learning approaches: word embeddings, large language models, and domain specific methods.

Distribution of public and in-house datasets for 47 RNA sequence analysis tasks is clearly explained using parentheses. For each task, first number represents count of public datasets, and second indicates count of in-house datasets utilized for that particular task. Thereby, distribution of datasets for 47 different tasks is as follows: RNA Cluster Analysis (2, 0), mRNA Identification (7, 0), Small Non-coding RNA Classification (3, 1), Circular RNA Identification (3, 0), Long Non-coding RNA Identification (9, 5), Pre-micro RNA Identification (0, 2), CRISPR/Cas9 single guide RNA Identification (9, 0), Enhancer Identification (1, 0), Promoter Identification (2, 0), RNA-Gene Association Prediction (0, 2), RNA-Disease Association Prediction (57, 17), Protein-RNA Interaction Prediction (13, 4), Protein-RNA Binding Sites Prediction (20, 3), Protein-RNA binding affinity prediction (1, 0), non-coding RNA Interaction Prediction (6, 2), RNA Sub-cellular Localization Prediction (1, 2), ac4C-Acetyl Cytidine Modification Prediction (1, 0), 5mU-Methyl Uridine Modification Prediction (4, 0), 2'-OmU Methyl Uridine Modification Prediction (1, 0), 6mA-Methyl Adenosine Modification Prediction (14, 3), 7mG-Methyl Guanosine Modification Prediction (1, 5), 5mC-Methyl Cytosine Modification Prediction (2, 0), Methylation Modification Prediction (8, 4), RNA-Splicing Sites Prediction (5, 0), Alternative Splicing Prediction (0, 1), RNA Functions Prediction (2, 10), RNA Structure Prediction (15, 6), Spatial Gene Expression Analysis (7, 0), Gene Expression Prediction (1, 3), Cell-Specific Gene Regulatory Networks Prediction (7, 0), 16S rRNA Taxonomic Classification (1, 1), 16S rRNA Gene Copy Number Prediction (1, 0), Micro RNA Target Prediction (6, 1), Small Interfering RNA Target Prediction (3, 3), mRNA Degradation Prediction (2, 0), RNA-Seq Coverage Prediction (1, 0), and Cell-type Detection (19, 0).

Table 1
A Summary of Publicly Accessible Biological Databases, their Inherent Data Types, Species Diversity, and Statistics of Raw Sequences Related to Different Genomic and Proteomic Data.

Database Name	Release Date	Types of Data	Species	Organism	Sequences Statistics	Data Format
SPENCE	2022	ncRNAs	Homo sapiens	_	1700 patient samples, 6800 ncRNA transcripts, 29526 ncRNA-encoded peptides from 15 cancer types, 8,060 tumor-specific peptides, 4497 peptides with potential immunogenicity	.txt
m6A-Atlas v2	2022	mRNAs, lncRNAs, miRNAs	42 species	_	2813 samples, 16,868,200 m6A peaks, 797,091 m6A sites	.txt
RNALocate v2.0	2021	RNA	104 species	_	Number of entries: 213,260, Number of subcellular localization: 171	.txt
GENCODE Release 43	2021	ncRNAs	Animal, Homo sapiens, Mus musculus	_	63,086 genes, 19,411 protein-coding genes, 20,310 lncRNA genes, 7565 ncRNA genes, 14,716 pseudogenes, 254,070 transcripts, 89,581 Protein-coding transcripts, 21,774 Nonsense mediated decay transcripts, 59,927 Long non-coding RNA loci transcripts, 65,650 Total No of distinct translations, 13,620 Genes that have more than one distinct translations	_
Circad	2020	circRNAs	Homo sapiens, Mus musculus, Rattus rattus	_	Number of disease related circRNA: 1388, Number of diseases: 150, No. of circRNAs in: Homo sapiens = 1270, Mus musculus = 66, Rattus rattus = 42	_
MNDR3.0	2020	lncRNAs, piRNAs, circRNAs, miRNAs, tRNAs, snoRNAs	117 species	_	Experimental data: 343,273 All RNA-disease entries, Predicted data: 237,329 entries miRNA-disease information, 348,176 entries lncRNA-disease information, 362,454 entries circRNA-disease information, 48,779 entriespiRNA-disease information	.txt
cantataDB 2.0	2020	lncRNAs	39 species	_	239,631 lncRNAs	FASTA, .gtf
EVLncRNAs 2.0	2020	RNA	124 species	_	4010 lncRNAs, 1082 Diseases, 11,257 lncRNA-disease associations, 1665 Function Annotations (excluding interactions), 6244 Interactions, 37 Peptide-coding, 8 Structure, 33 Exosomal, 188 CircRNAs, 1079 Drug/chemoresistance/stress	.xlsx
piRBase	2019	piRNAs	44 species	_	181 million unique piRNA sequences	FASTA, .bed, .csv, .tsv, .json, .txt
EuRBPDB	2019	RBPs	162 species	_	315,222 RBPs	.txt, .fa
PanglaoDB	2019	RNA	Animal, Homo sapiens, Mus musculus	_	Mus musculus: 1063 samples, 184 tissues, 4,459,768 cells, 8,651 clusters, Homo sapiens: 305 samples, 74 tissues, 1,126,580 cells, 1,248 clusters	.tar
CSCD	2018	circRNAs	Homo sapiens	_	samples &amp;gt;1000, including ~800 tissue samples and ~300 cell line samples, 1013461 cancer-specific circRNAs, 1533704 circRNAs normal samples and 354422 circRNAs from both cancer and normal samples	.txt
RefSeq (version 90)	2018	DNA, RNA, Proteins	_	_	23838836 entries	.csv, .json

Table 1 (*continued*)

Database Name	Release Date	Types of Data	Species	Organism	Sequences Statistics	Data Format
GENCODE.v28	2018	RNA, Proteins	Homo sapiens	_	58,381 Total No of Genes, 19,901 Protein-coding genes, 15,779 Long non-coding RNA genes, 7569 Small non-coding RNA genes, 147723 Pseudogenes, 10693- processed pseudogenes, 3519 - unprocessed pseudogenes, 218 - unitary pseudogenes, 38 - polymorphic pseudogenes, 18 - pseudogenes, 408 Immunoglobulin/T-cell receptor gene segments - protein coding segments, 237 - pseudogenes, 203,835 Total No of Transcripts, 82,335 Protein-coding transcripts, 56541 - full length protein-coding, 25,794 - partial length protein-coding, 14,889 Nonsense mediated decay transcripts, 28,468 Long non-coding RNA loci transcripts, 61,132 Total No of distinct translations, 13,641 Genes that have more than one distinct translations	.gtf, .gff, FASTA, .bed, .json, .tsv
GENCODE.vM18	2018	RNA, Proteins	_	Mouse	54,146 Total No of Genes, 21,978 Protein-coding genes, 12,726 Long non-coding RNA genes, 6108 Small non-coding RNA genes, 12,838 Pseudogenes, 9612 - processed pseudogenes, 2842 - unprocessed pseudogenes, 37 - unitary pseudogenes, 79 - polymorphic pseudogenes, 65 - pseudogenes, 494 Immunoglobulin/T-cell receptor gene segments - protein coding segments, 203 - pseudogenes, 136,535 Total No of Transcripts, 57,388 Protein-coding transcripts, 44,118 - full length protein-coding, 13270 - partial length protein-coding, 6679 Nonsense mediated decay transcripts, 17,855 Long non-coding RNA loci transcripts, 44,166 Total No of distinct translations, 10,491 Genes that have more than one distinct translations	.gtf, .gff, FASTA, .bed, .json, .tsv
lncRNASNP2	2018	RNA	_	Human, Mouse	10,205,295 SNPs in 141,353 human lncRNA transcripts of 90,062 lncRNA genes, 859,534 Cosmic Noncoding Variations and 315,234 TCGA cancer mutations	.xlsx
LncRNADisease v2.0	2018	lncRNAs, circRNAs	Animal, Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus	_	19,166 lncRNAs, 823 circRNAs, 529 diseases, 205,959 lncRNA-disease associations, 1004 circRNA-disease associations	.xlsx
CircRNADisease	2018	circRNAs	12 species	Human, Chicken, Cow, Mouse, Rat	4246 circRNAs, 330 DO diseases, 6998 circRNA-diseases, 7,159,865 mutation-circRNAs	.txt, .xlsx
CircBank	2018	circRNAs, miRNAs	Plants	Human, Mouse, Fly, Worm, Yeast	more than 140,000 human annotated circRNAs, 1439 associations between 1135 circRNAs and 82 cancers	.bed, .txt, .xlsx
bpRNA	2018	RNA	_	_	708,144 hairpins, 517,672 bulges, 317,046 multi loops, 538,670 internal loops, 57,686 pseudoknots, 2,075,928 stems, 229,468 unpaired regions, 1,019,586 segments	FASTA, .pdf, .jpg

(*continued on next page*)

Table 1 (*continued*)

Database Name	Release Date	Types of Data	Species	Organism	Sequences Statistics	Data Format
RNALocate	2017	RNA	65 species	_	42,190 Number of entries, 41 Number of subcellular localization, 23,100 RNAs	.txt, .xlsx, FASTA
RMBase2.0	2017	miRNAs	Homo sapiens, Mus musculus, Rhesus, Rattus, A.thaliana, S.cerevisiae, P.aeruginosa, Escherichia coli, S.pombe	Chimpanzee, Pig, Zebrafish, Fly	5411 m1A, 988 m5C, 1373355 m6A, 5096 2'-O-Me, 9570 pseudoU, 2824 others	.txt
miRmine	2016	miRNAs	Homo sapiens	_	2822 cell lines, 2822 tissues	excel, .csv, .pdf
CircInteractome	2016	RNA	Homo sapiens	_	no of entries: 65535	.xlsx
ATract	2016	RBPs	38 species	_	370 RBPs and 1583 RBP consensus binding motifs	.txt, .csv, .tsv
HMDAD	2016	DNA, RNA, Proteins	_	_	483 disease-microbe entries which include 39 diseases and 292 microbes	.txt
dbDEMC	2016	miRNAs	_	Human, Mouse, Rat	3268 miRNAs, 40 cancer types, 149 cancer subtypes, 403 datasets, 807 experiments, 46388 samples	.txt
NONCODEV5	2016	lncRNAs	Arabidopsis, Caenorhabditis elegans	15 organisms	354,855 lncRNA genes, 548,640 lncRNA transcripts	FASTA
RMBase	2016	RNA	62 species	_	1,074,100 RNA modification, 73 types of RNA	.tar.gz
circRNADb	2015	circRNAs	Homo sapiens	_	32,914 annotated exonic circRNAs	FASTA, .tsv
DisGeNET	2015	DNA, RNA, Protein	Animals	Human	1,134,942 GDAs between 21,671 Genes, 30,170 diseases, and traits, 369,554 VDAs between 194,515 variants and 14,155 diseases and traits	.txt, RDF, SQL Dump
NDB	2014	RNA, DNA, Protein	Homo sapiens	_	17894 3D structures containing nucleic acids	.csv, .json
LNCipedia	2013	lncRNAs	Homo sapiens	_	127,802 transcripts and 56,946 genes	.bed, FASTA, .gff, .gtf
RefSeq (version 60)	2013	DNA, RNA, Proteins	_	_	4243209 entries	.csv, .json
doRiNA	2013	RNA	Homo sapiens, Mus Musculus, Caenorhabditis elegans, Drosophila melanogaster	_	_	.bed
lncRNADisease	2013	lncRNAs, circRNAs	Animal, Homo sapiens, Mus musculus, Rattus norvegicus, Oryctolagus cuniculus	_	6,066 lncRNAs, 10,732 circRNAs, 566 diseases, 13,191 lncRNA-disease associations, 12,249 circRNA-disease associations	.tsv, .xlsx
miRCancer	2013	miRNAs	_	34 organisms	57984 miRNAs, 196 cancers, 9080 miR-Cancers	.txt

A comprehensive analysis of Table 2 demonstrates that a total of 130 public and 45 in-house datasets are used to develop word embeddings and language model-based predictive pipelines for 8 RNA sequence analysis tasks. These tasks include long non-coding RNA identification, RNA-disease association prediction, protein-RNA binding sites prediction, non-coding RNA interaction, RNA sub-cellular localization prediction, 6mA-methyl adenosine modification prediction, 7mG-methyl guanosine modification prediction, methylation modification prediction, and RNA structure prediction. Notably, only 6 public datasets have commonly used by both kinds of predictive pipelines for 2 tasks namely RNA-disease association prediction and non coding RNA interaction prediction. Also,

Table 1 (*continued*)

Database Name	Release Date	Types of Data	Species	Organism	Sequences Statistics	Data Format
Encori	2013	mRNAs, miRNAs, ceRNAs, lncRNAs	23 species	Human, Mouse	2,725 CLIP-seq datasets, 100 Degradome-seq datasets, 59 RNA-RNA interactome datasets, RNA-seq data: more than 10,800 samples from 32 cancer types, miRNA-seq data: 10,500 samples from 32 cancer types, Disease data: 1,800,000 mutations from 531 disease types, miRNA-ncRNA(CLIP): 460,000 interactions, miRNA-mRNA(CLIP): 1,200,000 interactions, RBP-mRNA:1,290,000 interactions, RBP-ncRNA: 1,600,000 interactions, RNA-RNA: gt;3,700,000 interactions, miRNA-ncRNA(degradome): 32,000 interactions, miRNA-mRNA(degradome): 459,000 interactions, ceRNA: 2,900,000 pairs, function annotation: gt;34,000 functional terms from 21 categories, Pan-Cancer: Differential Expression, Survival Analysis, CoExpression	.txt, .xlsx
CircBase	2013	circRNAs	Homo sapiens, Mus musculus, Caenorhabditis elegans, Latimeria chalumnae, Latimeria menadoensis	_	Human: 8483 circRNAs, Caenorhabditis elegans: 2399 circRNAs, Drosophila melanogaster: 5795 circRNAs	FASTA, .txt, .xlsx, .bed
EPDnew	2013	RNA	Animals, Plants, Fungi, Invertebrates	_	Animal: 13,1870 promoters, Plants: 39,784 promoters, Fungi: 9919 promoters, Invertebrates: 5597 promoters	.bed, .dat, .fps, .bb, .idx, FASTA
PLncDB 2.0	2013	lncRNAs	80 species	_	1246372 lncRNAs, 13834 RNA-Seq datasets	.fa, .txt, .gff3
ClinVar	2013	DNA, RNA, Protein	Animals	Human	4,391,341 Records, 92,225 Total Genes	.xml, .tsv, VCF
GENCODE.v17	2012	RNA, Proteins	Homo sapiens	_	57,281 Total No of Genes, 20,330 Protein-coding genes, 13,333 Long non-coding RNA genes, 9078 Small non-coding RNA genes, 14154 Pseudogenes, 29 polymorphic pseudogenes, 13,897 pseudogenes, Immunoglobulin/T-cell receptor gene segments; 386 - protein coding segments, 228 - pseudogenes, 194,871 Total No of Transcripts, 81,565 Protein-coding transcripts, 56,950 full length protein-coding, 24,615 partial length protein-coding, 12,913 Nonsense mediated decay transcripts, 22,631 Long non-coding RNA loci transcripts, 61,102 Total No of distinct translations, 13,569 Genes that have more than one distinct translations	.gtf, .gff, FASTA, .bed, .json, .tsv
RNAcentral	2011	ncRNAs	_	_	96,670 sequences	.txt, FASTA, .json
miR2Disease	2009	miRNAs	Animal, Homo sapiens	_	349 miRNAs, 163 diseases, 3273 entries	.txt
HMDD	2008	miRNAs	_	Human	53,530 miRNA-disease association entries which include 1,817 human miRNA genes, 79 virus-derived miRNAs, 2,360 diseases from 37,090 papers	.txt, .xlsx

(*continued on next page*)

Table 1 (*continued*)

Database Name	Release Date	Types of Data	Species	Organism	Sequences Statistics	Data Format
dbGap	2007	RNA	_	_	12815 phenotype datasets, 430727 datasets, 4.64 million samples	.xml, .csv
HGMD	2007	DNA, RNA, Protein	Animal, Homo sapiens	_	Mutation totals: (public release for academic/non-profits only): 291,339 or HGMD Professional release 2023.4: 504,008	.txt
TarBase	2006	miRNAs	24 species	_	5,878,998 interactions, 103 tissues, 3300 unique miRNAs, 57 cell types	.tsv.gz
Gencode	2006	DNA, RNA, Protein	Animals, Homo sapiens, Mus musculus	_	Homo sapiens: Total Genes = 63086, Total Transcripts = 254070, Total distinct Translations = 65650, Mus musculus: Total Genes = 57132, Total Transcripts = 149138, Total distinct Translations = 44819	.txt
NCBI	2005	DNA, RNA, Protein	Animals, Homo sapiens, Mus musculus	_	35,608 CCDS IDs that correspond to 19,107 Genes, with 48,062 Protein Sequences	FASTA
GtRNAdb	2005	tRNAs	740 species	_	Eukaryota: 599 Number of Genomes, 74,048 Number of tRNA Genes, Archaea: 220 Number of Genomes, 10,476 Number of tRNA Genes, Bacteria: 4,038 Number of Genomes, 242,068 Number of tRNA Genes	.fa, .bed, .txt, .gtf, .tsv.gz
NPInter V4.0	2005	lncRNAs, miRNAs, circRNAs, snoRNAs, snRNAs	Homo sapiens, Mus musculus, Saccharomyces cerevisiae, Agrobacterium tumefaciens, Escherichia coli, Caenorhabditis elegans, Drosophila melanogaster, Kaposi sarcoma-associated herpesvirus	_	658171 lncRNA interactions, 488025 miRNA interactions, 61700 snoRNA interactions, 12789 snRNA interactions, 335 circRNA interactions, 488315 RNA-Protein interactions	.txt, .xlsx, .tsv
miRBase	2004	miRNAs	_	271 organisms	38 589 hairpin precursors and 48 860 mature microRNAs	.gff3, .dat, FASTA
Rfam	2003	RNA	_	_	4170 families, 3,026,773 regions, ENA 133/134 Rfamseq	.txt, .fa, .tar.gz
CTD	2003	mRNAs	_	632 organisms	2,915,515 Chemical–gene interactions, 406,571 Phenotype–based interactions, 32,694,093 Gene–disease associations, 3,489,469 Chemical–disease associations, 6,577,078 Chemical–GO associations, 1,570,026 Chemical–pathway associations, 305,622 Disease–pathway associations, 1,358,371 Gene–gene interactions, 39,776,068 Gene–GO annotations, 135,792 Gene–pathway annotations, 3,133,281 GO–disease associations, 17,667 Chemicals with curated data, 7,285 Diseases with curated data, 55,128 Genes with curated data	.csv, .tsv, .xml
ENCODE3	2003	scRNAs, siRNAs, miRNAs, small RNAs	Homo sapiens, Mus Musculus, Caenorhabditis elegans, Drosophila melanogaster	_	9000 high-throughput sequencing libraries from assays	.txt, .hic, .fastq, .bed

Table 1 (*continued*)

Database Name	Release Date	Types of Data	Species	Organism	Sequences Statistics	Data Format
ENCODE	2003	DNA, RNA, Protein	Animals, Homo sapiens, Mus musculus	_	17238 sequences	FASTA, BAM, BigWig, .bed, VCF
FANTOM5	2002	lncRNAs, miRNAs, circRNAs, snoRNAs, snRNAs	_	Human, Mouse, Dog, Chicken, Rat, Rhesus Monkey	_	.bed, .txt, .xlsx
GEO	2000	DNA, RNA, Protein	21 species	_	Samples = 7209691	SOFT, MINiML, .txt
ENSEMBL	1999	DNA, RNA, Protein	Animals, Homo sapiens, Mus musculus, Danio rerio, Sus scrofa	_	44,048 Genomes, 1014 Ensembl Fungi Genomes, 78 Ensembl Metazoa Genomes for invertebrate species, 236 Genomes for vertebrate Species, 67 Ensembl Plants Genomes, 237 Ensembl Protists Genomes	FASTA, .gtf, .gff, MySQL Dump
KEGG	1995	DNA, RNA, Protein	Animals, Plants, Fungi, Protists, Bacteria, Archaea	14 organisms	Genes: 53,674,741, Addendum Proteins: 4,181, Viral Genes: 688,823, Viral mature Peptides: 377	KGML, FASTA, .txt
EMBL-EBI	1994	DNA, RNA, Protein	_	_	_	.xml, FASTA, .txt, .tsv, .json
OMIM	1960	DNA, RNA, Protein	Animals	Homo sapiens	17,290 Gene descriptions, 18 Gene and Phenotypes combined, 6859 Phenotype description molecular basis known, 1502 Phenotype description molecular basis unknown, 1736 mainly Phenotypes with suspected mendelian basis	.txt

5 public datasets for RNA-disease association prediction and only 1 dataset for non-coding RNA interaction prediction are commonly used by both kinds of predictive pipelines.

Additionally, 158 public and 50 in-house datasets are leveraged to develop word embedding and domain-specific representation learning based predictive pipelines across 13 RNA sequence analysis tasks encompassing circular RNA identification, long non-coding RNA identification, RNA-disease association prediction, coding RNA-protein interaction prediction, protein-RNA binding sites prediction, non-coding RNA interaction, 5mU-methyl uridine modification prediction, 6mA-methyl adenosine modification prediction, 7mG-methyl guanosine modification prediction, 5mC-methyl cytosine modification prediction, methylation modification prediction, RNA structure prediction, and microRNA target prediction. However, only 5 public datasets are commonly employed by both kinds of predictive pipelines for 2 specific tasks namely coding RNA-Protein interaction prediction and RNA-protein binding sites prediction. Also, 4 public datasets for coding RNA-protein interaction prediction and 1 public dataset for protein-RNA binding sites prediction are commonly used by both kinds of predictive pipelines.

Furthermore, an in-depth analysis of Table 2 reveals that 151 public and 55 in-house datasets are employed for developing language models and domain-specific approaches based predictive pipelines for 10 RNA sequences analysis tasks namely long non-coding RNA identification, RNA-disease association prediction, protein-RNA binding sites prediction, non-coding RNA interaction, 6mA-methyl adenosine modification prediction, 7mG-methyl guanosine modification prediction, methylation modification prediction, RNA function prediction, RNA structure prediction, and cell-type detection. Notably, only 19 public datasets are commonly utilized by both language models and domain-specific representation learning methods based predictive pipelines for 4 RNA sequence analysis tasks namely CRISPR/Cas9 single guide RNA identification, RNA-disease association prediction, 6mA-methyl adenosine modification prediction, and 7mG-methyl guanosine modification prediction. Specifically, 6 public datasets for CRISPR/CAS9 single guide RNA identification, 1 public for RNA-disease association prediction, 11 public for 6mA-methyl adenosine modification prediction, and 1 public datasets for 7mG-methyl guanosine modification prediction are commonly used by both kinds of predictive pipelines.

While all three distinct types of representation learning-based predictive pipelines are employed across 6 different RNA sequence analysis tasks including long non-coding RNA identification, RNA-disease association, protein-RNA binding sites prediction, non-coding RNA interaction prediction, 6mA-methyl adenine modification prediction, and RNA structure prediction. Surprisingly, not a single dataset is commonly employed by all three kinds of predictive pipelines as they are evaluated on separate datasets for each task. This trend underscores that researchers have predominantly focused on developing new datasets for each type of predictive pipeline, rather than utilizing existing datasets. Thus, RNA sequence analysis domain lacks in rigorous fair performance comparison of predictive pipelines.

Table 2
Overview of 236 Public and 74 In-house Datasets used Across 37 Different RNA Sequence Analysis Tasks.

Task Name	Datasets used in Language Models		Datasets used in word embeddings		Datasets used in other methods	
	Public	In-house	Public	In-house	Public	In-house
RNA Cluster Analysis	Akiyama et al. TrainSet-A [8], Akiyama et al. TrainSet-B [8]	–	–	–	–	–
mRNA Identification	MLOS Flu Vaccines (Sanofi-Aventis) Dataset [107], Nieuwkoop et al. Dataset [107], Wint et al. Dataset [107], lixiProtein Expression Dataset [107], Groher et al. Dataset [107], Diez et al. Dataset [107], RYOS-I Dataset [107]	–	–	–	–	–
Small Non-coding RNA Classification	–	–	Aoki et al. Dataset [108], Deng et al. Dataset [109]	Non-Coding RNA Classification Dataset [110]	–	–
Circular RNA Identification	–	–	circRNAs [111,14], circRNA–Protein associations [112], Protein–Protein interactions [112,113]	–	Niu et al. Dataset [13]	–
Long Non-coding RNA Identification	Arabidopsis thaliana Dataset [114], Brassica napus Dataset [114], Brassica oleracea Dataset [114], Brassica rapa Dataset [114], Glycine max Dataset [114], Oryza sativa Dataset [114], Zea mays Dataset [114]	Dai et al. Dataset [115]	–	Human 1 [116], Human 2 [116], Mouse [116]	Tian et al. Dataset [117], Musleh et al. Dataset [118]	Nadir et al. Dataset [119]
Pre-micro RNA Identification	Gupta et al. Dataset [120], Raad et al. Dataset [121]	–	–	–	–	–
CRISPR/Cas9 single guide RNA Identification	–	–	–	–	WT Dataset [122], ESP Dataset [122], HF Dataset [122], xCas Dataset [122], SpCas9 Dataset [122], Sniper Dataset [122], HCT116 Dataset [122], HELA Dataset [122], HL60 Dataset [122]	–
Enhancer RNA Identification	Zhang et al. Dataset [123]	–	–	–	–	–
Promoter Identification	Mai et al. Dataset [124], Wang et al. Dataset [125]	–	–	–	–	–
RNA-Gene Association Prediction	–	–	–	Xia et al. Dataset [126], Yoon et al. Dataset [127]	–	–

Table 2 (*continued*)

Task Name	Datasets used in Language Models		Datasets used in word embeddings		Datasets used in other methods	
	Public	In-house	Public	In-house	Public	In-house
RNA-Disease Association Prediction	Zou et al. Dataset [128], MDAv2.0 Dataset [129], MDAv3.2 Dataset [129], Dai et al. Data2 Dataset [130], Ning et al. Dataset (1,2) [131], HMDD Dataset [132], HMDAD Dataset [132], LncRNADisease v2017 Dataset [132], Wu et al. Dataset (1,2,3) [133] Fu et al. Dataset [134], Zhou et al. Dataset [134], Li et al. Dataset [135], Li et al. Dataset (1,2) [136]	Wu et al. Dataset (1,2) [137], Ma et al. Dataset [138], Awn et al. Dataset [139]	Lu et al. Dataset [140], Ding et al. Dataset [141], Jindal et al. Dataset [141], Wang et al. Dataset [142], Human PPI [143], Disease–gene interaction Dataset [143], miRNA–Gene Network [143], miRNA–Disease Network [143], Duan et al. Dataset (1,2,3) [144]	Sun et al. Dataset [145], Zheng et al. Dataset [146]	Tian et al. Dataset [147], Ruan et al. Dataset [148], Xu et al. Dataset [149], Ji et al. Dataset [150], Li et al. Dataset (DS1, DS2) [151], Tang et al. Dataset [151], Huang et al. Dataset [152], Cao et al. Dataset [153], Gong et al. Dataset [154], Lan et al. Dataset (1,2,3,4,5) [155], Li et al. Dataset (1,2) [155], Lu et al. Dataset (1,2) [156], Zhang et al. Dataset [156], lncRNADisease Dataset [157], MNDR Dataset [157], Li et al. Dataset [158], Ma et al. Dataset [158], Xia et al. Dataset [158], CircR2Disease Dataset [159], circRNADisease Dataset [159], Circ2Disease Dataset [159], circAtlas Dataset [159]	Kang et al. Dataset (1,2,3) [160], Fu et al. Dataset (1,2) [161], Lu et al. Dataset [161], Yao et al. Dataset [162], Chen et al. Dataset (1,2) [163], Wang et al. Dataset [164], Liang et al. Dataset [165]
Coding RNA-Protein Interaction Prediction	–	–	NPInter2.0 [166], NPInter2.0_lncRNA [166], RPI7317 [166], RPI2241 [166], RPI38317 [166], Li et al. Dataset [167], Zhao et al. Dataset (1,2) [168]	Wei et al. Dataset [169], RPI369 [170], RPI1807 [170], RPI488 [170]	RPI369 Dataset [171], RPI488 Dataset [171], RPI1446 Dataset [171], RPI1807 Dataset [171], RPI2241 Dataset [171]	–
Protein-RNA Binding Sites Prediction	Non-Redundant Dataset [172], circRNA fragment Dataset 1 [173], Full length circRNA Dataset [173], circRNA fragment Dataset 2 [173], Linear RNA fragment Dataset [173], Protein Dataset [174], WTAP [175], FXR1 [175], C17ORF85 [175], QKI [175], TAF15 [175], AUF1 [175]	Jia et al. Dataset [176], Zhang et al. Dataset [176]	37 RBP Datasets [177], IGF2BP1 [178], IGF2BP3 [178], LIN28A [178], LIN28B [178], Stražar et al. Dataset [179]	–	RBP-120 Dataset, Maticzka et al. Dataset [180], RBP-24 Dataset [180]	Liu et al. Dataset [181]
Protein-RNA binding affinity prediction	Shen et al. Benchmark Dataset [182]	–	–	–	–	–

(*continued on next page*)

Table 2 (*continued*)

Task Name	Datasets used in Language Models		Datasets used in word embeddings		Datasets used in other methods	
	Public	In-house	Public	In-house	Public	In-house
Non coding RNA Interaction Prediction	_	CircBank Dataset [183]	Zhao et al. Dataset [184], Wang et al. Dataset [185], CMI-9905 Liu et al. Dataset [185], CMI-9589 Liu et al. Dataset [185]	CMI-753 Dataset [186]	Fu et al. Dataset [187], Zhou et al. Dataset [187]	
RNA Sub-cellular Localization Prediction	Zeng et al. Dataset [19]	_	_	Asim et al. Dataset [18], Lin et al. Dataset [17]	_	_
ac4C-Acetyl Cytidine Modification Prediction	Wang et al. Dataset [188]	_	_	_	_	_
5mU-Methyl Uridine Modification Prediction	_	_	Feng and Chen et al. [189], Jiang et al. [189]	_	GSE78040 Dataset [190], GSE63753 Dataset [190]	_
2’-OmU Methyl Uridine Modification Prediction	_	Soylu et al. Dataset [191]	_	_	_	_
6mA-Methyl Adenosine Modification Prediction	Wang et al. Dataset [192], MultiRM Dataset [193], YTHDF2 PAR-CLIP Dataset [194], Wan et al. A101 Dataset [195]	Dao et al. Mouse Dataset [196]	Zhang et al. Dataset [197], S51 Dataset [198], H41 Dataset [198], M41 Dataset [198]	cDNA Sequence [199]	Tu et al. P Dataset [200], Tu et al. N Dataset [200], Wang et al. Dataset [201], m 6 A-Atlas Dataset [201], Dao et al. Human Dataset [196], Dao et al. Rat Dataset [196]	m6A-Seq Dataset [202]
7mG-Methyl Guanosine Modification Prediction	_	Benchmark Dataset [203], Independent Dataset [203], Dai et al. Dataset [204]	_	Chen et al. Dataset [205], Dai et al. Dataset [205]	Chen et al. Dataset [206]	_
5mC-Methyl Cytosine Modification Prediction	_	_	Hasan et al. Dataset [207]	_	Kurata et al. Dataset [208]	_
Methylation Modification Prediction	DS_song Dataset [209], N1-methyladenosine (m1A) Dataset [192], N6-methyladenosine (m6A) Dataset [192], Pseudo-uridine (pseU,Ψ) Dataset [192]	Zhang et al. M. musculus Dataset [10], Zhang et al. A. thaliana Dataset [10], Zhang et al. S. cerevisiae Dataset [10]	Chen et al. Dataset [210], Song et al. Dataset [210], m1A site Dataset [211], m6A site Dataset [211]	_	_	Wang et al. Dataset [212]
RNA-Splicing Sites Prediction	Chen et al. Dataset [213], SpliceAI-80nt [214], SpliceAI-256nt [214], SpliceAI-400nt [214], SpliceAI-2k [214]	_	_	_	_	_

Table 2 (*continued*)

Task Name	Datasets used in Language Models		Datasets used in word embeddings		Datasets used in other methods	
	Public	In-house	Public	In-house	Public	In-house
Alternative Splicing Prediction	_	_	_	Brawand et al Dataset [215]	_	_
RNA Functions Prediction	Shulgina et al. Dataset [216]	bpRNA-1 [11], PDB [217], bpRNA-1m TS0 [217], ArchiveII [217], ArchiveII600 Dataset [218], bpRNA TS0 Dataset [218], RNAcontact Test80 Dataset [218], HeLa Dataset [218], Random7600 Dataset [218], Human7600 Dataset [218]	_	_	miRNA2GO-337 [219]	_
RNA Structure Prediction	Rfam_TR0 Dataset [220], Rfam_VL0 Dataset [220], Rfam_TS0 Dataset [220], Szikszai et al. Dataset [221], Zhang et al. Dataset (1,2) [222], Kalicki et al. Dataset [223], RNA-Puzzles [224], PDB Dataset [224], PT_128 Dataset [225], PT_512 Dataset [225]	bpRNA-1m Dataset (TR0) [226], PDB Dataset [226], RNAStralign Dataset [227]	_	American Gut microbiome [228], Gevers et al.'s Crohn's disease Dataset [228], SILVA 16S rRNA Dataset [228]	Stralign [229], ArchiveII [229], RNAStralign [230], ncRNA benchmark [230]	_
Spatial Gene Expression Analysis	hESC Dataset [231], hHEP Dataset [231], mDC Dataset [231], mESC Dataset [231], mHSC-E Dataset [231], mHSC-GM Dataset [231], mHSC-L Dataset [231]	_	_	_	_	_
Gene Expression Prediction	Khan et al. Dataset [232]	PBMC scRNA-Seq Dataset [233], TCGA RNA-Seq Dataset [233], Babjac et al. Dataset [234]	_	_	_	_
Cell-Specific Gene Regulatory Networks Prediction	hESC(1,2) Dataset [235], mESC(1,2) Dataset [235], mESCs Dataset [235], Bone Dataset [235], Dendritic Dataset [235]	_	_	_	_	_
16S rRNA Taxonomic Classification	_	_	16S rRNA amplicon Sequences [236]	McDonald et al. Greengenes Dataset {ziemski2021beating}	_	_
16S rRNA Gene Copy Number Prediction	_	_	_	_	Miao et al. 16S rRNA gene Dataset [237]	_

(*continued on next page*)

Table 2 (*continued*)

Task Name	Datasets used in Language Models		Datasets used in word embeddings		Datasets used in other methods	
	Public	In-house	Public	In-house	Public	In-house
Micro RNA Target Prediction	miRAW Dataset [238], DeepMirTar Dataset [238], deepTargetPro Dataset [238]	Pla et al. miRAW Dataset [239]	miRAW Dataset [240], DeepMirTar [240], DeepMirTarIn [240]	_	_	_
Small Interfering RNA Target Prediction	Huesken et al. Dataset [241], Reynold et al. Dataset [241], Katoh et al. Dataset [241]	Xu et al. Dataset (1,2,3) [241]	_	_	_	_
mRNA Degradation Prediction	OpenVaccine challenge Dataset [26], In vitro half-life Dataset [26]	_	_	_	_	_
RNA-Seq Coverage Prediction	Linder et al. Dataset [27]	_	_	_	_	_
Cell-type Detection	Multiple Sclerosis Dataset [242], Myeloid Dataset [242], hPancreas Dataset [242], PBMC 10K Dataset [242], Perirhinal Cortex Dataset [242], Immune human Dataset [242], COVID-19 Dataset [242], Adamson perturbation Dataset [242], Norman perturbation Dataset [242], Multiome PBMC Dataset [242], BMMC Dataset [242], ASAP PBMC Dataset [242]	_	_	_	Sim Dataset (1,2) [243], Specter Dataset [243], 10X_10K Dataset [243], SMAGE Dataset [243], Spleen Dataset [243], BMNC Dataset [243]	_

7. A brief look on representation learning and predictors used in RNA sequence analysis predictive pipelines

This section delves into 16 widely used word embedding methods, 8 language models, and 35 machine and deep learning predictors used in 47 different RNA sequence analysis tasks.

7.1. RNA sequence representation learning using word embeddings

In the realm of Natural Language Processing (NLP), the advent of word embedding methods have revolutionized efficacy of AI-driven applications. These approaches capture syntactic and semantic relationships of words to generate similar vectors for similar words and dissimilar vectors for dissimilar words. For example, words like good, better and best represent a same concept, so their vectors will be similar to each other. On the other hand vectors of words like good and bad will be dissimilar because both words are opposite and represent different concepts. These approaches have also introduced the concept of transfer learning in NLP domain. Primarily, statistical vectors of words are generated by training word embeddings models on large unlabeled textual corpora. Similar to computer vision domain, where models are first trained on imagenet data, word embeddings also provides pretrained weights at input layer of deep learning models. Following the promising performance of various word embedding approaches on different NLP tasks [244] [245] [246] [247], researchers have increasingly adopted these approaches for genomics and proteomics sequence analysis tasks that share significant similarities with NLP tasks. As is shown in Fig. 5, overall 16 different word embedding approaches used in RNA sequence analysis can be classified into 2 broad categories: non-graph based, and graph based word embedding approaches.

Non-graph based word embedding approaches discretize RNA sequences into overlapping or non-overlapping k-mers. Overlapping k-mers are generated by sliding a fixed-size window across the sequence with a stride size smaller than the size of the window. For example, if the window size is 3 and the stride size is 1, the resulting k-mers overlap by 2 positions. Non-overlapping k-mers are generated by sliding a fixed-size window with the stride size equal to the size of the window. This means that each k-mer starts immediately

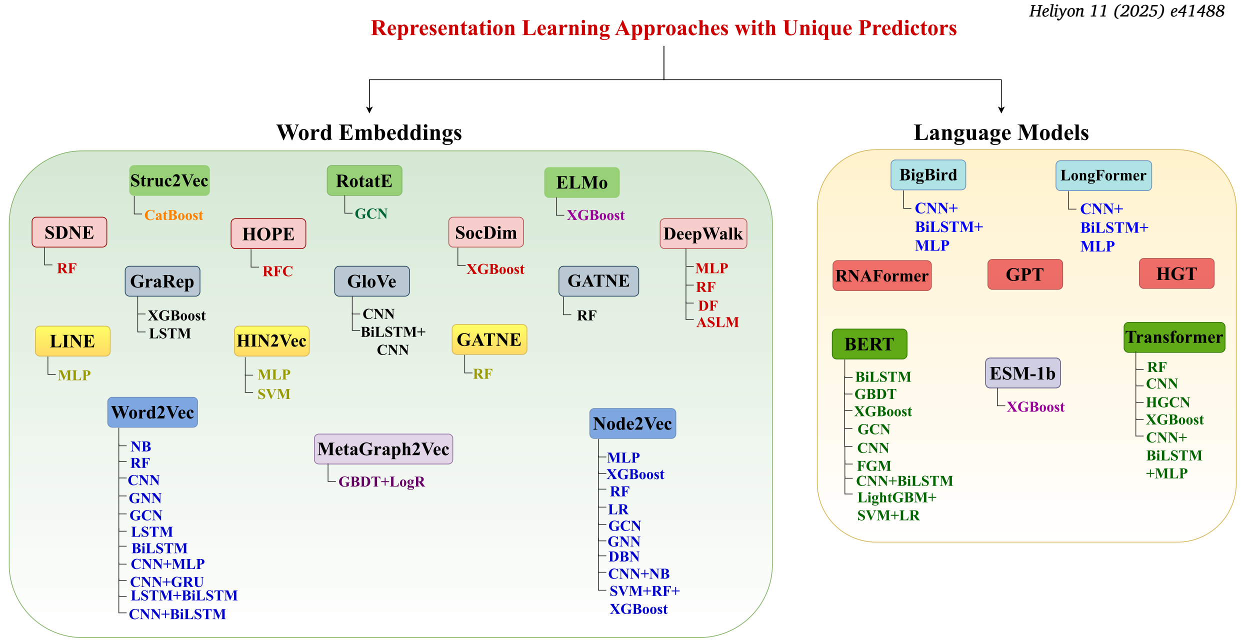

Fig. 5. Utilization of 16 Different Word Embedding Methods and 8 Large Language Models namely BigBird, LongFormer, RNAFormer, Generative Pre-trained Transformers (GPT), Heterogeneous Graph Transformer (HCT), Bidirectional Encoder Representations from Transformers (BERT), ESM-1b, and Transformer in Diverse RNA Sequence Analysis Pipelines based on a Variety of Machine and Deep Learning Algorithms such that RFC: Rotation Forest Algorithm, RF: Random Forest, CNN: Convolutional Neural Network, GNN: Graph Neural Network, XGBoost: Xtreme Gradient Boosting, MLP: Multilayer Perceptron, GCN: Graph Convolutional Network, LogR: Logistic Regression, LSTM: Long Short Term Memory, GBDT: Gradient Boosting Decision Trees, BiLSTM: Bidirectional Long Short Term Memory, SVM: Support Vector Machine, GBU: Gated Recurrent Unit, NB: Naive Bayes, NNRM: Neural Network Regression Model, DF: Deep Forest, ASLM: Adaptive subspace learning model, ERM: ElasticNet Regression Model, DNN: Deep Neural Network, HGCN: Hyper Graph Convolutional Network.

after the previous k-mer ends, with no overlap. The size of the k-mer is determined by the size of the window. Researchers often generate pretrained embeddings using different k-mer sizes and select the k-mer size that performs best on downstream tasks. After generating k-mers, these k-mers are passed to word embedding models for representation generation. Specifically, Word2vec [109,108,111,127,126,248,139,249,168,177,178,250,179,251,7,187,184,185,17,197–199,205,207,210,215,228,252,236,240,253] has 2 variants namely: 1) Continuous bag of words paradigm (CBoW), 2) SkipGram. In CBoW, the context of neighboring k-mers are used to predict a target k-mer whereas SkipGram predicts neighboring k-mers by using a target k-mer. For better understanding lets take a toy RNA sequence "AGUCCCU" with k = 3, four k-mers are generated such as AGU, GUC, UCC, CCU. Assume "GUC" is target k-mer and window size equal to 1, neighboring k-mers are "AGU" and "UCC". In this case, CBOW model predicts target k-mer "GUC" using neighboring k-mers "AGU" and "UCC", while Skip-gram model predicts neighboring k-mers ("AGU" and "UCC") based on target k-mer "GUC". Primarily, Word2Vec is a neural network-based architecture that consists of an input layer, a hidden layer, and an output layer. At input layer, each k-mer is initialized with a random d-dimensional vector, which is then passed to hidden layer to learn relationships between k-mers. These relationships are passed to output layer to estimates probability/ies of output k-mers based on context of input k-mers. The predicted probabilities are further used to compute loss value. This shallow neural network is trained to maximize the probability of the next k-mer given the context.

Furthermore, GloVe [116,211] learns k-mer embeddings by factorizing the co-occurrence matrix. Co-occurrence matrix represents the number of times $k - mer_i$ appears in the context of $k - mer_j$ within a fixed window size. This matrix captures how frequently k-mer appear together in the entire corpus. Then, it calculates the probability of $k - mer_i$ appearing in the context of $k - mer_j$. GloVe's objective is to find k-mer vectors and context vectors such that their dot product approximates the logarithm of co-occurrence probability. Unlike Word2vec and Glove that generate context independent embeddings that assign a single vector to each k-mer, Embeddings from Language Models (ELMO) [172,254,255] generates different embeddings for k-mer based on its context. ELMo uses a deep bidirectional language model (BiLM) that consists of multiple layers of Long Short-Term Memory (LSTM) networks. This model reads the sequence in both forward and backward directions to capture the context of each k-mer from both sides. The model is trained on a large amount of sequences to predict k-mers based on their context. After training, it provides embeddings at multiple layers of the network. Each layer captures different aspects of the k-mers scientific meaning.

On the other hand, rather than utilizing unlabeled data as it is, graph-based embedding methods first map data into a graphical space. Based on the relationships between nodes in the graph, these methods capture diverse types of information and generate new data on which a further model is trained. Similar to non-graph-based methods, first k-mers are generated and a graph is constructed using the relationships between k-mers. For example, if the input corpus has a sequence of k-mers such as AC, CG, GT, TC, etc., a sliding window of size two with stride size one is used to generate pairs of k-mers like (AC, CG), (CG, GT), (GT, TC), and so on. In the constructed graph, k-mers represent nodes, and relationships between k-mers represent links between nodes. Random walk based embedding methods like Node2vec [145,256,146,143], DeepWalk [140–142,257] perform random walks on this graph to generate new samples in form of sequences of nodes connected by edges, also called meta-paths. Apart from target k-mer and context sampling, a small subset of k-mers that are not part of the context are selected as negative samples. These new samples are used to train Word2Vec Skipgram model to generate statistical vectors of k-mers. Although both Node2vec, DeepWalk working seems quite similar, however, both differ by the type of random walk and captured information. Node2vec applies a biased random walk strategy to explore diverse neighborhoods of k-mers. It combines breadth-first and depth-first search strategies using two parameters (p) and (q), to control the likelihood of revisiting a k-mer and exploring new k-mers, respectively. Node2vec captures both local and global structures of graph. Whereas, DeepWalk perform uniform random walks where each step in the walk has an equal probability of moving to any of the neighboring k-mers. DeepWalk has no additional parameters to control the walk behavior, hence it only captures local structures of graph.

Furthermore, HIN2vec [112] makes use of random walk and meta-paths paradigm to generate training data in the explicit form of (a, b, B(a, b, z)) where a and b denote two k-mers, z denotes the relationship among two k-mers, and B(a, b, z) denotes a binary value representing whether there exist a relationship z among a and b k-mers. As each meta-path represents a specific pattern of relationships within the network, HIN2vec [112] mainly targets multiple prediction tasks to capture various types of relationships between k-mers. Instead of learning separate models for each type of relationship, HIN2Vec [112] jointly learns a single three-layer feedforward neural network model that can handle all the prediction tasks. For any given pair of k-mers, the model predicts a set of target relationships defined by the meta-paths. These predictions involve estimating the probability that a relationship exists between the k-mers according to the specified meta-path. After iterative training of the neural network using back-propagation and gradient descent, optimized dense k-mers vectors are treated as final embeddings. Another approach Struc2Vec [186] constructs a multi-layer graph where each layer represents a different level of structural similarity which allows the model to learn embeddings that reflect the structural roles of k-mers in the graph. K-mers in different layers of the hierarchical graph are connected with weighted edges. The weight of these edges is determined by the structural distance that quantifies the number of edges connected to k-mer and their neighbors. Struc2vec [186] employs a biased random walk technique to sample paths within the hierarchical graph, where the probability of moving from one k-mer to another is higher if their structural distance is smaller. The random walk ensures that the sampling process captures local topological structures such as k-mers degree, neighboring k-mers, and neighborhood degree effectively while ignoring the specific positions of k-mers in the graph. Struc2vec leverages these local topological structures to generate embeddings that reflect the structural properties of k-mers.

In addition, General Attributed Multiplex Heterogeneous Network Embedding (GATNE) [256] method make use of random walks to generate new sequences of k-mers which serve as training data. GATNE considers all nodes and edges of different types and employs a combination of multi-layer network to effectively capture the complex relationships. The method starts with a base embedding layer that generates a shared embedding for each k-mer, irrespective of the edge type. This base embedding serves as a common feature representation across all connections. Additionally, GATNE [256] includes edge-specific embedding layers for each type of relationship which allows it to learn the unique characteristics of different connections. It uses an attention mechanism layer to weigh the importance of various neighbors and relationships, and eventually aggregate information from the most relevant ones. The final combination layer integrates the base embeddings and the edge-specific embeddings using the attention scores, resulting in a comprehensive, low-dimensional embedding for each k-mer. Another approach called MetaGraph2Vec [144] treats nodes and edges as of different types. It builds a metagraph that specifies the types of nodes and edges which should be considered in the random walks to ensure that the walks capture the complex and meaningful relationships among different types of k-mers. Then, it performs, random walks guided by metagraph to generate k-mer sequences and train skip-gram model. Random Walk with Restart (RWR) [147] approach generates node embeddings by simulating a random walk that occasionally restarts from the initial node. This method is particularly useful for capturing the local and global structure of the graph. RWR [147] begins by selecting a starting node, often referred to as the “seed” node. In a k-mers graph, this could be any k-mer of interest. The random walk is initialized from this node. At each step of the walk, the algorithm moves to a neighboring node based on transition probabilities. These probabilities are typically derived from the edge weights between nodes. For instance, if a k-mer has a high similarity or frequent occurrence with another k-mer, the transition probability between these nodes will be higher. At each step, there is a predefined probability that the walk will restart from the initial seed node. This ensures that the walk does not drift too far from the starting point, maintaining a balance between exploring the graph and focusing on the local neighborhood of the seed node. The random walk continues until it reaches a steady state, where the probability distribution over the nodes no longer changes significantly. This steady-state distribution represents the importance or influence of each node relative to the seed node. Once the steady state is achieved, the resulting probability distribution is used to generate the node embeddings. Each node’s embedding is a vector that captures its relationship with the seed node and other nodes in the graph.

Beyond random walks, some graph embedding methods like HOPE [258], LINE [219], SDNE [154] make use of proximity information to learn low-dimensional vector representations of k-mers. Proximity information capture the notion of how related or connected two k-mers are based on their attributes, relationships, or interactions. Given a k-mers graph, High Order Proximity preserved Embedding (HOPE) [258] constructs a high-order proximity matrix (S). This matrix quantifies the similarity between directly connected k-mers as well as in-directly connected k-mers on the basis of number of distinct paths of length (k) between k-mers (i) and (j). Then, HOPE [258] decomposes the high-order proximity matrix (S) into two smaller matrices (U_s) and (U_t) for source and target k-mer embeddings, respectively. The source k-mer embedding encodes how a k-mer influences others, while the target k-mer embedding encodes how a k-mer is influenced by others. Even in undirected graphs, this dual representation allows for capturing more complex relationships and dependencies between k-mers. The decomposition is done in such a way that the product of (U_s) and (U_t) approximates the original high-order proximity matrix (S). The optimization objective is to minimize the difference between the high-order proximity matrix (S) and the product of the two embedding matrices (U_s) and (U_t). Unlike HOPE, Large-scale Information

Network Embedding (LINE) [219] encodes first-order proximity information by capturing the direct interactions between k-mers. Also, it encodes second-order proximity information by capturing the similarity in the k-mers neighborhood structures. The method optimizes the embeddings such that k-mers with similar contexts have similar embeddings. This is achieved by treating the neighborhood structure as a probability distribution and minimizing the Kullback-Leibler divergence between the actual and the predicted distributions. Another method Structural Deep Network Embedding (SDNE) [154] represents the graph by an adjacency matrix and employs a deep autoencoder to compresses the input data (adjacency matrix) into a lower-dimensional representation and reconstruct the original adjacency matrix from this compressed representation. Apart from learning embeddings of adjacency matrix, SDNE [154] preserves first-order proximity information by minimizing the reconstruction error between the original adjacency matrix and the reconstructed adjacency matrix. It also preserves second-order proximity information using the Laplacian Eigenmaps objective, which ensures that nodes with similar neighbors have similar embeddings. The SDNE [154] combines the first-order proximity objective with the second-order proximity Laplacian Eigenmaps objective into a unified loss function. This loss function is then optimized to learn the embeddings.

Apart from random walk, and proximity information, some graph embedding methods make use of matrix factorization techniques like Singular Value Decomposition [259,157] to learn final k-mers embeddings. GeneticSeq2Vec (or GraRep) [18,189] generates an adjacency matrix of k-mers graph, where each entry indicates whether a pair of k-mers (nodes) are connected. To capture more complex relationships between k-mers at different distances, k-hop proximity matrices are generated which represent connections that span multiple steps in the graph. These k-hop proximity matrices are factorized using Singular Value Decomposition (SVD) [259,157] to produce lower-dimensional representations. This step helps in capturing the essential features and relationships of the k-mers. The representations from different k-hop matrices are concatenated to form a comprehensive feature vector for each k-mer to ensure that both local and global relationships are captured. Also, SocDim (Social Dimensions) [189] operates on a graph with k-mers as nodes by extracting social dimensions that capture the community structure of the graph. It first identifies communities in the graph and then represents each k-mer as a vector of its affiliations to these communities. SocDim [189] measures the quality of the community detection using a metric called modularity, which quantifies the strength of division of a network into communities. Modularity is calculated by comparing the actual edge density within communities to the expected edge density if edges were distributed randomly. Actual edge density is a measure of how densely the edges are distributed in a graph relative to the number of possible edges. It is calculated as the ratio of the number of actual edges present in the graph to the total number of possible edges. Mathematically, the modularity matrix (B) is derived from the adjacency matrix (A) and degree vector (d). It adjusts the adjacency matrix to reflect the community structure by subtracting the expected edge density. Afterwards, it extracts the principal components of the modularity matrix (B) to identify the most significant community structures. This is done by performing eigenvector decomposition on (B) to obtain the leading eigenvectors. These eigenvectors represent the social dimensions of the network. Leading eigenvectors obtained from the modularity matrix (B) are used as the node embeddings.

Furthermore one unique approach called RotatE [153] operates on a knowledge graph with k-mers as nodes by representing relations as rotations in a complex space. It models each relationship as a rotation from the source k-mer to the target k-mer. The embeddings are learned by optimizing a scoring function that measures the plausibility of each triplet (source, relation, target). The objective is to capture the relational patterns in the graph, such as symmetry, antisymmetry, inversion, and composition.

In RNA sequence analysis landscape, word embedding methods are employed in two different ways to generate pre-trained embeddings. First approach breaks down RNA sequences into k-mers and generates k-mers embeddings. Alternatively, second approach generates embeddings for entire RNA sequences, which can be further applied in two distinct ways for homogeneous and heterogeneous networks. Homogeneous network deals with a same type biomolecule (RNA). In contrast, heterogeneous networks involve multiple types of biomolecules, such as miRNAs, lncRNAs, circRNAs, protein, and diseases. In heterogeneous graphs, nodes represent biomolecules and their interactions or associations are represented as edges. Heterogeneous networks are more complex than homogeneous network and extracts more detailed and comprehensive relationships through graph-based embedding methods. Specifically, 41 RNA sequence analysis predictive pipelines employ first approach to generate embeddings for 19 different RNA sequence analysis tasks [109,108,111,116,127,126,146,139,258,249,167,166,169,168,260,170,177,178,250,179,251,7,184,185,17,18,189,197,255,198,199,205,207,210,211,215,228,252,236,240,253]. On the other hand, 17 predictive pipelines leverage second approach to generate embeddings for 7 different RNA sequence analysis tasks including circular RNA identification [112,14], miRNA-disease associations prediction [154,256,142,153,141,140,143,145], lncRNA-disease association prediction [261,259,144,262,248], circRNA-disease association prediction [257], circRNA-miRNA interactions prediction [186], and RNA function prediction [219].

7.2. RNA sequence representation learning using language models

In the rapidly advancing field of Natural Language Processing (NLP), the introduction of the Transformer model has marked a significant milestone as it has established a new standard for future language model innovations [216,107]. The Transformer [120] and distinct language models including BERT [241], GPT-3 [242], and ESM-1 [172], have greatly expanded the capabilities of machines in understanding and generating human language [216,107]. These models are not only remarkable for their text comprehension and generation abilities but also for their applications in various domains, including genomics and proteomics sequence analysis [172]. By creating highly effective representations of biological sequences, these models are transforming numerous genomics and proteomics sequence analysis tasks [172]. To aid RNA sequence analysis researchers, we provide an overview of the key features, benefits, and drawbacks of 8 most commonly used sophisticated large language models: Transformer [120], BERT [107], GPT-3 [216], Hetergeneous Graph Transformer (HGT) [128], BigBird [115], LongFormer [263], RNAFormer [220], and ESM-1b [172], mentioned in Fig. 5. Table 3 illustrates 8 distinct language models and their variants, organized into 4 categories based on their underlying architectures.

Table 3
A Summary of 8 Contemporary Language Models utilized in RNA Sequence Analysis tasks.

Architecture Type	Language Model, Release Year	Language Model Variants	Number of Layers in Encoders	Number of Layers in Decoders
Encoder-Decoder	Longformer [263], 2020	Base	6	6
		Large	12	12
	BigBird [264], 2020	BigBird-ITC (Base)	12	12
		BigBird-ITC (Large)	24	24
		BigBird-ETC (Base)	12	12
		BigBird-ETC (Large)	24	24
	Transformer, [265], 2017	Base	6	6
		Big	6	6
Encoder-Only	BERT, [266], 2019	Base	12	_
		Large	24	_
Decoder-Only	GPT, 2018	GPT-1 [267]	_	12
		GPT-2 small [268]	_	12
		GPT-2 medium [268]	_	24
		GPT-2 Large [268]	_	36
		GPT-3 [269]	_	96
		GPT-4 [270]	_	120
Special Transformer Variants	ESM-1, 2021	ESM-1b [271]	33	_
		ESM-1v [272]	33	_
		ESM-MSA/ MSA Transformer [273]	12	_
	RNAformer [274], 2023	32 D	32 Residual convolution blocks (each block: 6 layers)	
		64 D	64 Residual convolution blocks (each block: 6 layers)	
		128 D	128 Residual convolution blocks (each block: 6 layers)	
		256 D	256 Residual convolution blocks (each block: 6 layers)	
	Heterogeneous Graph Transformer [275], 2020	_	256 Residal GNN blocks (each block: 3 layers)	

These categories include encoder-decoder architecture, encoder-only architecture, decoder-only architecture, and special transformer variants. Moreover, Table 3 outlines number of layers in language model architecture and specifies number of encoders or decoders along with their respective layers.

The Transformer model [120,121,276,129,137,133,138,135,136,175,182,19,195,203,209,221,224,225,227,217,231–233,277,235,238,26,27], introduced by Vaswani et al. [265] in 2017, represents a significant departure from previous models that relied on recurrent or convolutional neural networks for processing sequential data. This model employs a unique architecture centered on attention mechanisms to manage long-range dependencies and grasp the context and semantics of sequences more effectively [265,120]. Notable innovations of the Transformer include positional encoding and self-attention mechanisms [265,120]. Positional encoding assigns a unique identifier to each nucleotide or group of nucleotides which helps the model recognize the order and context of sequences. The self-attention mechanism allows the model to evaluate the importance of each nucleotide in relation to others and enhances model's ability to process and predict scientific language patterns [265,120]. The primary advantage of the Transformer lies in its training and inference efficiency due to parallel sequence processing [265,120]. However, it demands substantial computational resources, which can be a constraint in resource-limited settings. Despite this, its flexibility and scalability in handling diverse genomics tasks make it a favored choice in many advanced AI applications [120].

Bidirectional Encoder Representations from Transformers (BERT) [8,107,114,123,125,124,131,132,139,173,174,176,278–280,183,16,188,191,281,204,193,192,282,10,213,214,226,222,223,11,218,234,239,241,283,284], introduced by Google in 2018 [266], is pretrained on extensive text corpora such as Wikipedia and books [266]. BERT has transformed NLP tasks through its transformer-based architecture, which allows the model to consider the context of words bi-directionally, rather than uni-directionally [266]. What sets BERT apart is its deep bidirectional nature achieved using the transformer model and specific techniques like Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [266]. This enables BERT to understand the context of a word based on all surrounding words in a sentence, not just those that come before it. It excels at capturing the semantics and contextual information of input text through self-supervised learning tasks such as MLM and NSP [266]. In RNA sequence analysis, BERT is employed to transform RNA sequences into a statistical feature space and is subsequently fine-tuned for specific downstream tasks. BERT captures the semantics of RNA sequences by dynamically learning their representations using a multihead self-attention mechanism. By leveraging transfer learning, BERT is pretrained on a large corpus and then fine-tuned for specific RNA sequence analysis tasks, allowing it to adapt to diverse applications [131]. During pretraining, BERT uses MLM and NSP tasks to learn the contextual relationships between nucleotides in RNA sequences [131].

The main advantages of BERT include its high accuracy and efficiency across various RNA sequence analysis tasks. This is due to its robust handling of context and bidirectional training [131]. BERT effectively captures both discriminative and semantic relationships of nucleotides which makes it highly effective in characterizing RNA sequences [131]. BERT-based models have shown superior performance compared to traditional methods in RNA sequence analysis tasks such as enhancer identification and strength prediction [131]. Additionally, BERT can be adapted to specific applications by pretraining on domain-specific custom corpora [131]. However, BERT is a large model requiring substantial computational resources for training and inference on extensive datasets. Its optimal performance is achieved when trained on large and diverse datasets, which may not always be available for specific RNA sequence analysis tasks. While BERT delivers state-of-the-art results in many scenarios, it requires fine-tuning for specific tasks, which can be resource-intensive. Furthermore, BERT's performance can degrade with longer texts, and its complex architecture makes it challenging to interpret the learned representations and understand the underlying biological mechanisms [131].

GPT-3 [216,242], developed by OpenAI, is among the most advanced AI language models available today [269]. It is renowned for its remarkable ability to generate text that closely resembles human writing, marking a significant milestone in natural language processing. GPT-3 is built on the transformer architecture, which uses self-attention mechanisms to process input data [269]. While GPT-2 featured 1.5 billion parameters, GPT-3 takes a quantum leap with 175 billion parameters. This vast increase in parameters significantly enhances its capacity to produce coherent and contextually appropriate text [269]. Unlike BERT and XLNet, GPT-3 maintains an autoregressive model which allows to predict the next nucleotide in a sequence based on the preceding nucleotides, whereas BERT employs bidirectional context [269].

One of GPT-3's key innovations is its use of alternating dense and locally banded sparse attention patterns. Dense attention considers all input nucleotides at once, while sparse attention focuses on a subset which makes the model more efficient and scalable. This approach allows GPT-3 to manage long-range dependencies while maintaining computational efficiency [269]. A standout feature of GPT-3 is its impressive performance in few-shot settings. Unlike models that require extensive fine-tuning with large amounts of task-specific data, GPT-3 can excel in new tasks with minimal sequences. This flexibility offers a notable advantage over models like BERT, which typically need substantial fine-tuning for each specific task. GPT-3 demonstrates strong performance across various tasks, often matching or surpassing that of fine-tuned models, which makes it a versatile tool for a wide array of applications [269].

Heterogeneous Graph Transformer (HGT) [128,130,256,285] is a graph neural network architecture designed to handle heterogeneity and dynamics in large-scale graphs. HGT addresses the challenges of heterogeneous graphs by introducing node-type and edge-type dependent attention mechanisms. It parameterizes weight matrices based on meta relation triplets which allow nodes and edges of different types to maintain specific representation spaces. HGT utilizes message passing across layers to incorporate information from high-order neighbors of different types. This enables the model to capture complex relationships and dependencies in the graph. HGT incorporates Relative Temporal Encoding (RTE) to model structural temporal dependencies in the graph. It enables the model to learn the temporal evolution of the graph, even with unseen and future timestamps. HGT uses meta relation triplets to parameterize weight matrices which enables the attention calculation over each edge. This feature enables the model to capture important relationships and interactions between different types of nodes. HGT can automatically learn and extract “meta paths” that are important for downstream tasks without the need for manual design. This flexibility allows the model to adapt to different graph structures and tasks. HGT allows the integration of diverse data sources and captures the specific characteristics of different types of nodes and edges through dedicated representations. However, use of multiple projection weights and attention heads for dedicated representations requires careful parameter tuning to achieve optimal performance, which can be a tedious process. Additionally, training HGT on large-scale graphs demands significant computational resources.

The ESM-1b language model [172] possesses a unique working approach that distinguishes it from other language models. ESM-1b is a single-sequence language model explicitly designed for protein sequence analysis. It is trained on vast databases of unaligned and unrelated protein sequences through the use of masked language modeling. ESM-1b design incorporates the physicochemical attributes of amino acids in its representations which allow it to encode essential biochemical knowledge. Unlike other domain-specific language models that rely on next token prediction or multiple sequence alignments (MSAs), ESM-1b focuses on single-sequence training and does not require MSAs during inference. ESM-1b has proven to be competitive in predicting variant effects, making it a valuable tool for examining RNA sequences. The model is capable of capturing a broad range of protein variations and properties, enabling it to handle diverse RNA sequences. Furthermore, by integrating physicochemical properties, ESM-1b can encode crucial biochemical information pertinent to RNA sequence analysis.

The RNAformer [220] is a deep learning architecture that is inspired by the renowned protein structure prediction algorithm, Alphafold. It is designed for the purpose of predicting RNA secondary structures. The RNAformer utilizes a data-driven approach to make predictions. It makes use of a 2D latent space representation and axial attention mechanisms to capture long-range interactions and dependencies within the RNA sequence. The model aims to learn the underlying biophysical dynamics of the folding process without relying on additional information like multiple sequence alignments (MSAs). The RNAformer is composed of multiple RNAformer blocks, each incorporating row-wise and column-wise axial attention layers, followed by a transition convolutional layer. The axial attention mechanism enables the model to efficiently process higher-dimensional data and capture dependencies along each axis independently. The transition convolutional layer assists in modeling local structures like stem-loops. Residual connections, pre-layer normalization, and dropout are applied to enhance training and prediction accuracy. The RNAformer makes use of a 2D latent space representation of the RNA sequence which allows the model to capture the pairing between nucleotides and leverage the advantages of deep learning methods. The axial attention mechanism in the RNAformer allows for efficient processing of long-range interactions and dependencies within the RNA sequence. It helps the model capture the structural characteristics of the RNA secondary structure. RNAformer has achieved state-of-the-art accuracy on benchmark datasets for RNA secondary structure prediction. It outperforms previous de novo prediction methods and performs on par with current homology modeling methods, demonstrating its effectiveness in capturing the folding dynamics of RNA.

BigBird [115] is an innovative deep learning model that showcases unique features designed for efficient learning of nucleotide embeddings. It treats each base of the RNA sequence as a token and always includes a [CLS] token at the beginning of every sequence. Additionally, it employs the MLM pre-training framework, during which a portion of the tokens are replaced with [MASK] tokens. To reduce computational complexity and memory requirements, BigBird utilizes a sparse attention mechanism that incorporates three distinct attention components: random attention, window local attention, and global attention. In random attention, each query block randomly selects a specified number of key blocks to attend to, introducing a degree of randomness to capture diverse dependencies within the RNA sequence. Window local attention ensures that each query block attends to a specific window of key blocks, which is centered around the query block, and all query blocks attend to key blocks within the window range. This component is useful for capturing local dependencies and structural characteristics within the RNA sequence. Global attention allows one query and key block to attend to every other block, which helps capture the global context and dependencies across the entire RNA sequence. For handling long sequences, BigBird utilizes a sampling subsequence approach, dividing the long sequence into smaller subsequences or windows, enabling the model to process and attend to smaller chunks of data at a time. This approach helps handle longer sequences efficiently and avoids memory constraints. The model generates RNA sequence representations by utilizing the output embedding of the [CLS] tokens. This provides a concise and informative representation of each subsequence. BigBird generates different types of embeddings for each RNA sequence, such as Bigbird256, and Bigbird768 embeddings. These embeddings capture different levels of information and can be used for various downstream tasks. In summary, BigBird's sparse attention mechanism, efficient handling of long RNA sequences, and multiple embeddings make it a powerful model for learning nucleotide embeddings and analyzing RNA sequences effectively.

LongFormer [263] presents several innovative components and features that enable it to process lengthy sequences efficiently and learn effective nucleotide embeddings. LongFormer addresses the limitation of quadratic attention scaling in traditional Transformers by introducing an attention mechanism that scales linearly with the sequence length. This allows LongFormer to handle long sequences with thousands of tokens or more. LongFormer incorporates a local windowed attention mechanism, which attends to a specific window of tokens within the sequence. This local attention captures contextual information and dependencies within the windowed region. LongFormer combines the local windowed attention with a task-motivated global attention mechanism. The global attention allows the model to capture broader context and dependencies across the entire sequence, enhancing its understanding of the nucleotide sequence. LongFormer can be pretrained using a masked language modeling (MLM) objective, similar to other Transformer models where some tokens are masked in the input sequence, and the model is trained to predict the original values of these masked tokens. This pre-training process helps LongFormer learn representations that capture the underlying patterns and dependencies in the nucleotide sequence. Pre-trained models can then be fine-tuned on specific downstream tasks, such as enhancer RNA identification and promoter RNA identification. LongFormer is specifically designed to handle long sequences efficiently. It adopts strategies like sampling sub-sequences and incorporating global and local attention mechanisms to process lengthy nucleotide sequences effectively. It can effectively capture cross-partition information without the need for complex architectures or partitioning the sequences into smaller sequences. LongFormer also introduces a variant called Longformer-Encoder-Decoder (LED), which follows an encoder-decoder architecture similar to the original Transformer model. LED is suitable for sequence-to-sequence tasks like gene prediction, RNA splicing, genetic variant detection, motif detection, allowing LongFormer to scale efficiently for such tasks. By incorporating these unique components and features, LongFormer can effectively learn nucleotide embeddings by capturing dependencies, contextual information, and long-range dependencies within the sequence.

7.3. Machine and deep learning predictors

Machine learning and deep learning algorithms rely on statistical vectors to identify useful patterns for particular sequence analysis tasks. A thorough review of 172 studies indicates that, 8 language models and 16 word embedding have been employed to generate statistical vectors of genetic sequences to feed 44 unique algorithms for 47 distinct RNA sequence analysis tasks. From 44 algorithm, 13 machine learning algorithms include Support Vector Machine (SVM) [169,200,212], Naive Bayes (NB) [236], Logistic Regression [261,177], ElasticNet Regression Model (ERM) [202], Rotation Forest Algorithm [258], Random Forest [170], Xtreme Gradient Boosting [172], Gradient Boosting Decision Trees (GBDT) [280], Deep Forest [141], AdaBoost [163], CatBoost [165], and MultiLayer Perceptron (MLP) [249,167]. Furthermore, 9 deep learning algorithms include Convolutional Neural Network [173], Graph Neural Network [216], Graph Convolutional Network [147], Long Short Term Memory [18], Bidirectional Long Short Term Memory [176], Gated Recurrent Unit [17], Neural Network Regression Model [262], Adaptive subspace learning model [257], and Deep Neural Network [185]. Similarly, 5 algorithms including GPT-3 [242], ESM-1b [172], Heterogeneous Graph Transformer (HGT) [128], BERT [131], and Transformer [137] belong to language modeling algorithms. Besides machine and deep learning algorithms, 7 algorithms have utilized two or more machine learning algorithm namely CatBoost + ET + LightGBM + RF + XGBoost + LR [162], GBDT + LR [144], SVM + RF + XGBoost + GBDT + AdaBoost + MLP [163], SVM + LogR [286], XGBoost + LightGBM + RF + ET + CatBoost [165], SVM + Ridge Regression [237], and LightGBM + SVM + LR [204], 7 algorithms have employed more than 1 deep learning algorithm such as CNN + RNN [228,122], LSTM + CNN [205], CNN + DNN [203], BiLSTM + CNN [13,111,116,180,179,251,287], CNN + GRU [17], CNN + BiGRU [196], and BiLSTM + LSTM [250] and 3 algorithms reap benefits of both machine and deep learning algorithms namely BiLSTM + LogR [181], CNN + GuasianNB [145], and AdaBoost + CNN + LightGBM [157]. This organized prediction approach simplifies the selection of the most appropriate method for a specific RNA sequence analysis task. Additionally, it enables comparative analyses both within and across various algorithm categories and facilitates in informed decision-making and assessment of algorithm strengths and weaknesses. Let's take a brief look into the functional paradigms of 35 different algorithm.

From machine learning algorithms, Support Vector Machine (SVM) [169,200,212] algorithm works by finding the optimal hyperplane that maximizes the margin between different classes. For non-linear classification tasks, SVMs use kernel functions to map data into higher-dimensional spaces where a linear separation is possible. SVMs are particularly effective in high-dimensional spaces and can handle cases where the number of dimensions exceeds the number of samples. They are versatile and robust, performing well even with non-linearly separable data by using soft margins. However, SVMs can be computationally intensive, requiring significant time and memory resources, especially with large datasets. They also require careful tuning of parameters such as the kernel type and regularization parameter and do not inherently provide probabilistic outputs for their predictions.

Naïve Bayes (NB) [236] algorithm is fundamentally based on Bayes' theorem, which calculates the posterior probability of a class given a set of features. This method operates under the "naïve" assumption that features are conditionally independent given the class label, which simplifies the computation. One of the main advantages of Naïve Bayes is its simplicity and computational efficiency, making it particularly suitable for real-time applications. It scales well with large datasets and can effectively handle irrelevant features. However, the independence assumption often does not hold true in real-world scenarios, which can negatively impact performance. Additionally, Naïve Bayes may be less effective for complex relationships between features and class labels and is sensitive to the quality of the data. Logistic regression (LogR) [261,177] algorithm computes probability of a specific class or event occurring and translates this probability into binary outcomes using a logistic function. The main advantage of logistic regression lies in its simplicity and interpretability, making it easy to implement and providing insights into the relationship between features and the outcome variable. It is also computationally efficient and can handle large datasets with numerous features, offering probabilistic outputs that aid in making informed decisions. However, logistic regression assumes a linear relationship between independent variables and the log-odds of the dependent variable, which may not always hold true. This assumption limits its flexibility in capturing complex, non-linear relationships. Additionally, logistic regression can be prone to overfitting, especially with high-dimensional data, and is sensitive to outliers. It also performs best with balanced datasets, and significant class imbalances may require additional techniques to maintain performance.

Elastic-Net regression [202] is a regularization based algorithm that combines penalties of both Lasso and Ridge regression methods in order to address some of their limitations. In foundational linear regression algorithm, the goal is to find the best-fitting line that predicts the relationship between the independent variables and the dependent variable. However, when there are large number of independent variables and multi-collinearity is present, ordinary least squares regression can lead to overfitting and poor performance. The working paradigm of Elastic-Net regression involves adding two penalty terms to the standard regression equation: one that is proportional to the absolute value of the coefficients (L1 penalty) and one that is proportional to the square of the coefficients (L2 penalty). This combination allows Elastic-Net regression to effectively select a subset of important variables and also handles multicollinearity. One advantage of Elastic-Net regression is that it can handle highly correlated variables better than Lasso regression, which tends to select only one variable from a group of correlated variables. This makes Elastic-Net regression a more robust model for real-world data sets where multicollinearity is common. However, one disadvantage of Elastic-Net regression is that it introduces two tuning parameters that need to be optimized through cross-validation, which can make the model more complex and computationally intensive compared to simpler regression methods.

In tree based algorithms paradigm, in RNA sequence analysis landscape, foundational decision tree algorithm paradigm is extended to develop 8 algorithms including Rotation Forest algorithm [258], Random Forest [119,256,146,154,170,187,197,252], Deep Forest (DF) [141], Xtreme Gradient Boosting (XGBoost) [120,259,172,183,189], Gradient Boosting Decision Trees (GBDT) [280], AdaBoost [187], and CatBoost [186,118]. Rotation Forest algorithm builds multiple decision trees using different subsets of features and subsequently combines their predictions. The core idea is to apply Principal Component Analysis (PCA) to each subset of features before training each individual tree. This process ensures that the diversity among the trees is maximized, which is crucial for the strength of ensemble methods. Its advantages include enhanced diversity, improved accuracy, robustness to overfitting, and effective feature utilization. However, it also has disadvantages such as computational complexity, the need for careful hyperparameter tuning, reduced interpretability, and high memory usage. Random Forest (RF) algorithm [119] is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes as the prediction. RF is known for its robustness to overfitting, feature importance estimation, and ability to handle high-dimensional data with ease [119]. However, RF may not perform as well when dealing with imbalanced datasets or when there are many irrelevant features present in the data. Deep Forest (DF) [141] algorithm is another ensemble learning method that utilizes a cascade structure of multiple random forests to make predictions. DFs are capable of learning hierarchical representations of data and can capture complex patterns in high-dimensional spaces effectively [141]. Nonetheless, the main drawback of DF lies in its computational complexity and the need for substantial computational resources, which can limit its practicality in large-scale RNA sequence analysis projects.

Gradient Boosting [162] minimizes a specified loss function by using gradient descent to determine the optimal direction and step size for model improvement. In Gradient Boosting, each new model is trained to correct the residual errors of the combined ensemble of all previous models. This iterative process continues until further improvements are minimal, and effectively reduces both bias and variance. XGBoost (Extreme Gradient Boosting) [162] is an extension of Gradient Boosting that emphasizes speed and performance. Its core working difference lies in its use of regularized model formalization to control overfitting. XGBoost incorporates advanced features such as tree pruning, which eliminates unnecessary branches to reduce overfitting, and supports parallel processing for faster computations. It also includes both L1 and L2 regularization to manage model complexity. Additionally, XGBoost efficiently handles missing values by learning the best path for dealing with them during the training process, ensuring robust and accurate predictions. Gradient Boosting Decision Trees (GBDT) [280] algorithm operates on the principle of sequentially building models, each one correcting the errors of its predecessor. The primary functional difference of GBDT lies in its use of gradient descent to optimize a chosen loss function, such as mean squared error or logistic loss. Each new tree in GBDT is trained to fit the residuals (errors) of the previous tree, thereby incrementally improving the model's accuracy. GBDT also employs techniques like shrinkage (learning rate) to control the contribution of each tree and sub-sampling to prevent overfitting by training on different subsets of the data.

AdaBoost (Adaptive Boosting) [163] sets itself apart by focusing on combining multiple weak learners, typically decision stumps, to form a strong algorithm. Its working mechanism involves adjusting the weights of instances based on their prediction results. Wrong predicted instances are given higher weights which makes them more prominent in subsequent iterations, while correctly predicted instances are given lower weights. This adaptive process ensures that the model focuses on the harder-to-classify instances, thereby improving overall accuracy. AdaBoost's unique approach to handling weights and concentrating on difficult cases makes it particularly effective for scenarios where simple models need to be boosted into powerful ensembles. CatBoost (Categorical Boosting) [165] is designed specifically for handling categorical data efficiently, distinguishing it from other boosting algorithms. Its primary functional advantage is its ability to process categorical features without extensive preprocessing like one-hot encoding. CatBoost uses an innovative technique called ordered boosting, which maintains a strict ordering of training examples to reduce overfitting. Moreover, it builds symmetric trees, ensuring balanced and faster predictions. CatBoost also has built-in support for handling missing values seamlessly during training, making it highly suitable for real-world datasets that often include categorical and missing data.

Multi-Layer Perceptron (MLP) [8,112,140,128,148–150,142,155,160,159,167,178,254,255,193,210,209,282,10,219,220,229,242] algorithm consists of multiple layers of nodes or neurons. Each node acts as a perceptron, utilizing a nonlinear activation function. MLPs are trained using backpropagation, a method that adjusts the weights of the connections to minimize the error. One of the key strengths of MLPs is their ability to approximate any continuous function, making them powerful tools for complex tasks. They are flexible and capable of handling a wide variety of problems, from classification to regression. However, training MLPs can require substantial computational resources and time, particularly with large datasets and deep architectures. MLPs are also prone to overfitting, necessitating the use of regularization techniques. The process of tuning hyperparameters, such as the number of layers, neurons per layer, and learning rate, is critical and can be challenging.

Among all categories, deep learning algorithms are most extensively used for efficient RNA sequence analysis. A total of 9 deep learning algorithms are most commonly used by scientific community for RNA sequence analysis. Convolutional Neural Network (CNN) [108,121,151,152,161,158,173,7,190,191,201,198,199,207,211,215,230,227,11,218,232,252,238] is designed to process structured grid-like data, such as images. In RNA sequence analysis, CNNs can be applied to RNA sequence analysis tasks to capture spatial dependencies in data. They are effective for tasks that require feature hierarchies and translation invariance [190,191]. However, CNNs may struggle with capturing long-range dependencies in sequences, which can be crucial in RNA analysis where distant nucleotides may interact. Graph Neural Network (GNN) [240,166,260,216] is a type of neural network designed to operate on graph-structured data. GNNs are suitable for tasks involving relational data, such as molecular structures that makes them applicable to RNA sequence analysis for tasks like clustering [216]. GNNs can effectively capture dependencies between nodes in a graph and are capable of learning representations that incorporate both local and global information [216]. However, GNNs may encounter challenges in efficiently scaling to large graphs, and interpreting the learned representations in GNNs can be complex, limiting their interpretability. Graph Convolutional Network (GCN) [147,153,143,156,184,171] is a type of neural network designed to operate on graph-structured data. GCNs can leverage graph structures to learn representations of nodes and edges, enabling tasks like node classification and link prediction in RNA sequences [184]. However, GCNs may require meticulous graph construction and preprocessing, and they can be computationally intensive, especially for large graphs, which can hinder their scalability. Hypergraph convolutional Networks (HGCN) [129] are extended GNNs which are designed for hypergraphs to capture complex relationships. These network captures local and global information of hyperedges and their connected node which can used in various tasks including miRNA-disease association prediction [129]. HGCN offers significant advantages in modeling complex relationships and capture higher order relationships but requires higher computational resources to aggregate information through hyperedges.

Long Short-Term Memory (LSTM) [127,18] is designed to overcome the vanishing gradient problem in traditional RNNs by introducing a memory cell that can maintain information over long sequences. It consists of three gates: input gate, forget gate, and output gate, that control the flow of information. LSTM can capture long-term dependencies in sequences and is suitable for tasks requiring memory of past information. LSTM is more complex and computationally expensive compared to GRU, making it slower to train and deploy. Bidirectional Long Short-Term Memory (BiLSTM) [109,126,248,176,279,188,225,253] is an extension of LSTM that processes sequences in both forward and backward directions. BiLSTMs are advantageous in RNA sequence analysis for tasks where contextual information from both past and future is essential [176]. BiLSTMs can capture dependencies in both directions and are effective in tasks requiring bidirectional context understanding [176]. However, BiLSTMs may be computationally intensive due to processing sequences in two directions, which can impact their training and inference speed. Gated Recurrent Unit (GRU) [17] is a simplified version of LSTM with only two gates - reset gate and update gate. It is computationally more efficient than LSTM as it is faster to train and may perform better on smaller datasets due to its simpler architecture. GRU may struggle with capturing long-term dependencies in sequences, leading to performance degradation on tasks requiring memory of distant information.

The Neural Network Regression model [262] is a precisely deep neural network based on multiple layers. It passes the input features vectors through two hidden layers with ReLU activation functions, which help capture complex, non-linear relationships. To mitigate overfitting, a dropout layer with a 0.02 probability is used between the hidden layers. The output layer consists of a single neuron with a sigmoid activation function, which generates a probability score indicating the likelihood of an association. The model employs binary cross-entropy loss to measure the error between predicted probabilities and actual labels, and it is optimized using

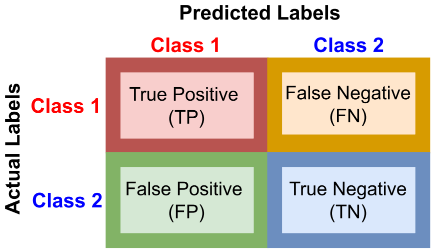

Fig. 6. Overview of Confusion Matrix.

the Adam optimizer. Precisely deep neural network regression model heavily relies on quality of input feature vectors and it may overfit easily.

The Adaptive Subspace Learning Predictor (NSL2CD) [257] is designed to discover hidden relationships between circular RNAs (circRNAs) and diseases by integrating multiple data sources. The core of its functionality lies in the use of projection matrices, which transform high-dimensional circRNA and disease features into a shared, lower-dimensional latent space. This transformation is achieved by multiplying each feature matrix with its corresponding projection matrix, thereby aligning the different types of data. The model then minimizes the regression error to ensure that the transformed features in the latent space closely resemble the original data. Regularization techniques like L1, 2-norm and graph Laplacian regularization are employed to maintain model simplicity and preserve the geometric structure of the data. An iterative optimization process fine-tunes the model parameters, gradually improving the accuracy of the projections and predictions. The final output is a predicted association matrix that highlights potential relationships between circRNAs and diseases. The process of projecting high-dimensional data into a lower-dimensional latent space can sometimes lead to the loss of important information, potentially affecting the model's accuracy. Additionally, the need for multiple data sources means that the model's performance is highly dependent on the quality and completeness of the input data. Furthermore, deep neural network (DNN) [185] algorithm is used for circRNA-miRNA association prediction. DNN algorithm is a multi-layer neural network designed to learn complex patterns from the feature representations of circRNA and miRNA sequences. It processes the input feature vectors through several hidden layers, applies non-linear transformations (ReLU), and outputs the probability of a circRNA-miRNA association.

For different RNA sequence analysis tasks, 5 contemporary language models namely GPT-3 [242], ESM-1b [172], Heterogeneous Graph transformer (HGT) [130,285], BERT [107,114,123,125,124,131,132,139,174,278,16,281,213,214,226,222,223,234,239,241, 284], and Transformer [276,137,133,138,135,136,175,182,19,195,203,221,224,217,231,233,277,235,26,27] have been used in two different settings. In first setting, the addition of classification layers to these language models adapts the general-purpose language models to specific classification tasks by learning to map the rich contextual embeddings to the desired output classes. In second setting, rich contextual embeddings of these language models are passed to standalone machine learning algorithms, deep learning algorithms, ensemble or hybrid algorithms for accurate classification of RNA sequences.

8. Uncovering evaluation measures for RNA sequence analysis predictive pipelines

Performance evaluation of AI-driven predictive pipelines for RNA sequence analysis undergoes through two experimental settings: 1) Train-test split [288,289], and 2) k-fold cross-validation [290,291]. In train-test split, data is splitted into two sets namely train and test set. In this setting, usually 70-80% of data is used for training and remaining 20-30% for testing. To prevent overfitting issues, a subset of training data, also known as validation set, is used to fine-tunes predictor hyperparameters [292]. On the other hand, k-fold cross-validation splits data into k-equal folds. Since k-fold cross-validation is an iterative process, another fold is reserved for testing while remaining k-1 folds are used for training. In this way, predictive pipeline is trained and tested for k-times on whole data.

AI-driven genomic sequence analysis tasks belong to five different types namely: 1) binary classification [293], 2) multi-class classification [293,20], 3) multi-label classification [294], 4) regression [295,296], 5) clustering [295,296]. Based on the nature of task, there are multiple evaluation measures for each type. This section provides an in-depth understanding of evaluation measures for binary/multi-class, multi-label, regression and clustering.

8.1. Binary/multi-class classification evaluation measures

In binary/multi-class classification, predicted label can either be positive or negative. In order to evaluate the performance of binary/multi-class predictive pipeline, precision (P) [297], recall (R) [293], F1-score (F1) [297], accuracy (Acc) [293], specificity (SP) [293], and Matthews correlation coefficient (MCC) [[293]] are most commonly used evaluation measure. These measures are calculated using confusion matrix. Fig. 6 depicts confusion matrix, comprised of four different entities: 1) True Positive (TP), 2) True Negative (TN), 3) False Positive (FP), 4) False Negative.(FN).

Among four entities, TP and TN specify the correct predictions of positive and negative classes respectively. However, FP and FN specify incorrect predictions of positive and negative classes respectively. Equation (6) embodies mathematical expressions for these evaluation measures.

$$f(x)-balanced=\begin{cases} Acc=\frac{TP+TN}{TP+FP+TN+FN} \\ P=\frac{TP}{TP+FP} \\ R=\frac{TP}{TP+FN} \\ F1=\frac{2*P*R}{P+R} \\ SP=\frac{TN}{TN+FP} \\ MCC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \tag{6}$$

These measures are commonly used for balance datasets. However, variants of these measures including weighted, micro, and macro are used for imbalanced datasets. To compensate for class imbalance problem, weighted precision (Wei-P) [298] is a ratio that computes sum of precision of each class weighted by its size by total number of weights for all classes. Precision of each class is proportion of positive prediction of the specific class, while relative weight assigns a weight score to each class based on the proportion in data. Similarly, weighted-recall (Wei-R) [298] and weighted F1-score (Wei-F1) [299] are computed by assigning weights of recall and F1-score to each class. Macro precision [300] is computed by calculating the precision of individual classes and then averaging these precisions. In the same way, Macro recall (Mac-R) [300] and Macro F1-score (Mac-F1) [300] are calculated by taking the average of all classes. Micro-precision (Mic-P) [300] calculates the proportion of all true positive instances by total number of predicted positive instances for all classes. In the same manner, Micro recall (Mic-R) [300] and Micro F1-score (Mic-F1) [300] calculate the score for all classes. Equation (7) signifies mathematical expressions for these evaluation measures.

$$f(x)-imbalanced=\begin{cases} Wei-P=\frac{\sum_{z=1}^{n}P^z.w^z}{\sum_{z=1}^{n}w^z} \\ Wei-R=\frac{\sum_{z=1}^{n}R^z.w^z}{\sum_{z=1}^{n}w^z} \\ Wei-F1=\frac{\sum_{z=1}^{n}F1-score^z.w^z}{\sum_{z=1}^{n}w^z} \\ Mac-P=\frac{1}{n}\sum_{z=1}^{n}P^z \\ Mac-R=\frac{1}{n}\sum_{z=1}^{n}R^z \\ Mac-F1=\frac{1}{n}\sum_{z=1}^{n}F1-Score^z \\ Mic-P=\frac{\sum_{z=1}^{n}TP^z}{\sum_{z=1}^{n}(TP^z+FP^z)} \\ Mic-R=\frac{\sum_{z=1}^{n}TP^z}{\sum_{z=1}^{n}(TP^z+FN^z)} \\ Mic-F1=\frac{\sum_{z=1}^{n}2.TP^z}{\sum_{z=1}^{n}(2.TP^z+FP^z+FN^z)} \end{cases} \tag{7}$$

Here, for class z, TP^z, FP^z, FN^z represents true positive, false positive, and false negative factors respectively. P^z, R^z, $F1-score^z$ denote precision, recall, and F1-score of class z. w^z is relative weight of class z and z is z^{th} class for n number of classes.

8.2. Multi-label classification

Performance evaluation of multi-label classification predictive pipelines is relatively arduous compared to binary and multi-class predictive pipelines. In multi-label predictive pipelines, instances have more than one label at a time. Therefore, among predicted labels, some labels can be correct, some can be incorrect, all can be correct or incorrect. Because of this partial correctness, it becomes difficult to evaluate multi-label predictive pipelines [301]. To cope with this issue, different evaluation measures have been introduced including precision (P) [294], recall (R) [294], accuracy (Acc) [294], and hamming loss (HL) [294]. Equation (8) represents mathematical expressions for these evaluation measures.

$$f(x)-multi-label=\begin{cases} P=\frac{1}{M}\sum_{z=1}^{M}\frac{|A^z\wedge P^z|}{|P^z|} \\ R=\frac{1}{M}\sum_{z=1}^{M}\frac{|A^z\wedge P^z|}{|A^z|} \\ Acc=\frac{1}{M}\sum_{z=1}^{M}\left|\frac{A^z\wedge P^z}{A^z\vee P^z}\right| \\ F1=\frac{1}{M}\sum_{z=1}^{M}\frac{2*|P(m^z)*R(m^z)|}{|P(m^z)+R(m^z)|} \\ HL=\frac{1}{Ml}\sum_{z=1}^{M}\sum_{k=1}^{l}\left[|(A_k{}^z\neq P_k{}^z)|\right] \end{cases} \tag{8}$$

M denotes the total number of instances, m^z represents z^{th} instance from M instances, actual and predicted class labels are denoted by A^z and P^z for m^z instance respectively. Instance length and class index are indicated by l and k respectively, $\vee$ and $\wedge$ signifies logical OR and AND operators. For imbalanced datasets, evaluation measures incorporate weighted, micro and macro variants. After, a thorough analysis of existing literature, it is inferred that most commonly used evaluation measures in AI-driven predictive pipelines for genomic sequence analysis are precision, recall, accuracy, specificity, sensitivity, MCC, and F1-score [290,294].

8.3. Regression evaluation measures

There is a fundamental difference between regression and classification tasks. Regression task predicts continuous values instead of class labels. Thus researchers introduced variety of evaluation measures to evaluate performance of regression-based predictive pipelines. These measures include Mean Square Error (MSE) [302], Root Mean Square Error (RMSE) [302], Mean Absolute Error (MAE) [303], Mean Absolute Percentage Error (MAPE) [304], Mean Bias Error (MBE) [303], R^2 Score [303], relative Root Mean Square Error (rRMSE) [305], relative Mean Square Error (rMSE) [305], relative Mean Absolute Error (rMAE) [305], and relative Mean Bias Error (rMBE) [305].

MAE computes absolute difference between predicted and actual values and then takes the average for all number of instances [303]. Where as, MSE calculates the average error by taking squared differences between predicted and actual values [302]. While, RMSE takes square root of MSE [302], and MBE calculates the average bias of the predictor pipeline by taking difference between actual and predicted values [303]. However, MAPE calculates percentage first using absolute difference between the actual and predicted values by actual values and then averages them [304]. Besides this, R^2 Score is a statistical measure that analyzes the relationship strength between the dependent and independent variables. It uses the squared difference of predicted and actual values by square difference of actual and average of actual values. [303]. The minimum error scores of MAE, MSE, MBE, and MAPE indicate that predictor pipeline will perform well while high score or R^2-squared signifies pipeline robustness. However, these error scores calculate N number of instances average error value.

Relative performance evaluation can enhance quality of performance assessments by diminishing data noise. In this evaluation, error score is calculated in percentage, ratio of particular error score by the average of actual values. Relative versions of these evaluation measures including rMAE, rMSE, rMBE, and rRMSE validate the pipeline performance relative to the average of the actual baseline. These measures are helpful for pipeline robustness analysis when tested on varying datasets. Equation (9) embodies mathematical expressions for these evaluation measures.

$$f(x) - regression = \begin{cases} MAE = \frac{1}{N} \sum_{z=1}^{N} |P^z - A^z| \\ MSE = \frac{1}{N} \sum_{z=1}^{N} (A^z - P^z)^2 \\ RMSE = \sqrt{\frac{1}{N} \sum_{z=1}^{N} (A^z - P^z)^2} \\ MBE = \frac{1}{N} \sum_{z=1}^{N} (P^z - A^z) \\ MAPE = \frac{1}{n} \sum_{z=1}^{N} \left| \frac{P^z - A^z}{A^z} \right| \times 100 \\ R^2 Squared = 1 - \frac{\sum_{z=1}^{N} (P-A)^2}{\sum_{z=1}^{N} (A - avg(A))^2} \\ rMAE = \frac{MAE}{\bar{A}} \times 100 \\ rMSE = \frac{MSE}{\bar{A}} \times 100 \\ rMBE = \frac{MBE}{\bar{A}} \times 100 \\ rRMSE = \frac{RMSE}{\bar{A}} \times 100 \end{cases} \tag{9}$$

Here, N is the instances, $\bar{A}$ represents average of total actual values, P^z and A^z are predicted, and actual values of instance z.

8.4. Clustering evaluation measures

Clustering tasks are different as compared to classification and regression. Clustering tasks aim to group data points which share common features. These tasks are based on unsupervised learning methods, that make clusters based on inherited features, similarity score, and data structure rather than labeled data [306]. New data points are assigned to that cluster which have maximum similarity, mutual information, and minimum intra-cluster distance. Different evaluation measures have been adopted to validate clustering-based predictive pipeline performance such as silhouette score (SS) [307], accuracy (Acc) [308], Dunn index (DI) [309], normalized mutual information (NMI) [308] and davies-Bouldin index (DBI) [310].

Accuracy is the ratio of correct predictions of instances to total instances of the data with calculating maximum match of predicted clusters [308]. NMI calculates an information gain score that computes mutual information by taking a mean of predicted and actual cluster entropies [308]. SS calculates the similarity score of an instance to its own cluster and dissimilarity between clusters [307]. DI measures proportion of similarity score by focusing on minimum distance within clusters to maximum distance in intra-class cluster [309]. DBI focuses on calculating the average similarity score by taking maximum ratio of average distance within the cluster to the distance between centroids [310]. SS calculates variance in cluster data while DBI evaluates how clusters are well segregated and compact. Minimum score of DBI is good for cluster-based predictive pipelines. However, DI computes how clusters are well separated and tightly bound to internal cluster structure and maximum score is good for cluster-based predictive pipeline. Equation (10) illustrates mathematical expressions for these evaluation measures.

$$f(x) - clustering = \begin{cases} Acc = \underset{max}{m} \frac{\sum_{z=1}^{n} 1\{y_z = m(c_z)\}}{n} \\ NMI = \frac{I(y_z, c_z)}{\frac{1}{2}\left[E(y_z) + E(c_z)\right]} \\ SS = \frac{min\{d(y_z)\} - a(y_z)}{max\{min\{d(y_z)\}, a(y_z)\}} \\ DBI = \frac{1}{n} \sum_{z=1}^{n} \underset{k \neq z}{max}\left(\frac{\bar{S}_z + \bar{S}_k}{d(c_z, c_k)}\right) \\ DI = \frac{min_{1 \leq z < k \leq n} d(c_z, c_k)}{max_{1 \leq l \leq n} d'(c)} \end{cases} \tag{10}$$

Here m is a mapping function, y^z is predicted cluster, among n clusters c^z and c^k refers the z^{th} and k^{th} clusters respectively. $I(y_z, c_z)$ signifies mutual information while $E(y_z)$ and $E(c_z)$ are predicted and actual cluster entropies respectively. $d(y_z)$ and $a(y_z)$ indicate average distance to other cluster centroids and in that clusters respectively. $d(c_z, c_k)$ represents inter-cluster distance while S_z and S_k denote the mean distance from all observations in cluster z and mean distance for median cluster k respectively.

9. Open-source RNA sequence analysis predictive pipelines

The public availability of source codes for predictive models, pretrained language models, and word embeddings significantly accelerate research efforts by eliminating the need to start from scratch. By leveraging existing predictive models and incorporating new strategies, researchers can develop new applications which result improved performance. Additionally, public access to these codes ensures transparency, reliability, and reproducibility in research. To benefit the research community and develop more precise, robust, and efficient AI-driven RNA sequence analysis predictive pipelines, this section provides an in-depth summary of open-source predictive pipelines developed using two contemporary representation learning methods namely word embeddings and large language models for 47 distinct RNA sequence analysis tasks. Our analysis reveals that, from 58 existing RNA sequence analysis studies, only 20 studies have made their predictive pipelines source codes publicly available for word embeddings AI applications. In addition, out of 70 existing RNA sequence analysis studies based on large language models, source code of only 45 studies are publicly available. Tables 4 and 5 provide information on open-source codes for RNA sequence analysis applications using word embeddings and large language models, respectively. These tables also summarize the representation learning methods, machine/deep learning predictors used, and include links to the respective source codes.

Table 4 summarizes these predictive pipelines in form of their respective representation learning approaches, machine or deep learning predictors, target RNA sequence analysis tasks, and links of source codes. A closer examination of Table 4 shows that a total of 6 unique word embedding approaches namely Word2Vec, GloVe, Transformer, LINE, Node2Vec, SDNE, and SVD have been used to develop 20 predictive pipelines are developed for 14 distinct RNA sequence analysis tasks. These tasks are sncRNA Prediction, cirRNA Prediction, lncRNA Prediction, RNA Sub-cellular Localization Prediction, RNA Functions Prediction, RNA-protein binding sites identification, RNA-protein interaction prediction, RNA-RNA Associations prediction, 5mC-Methyl Cytosine Modification Prediction, Methylation Modification Prediction, RNA-Disease Associations Prediction, RNA-Gene Association Prediction, miRNA Target Prediction, 16S rRNA Taxonomic Classification.

Specifically, a total of 3 open source RNA-protein binding sites identification studies have utilized Word2Vec representation learning along with 3 deep learning architectures namely CNN, CNN+BiLSTM and LSTM+BiLSTM. Moreover, for coding RNA-Protein interaction prediction, two open source predictive pipelines have utilized two unique word embeddings (Word2vec, Node2Vec) along with GNN and GCN classifiers. Moreover, a total of 5 open-source RNA-disease association prediction studies make use of 5 unique word embedding approaches namely Word2Vec, Node2Vec, GloVe, SNDE, and SVD along with RF, BiLSTM, XGBoost, and DBN. For 5mC-methyl cytosine modification prediction, 1 open source predictive pipeline make use of Node2Vec and 1 predictive pipeline make use of Word2vec embeddings along with CNN classifier. In addition, open-source predictive pipelines of micro RNA target prediction, 16S rnRNA taxonomic classification, small non coding RNA identification, circular RNA identification, and RNA subcellular localization prediction make use of Word2Vec embedding with 5 unique classifiers namely NB, GNN, BiLSTM, and CNN+BiLSTM. For RNA function prediction and long non-coding RNA identification, GloVe and LINE word embeddings along with hybrid CNN+BiLSTM and deep hierarchical model are used, respectively.

Table 5 provides a comprehensive summary of 45 open-source predictive pipelines based on large language models developed for various RNA sequence analysis tasks. Analysis of Table 5 reveals that these pipelines utilize five distinct large language models: Transformer, BERT, ESM-1b, Heterogeneous Graph Transformer, and GPT, along with 5 unique classifiers including MLP, CNN, XGBoost, BiLSTM, and Hybrid (CNN + BiLSTM + MLP). Collectively, these 45 predictive models cover 24 different RNA sequence analysis tasks. These tasks include RNA-Protein Binding Affinity Prediction, Cell-Specific Gene Regulatory Networks Prediction, Single-Cell Multi-Omics Analysis, mRNA Degradation Prediction, RNA-Disease Association Prediction, Enhancer RNA Identification, 6mA-Methyl Adenosine Modification Prediction, RNA Subcellular Localization Prediction, Pre-miRNA Prediction, Promoter Identification, RNA Cluster Analysis, RNA-Seq Coverage Prediction, RNA Structure Prediction, Spatial Gene Expression Analysis, CRISPR/Cas9 single guide RNA Prediction, microRNA- Target Prediction, RNA-Protein Interaction Prediction, RNA Splicing Sites Prediction, RNA Function and Structure Prediction, Long non coding RNA Prediction, miRNA Target Prediction, RNA-Protein Binding Sites Prediction,

Table 4
Summary of open-source word embedding based RNA Sequence Analysis models in existing studies.

Author, Year [ref]	Task	Embedding Approach	Classifier	Source Code
Deng et al., 2023 [109]	Small non coding RNA Prediction	Word2Vec	BiLSTM	https://github.com/YinggggJ/ABLNCPP
Chaabane et al., 2020 [111]	Circular RNA Prediction	Word2Vec	CNN + BiLSTM	https://github.com/UofLBioinformatics/circDeep
Liu et al., 2019 [116]	Long non coding RNA Prediction	GloVe	BiLSTM + CNN	https://github.com/www-bioinfo-org/CNCI
Zeng et al., 2023 [19]	RNA Sub-cellular Localization Prediction	Word2Vec	Transformer	https://github.com/CSUBioGroup/LncLocFormer
Wang et al., 2019 [219]	RNA Functions Prediction	LINE	Deep Hierarchical Model	https://github.com/JChander/DeepMiR2GO
Wang et al., 2021 [250]	RNA-Protein Binding Sites Identification	Word2Vec	BiLSTM+LSTM	https://github.com/wzf171/CRPBsites
Deng et al., 2020 [179]	RNA-Protein Binding Sites Identification	Word2Vec	CNN+BiLSTM	https://github.com/youzhiliu/DeepRKE/
Xiaoyong et al., 2018 [7]	RNA-Protein Binding Sites Identification	Word2Vec	CNN	https://github.com/xypan1232/iDeepV
Han et al., 2023 [166]	Coding RNA-Protein Interaction Prediction	Node2Vec	GNN	https://github.com/nwpu-903PR/ncRPI-LGAT
Shen et al., 2021 [260]	Coding RNA-Protein Interaction Prediction	Node2Vec	GNN	https://github.com/AshuiRUA/NPI-GNN
Zhao et al., 2022 [184]	Coding RNA-Protein Interaction Prediction	Word2Vec	GCN	https://github.com/zhaozhiya-20/SEBGLMA-semantic-embedded-bipartite-graph-network-for-predicting-lncRNA-miRNA-associations
Hasan et al., 2022 [207]	5mC-Methyl Cytosine Modification Prediction	Word2Vec	CNN	https://github.com/hasan022/Deepm5C
Wang et al., 2022 [211]	5mC-Methyl Cytosine Modification Prediction	GloVe	CNN	https://github.com/whl-cumt/EMDLP
Shi et al., 2019 [248]	RNA-Disease Associations Prediction	SDNE	RF	https://github.com/BioMedicalBigDataMining-Lab/NEMII
Shi et al., 2022 [248]	RNA-Disease Associations Prediction	Word2Vec	BiLSTM	https://github.com/hongshi940/HGNNLDA
Li et al., 2021 [259]	RNA-Disease Associations Prediction	SVD, Node2Vec	XGBoost	https://github.com/iALKing/SVDNVLDA
Madhavan et al., 2021 [262]	RNA-Disease Associations Prediction	Node2Vec	DBN	https://github.com/manumad/DBNLDA
Xie et al., 2021 [126]	RNA-Gene Association Prediction	Word2Vec	BiLSTM	https://github.com/Xshelton/SG_LSTM
Przybyszewski et al., 2023 [240]	Micro RNA Target Prediction	Word2Vec	GNN	https://github.com/SanoScience/graphtar
Wolo et al., 2019 [236]	16S rRNA Taxonomic Classification	Word2Vec	NB	https://github.com/EESI/microbiome_embeddings

2'-O-Methylation Modification Prediction, Methylation Modification Prediction, siRNA Target Prediction, and ac4C-Acetyl Cytidine Modification Prediction. An extensive analysis of Table 5 indicates that 17 Transformer based predictive pipelines are developed for 13 RNA sequence analysis task including RNA-Protein Binding Affinity Prediction, Cell-Specific Gene Regulatory Networks Prediction, Methylation Modification Prediction, RNA-Protein Binding Sites Prediction, Long non coding RNA Prediction, CRISPR/Cas9 single guide RNA Prediction, Spatial Gene Expression Analysis, RNA structure prediction, RNA-Seq Coverage Prediction, RNA Subcellular Localization Prediction, Pre-miRNA Prediction, mRNA Degradation Prediction, and RNA-Disease Association Prediction. Whereas 23 BERT based predictive pipelines are developed for 18 RNA sequence analysis tasks. Moreover, only 1 ESM-1b based predictive pipeline is developed for RNA-Protein Binding Sites Prediction, 1 GPT based predictive pipeline is developed for cell-type detection, and 2 heterogeneous graph transformer (HGT) based predictive pipelines are developed for two tasks namely RNA-Disease Association Prediction and microRNA- Target Prediction.

Language models based predictive pipelines can be used in different way: First is to train a language model from scratch on a large dataset, which is also known as self-training and second is to fine-tuning a pre-trained open-source language model for specific downstream tasks. A detailed analysis of existing studies shows that source codes for 24 BERT, 19 Transformer, 1 GPT, and 1 ELMo and ESM-1b based predictive pipelines are publicly available. Among the 24 BERT-based pipelines, 9 are self-trained for 9 different tasks including Single-Cell Multi-Omics Analysis [284], Enhancer RNA Identification task [123], 6mA-Methyl Adenosine Modification [194], Promoter Identification [124], RNA Cluster Analysis [8], RNA Structure Prediction [223], Splicing Sites Prediction [314], RNA Structure and Function Prediction [218], and miRNA Target Prediction [239]. Remaining 15 pre-trained BERT models are used for 11 different tasks namely RNA-Disease Association Prediction [131,132], 6mA-Methyl Adenosine Modification Prediction [281,312], Promoter Identification [125], RNA Structure [223], RNA-Protein Interaction Prediction [9], RNA-Protein Binding Sites Prediction [174,279], 2'-O-Methylation Modification Prediction [191], Methylation Modification Prediction [193,192], Long non coding RNA Prediction [114], siRNA Target Prediction [241], and ac4C-Acetyl Cytidine Modification Prediction [188]. Table 6 presents details of the protein data used to train BERT and 4 other language models, resulting in various pretrained versions.

In a nutshell, this section provides information about 65 open-source predictive pipelines developed by using 14 unique word embedding and 5 distinct large language models. This knowledge can facilitate development of a comprehensive, large-scale RNA sequence analysis framework to harness the capabilities of AI.

10. RNA sequence analysis predictive pipelines performance analysis

To assist computer scientists, this section sheds lights on the performance figures achieved by word embedding, language model, and domain specific representation learning methods based predictive pipelines across 47 distinct RNA sequence analysis tasks using diverse benchmark datasets. To aid researchers in developing new predictors, we have conducted a thorough literature review for each task and discussed current state-of-the-art predictors. In Section 3, we have categorized 47 RNA sequence analysis tasks into 10 distinct categories. Here, we have summarized the performance values of predictive pipelines developed for these tasks into 7 different Tables. Each Table corresponds to the predictive pipelines of tasks within a single category, except two Tables that include the summary of predictive pipelines developed for the tasks coming from 3 different categories and 2 different categories respectively. Moreover, this analysis highlights which tasks within each category offers more room for improvement through the development of more robust and effective predictive pipelines.

Table 7 summarizes crucial details of 9 RNA sequence analysis tasks classified under the goal of RNA categorization and identification. Overall, for RNA categorization and identification goal, 10 unique representation learning methods including BERT, Transformer, Word2vec, HIN2Vec, one-hot encoding, k-mer composition, GloVe, pseudo nucleotides composition, Transformer+BigBird+Longformer, nucleotide physico-chemical properties and occurrence frequency based representation learning approaches are used across 9 different tasks. In 21 predictive pipelines, along with different representation learning approaches, 13 unique classifiers namely BiLSTM, DenseNet, CNN, CNN+BiLSTM, MLP, SVM+LogR, CNN+BiLSTM+MLP, RF, CatBoost, XGBootst, BERT-self classifier, transformer-self classifier and CNN+RNN classifiers are used. Most commonly used representation learning approach is BERT followed by Transformers. A total of 6 studies have developed BERT based predictive pipelines with a self classifier for 5 different tasks namely RNA cluster analysis [8], mRNA identification [107], long non-coding RNA identification [114], enhancer RNA identification, [320] and promoter identification [125,124]. BERT with a self classifier based predictive pipelines has achieved state-of-the-art performance for 4 tasks namely RNA cluster analysis [8], mRNA identification [107], enhancer RNA identification, [320] and promoter identification [125,124]. Second most commonly used representation learning approaches are Transformer and Word2vec. Transformer is used with a self classifier for CRISPER/Cas9 single guide RNA identification [276], and with two classifiers namely XGBoost and CNN for pre-micro RNA identification [15,121]. Transformer is also combined with BigBird and Longformer representation learning approaches to feed statistical vectors to an ensemble (CNN+BiLSTM+MLP) classifier for long non-coding RNA identification [115]. Transformer with XGBoost classifier has yielded state-of-the-art performance [120] for pre-micro RNA identification. Word2vec approach is used with a hybrid (CNN+BiLSTM) classifier for circular RNA identification [111] and is used with CNN and BiLSTM classifiers for small non-coding RNA classification [108,109]. It is important to mention that for small non-coding RNA classification task, there exist three different benchmark datasets which differ from each other in terms of number of classes. Deng et al. [109] non-coding RNA classification dataset is comprised of 4 classes namely lncRNAs, misc-RNAs, rRNAs, and sRNA, Aoki et al. [108] dataset is comprised of 9 classes including snRNA, snoRNA C/D, snoRNA H/ACA, scaRNA miRNA, YRNA, Vault RNA, 5S rRNA, and tRNA, whereas Asim et al. [110] dataset is comprised of 13 classes namely miRNA, ribozymes, 5S rRNA, 5_8S_rRNA, HACA-box, CD-box, tRNA, scaRNA, IRES, $Intron\_gpI$, $Intron\_gpII$, riboswitch, and leader. Considering rich regulatory roles of non-coding RNAs, Asim et al. [110] dataset is more valuable as it allows to identify more types of non-coding RNAs.

Table 5
Summary of open-source language models based predictors in existing studies.

Author, Year [ref]	Task Name	Language Model	Classifier	Pre-train/ Self-train	Code link
Shen et al., 2024 [182]	RNA-Protein Binding Affinity Prediction	Transformer	_	Self-train	https://github.com/xilinshen/Reformer
Zhao et al., 2024 [311]	RNA-Protein Binding Affinity Prediction	Transformer	_	Self-train	https://github.com/pfnet-research/GenerRNA
Xu et al., 2023 [235]	Cell-Specific Gene Regulatory Networks Prediction	Transformer	_	Self-train	https://github.com/zhanglab-wbgcas/STGRNS
Yang et al., 2022 [284]	Single-Cell Multi-Omics Analysis	BERT	_	Self-train	https://github.com/TencentAILabHealthcare/scBERT
He at al., 2023 [26]	mRNA Degradation Prediction	Transformer	_	Self-train	https://github.com/Shujun-He/RNAdegformer
Zou et al., 2024 [128]	RNA-Disease Association Prediction	Heterogeneous Graph Transformer	_	Self-train	https://github.com/zht-code/HGTMDA
Li et al., 2024 [136]	RNA-Disease Association Prediction	Transformer	_	Self-train	https://github.com/ghli16/NAGTLDA
Yao et al., 2024 [135]	RNA-Disease Association Prediction	Transformer	_	Self-train	https://github.com/ydkvictory/GCNFORMER
Wu et al., 2023 [133]	RNA-Disease Association Prediction	Transformer	_	Self-train	https://github.com/jinyangwu/KGETCDA
Ning et al., 2023 [131]	RNA-Disease Association Prediction	BERT	_	Pre-train	https://github.com/zhiweining/BertNDA-main
Zhao et al., 2022 [134]	RNA-Disease Association Prediction	Transformer	_	Self-train	https://github.com/EchoChou-990919/LDAformer
Yang et al., 2022 [132]	RNA-Disease Association Prediction	BERT	_	Pre-train	https://github.com/Wolverinerine/GTGenie
Zhang et al., 2023 [123]	Enhancer RNA Identification	BERT	_	Self-train	https://github.com/lyli1013/DeepITEH
Zhang et al., 2024 [194]	6mA-Methyl Adenosine Modification Prediction	BERT	_	Self-train	https://github.com/TingheZhang/m6A-BERT
Li et al., 2023 [281]	6mA-Methyl Adenosine Modification Prediction	BERT	_	Pre-train	https://github.com/liqianyue/zeitgeist-/tree/master/m6A_BERT_Stacking
Le et al., 2022 [312]	6mA-Methyl Adenosine Modification Prediction	BERT	CNN	Pre-train	https://github.com/khanhlee/bert-dna
Zeng et al., 2023 [19]	RNA Subcellular Localization Prediction	Transformer	_	Self-train	https://github.com/CSUBio-Group/LncLocFormer
Raad et al., 2022 [121]	Pre-micro RNA Prediction	Transformer	CNN	Self-train	https://github.com/sinc-lab/miRe2e
Wang et al., 2023 [125]	Promoter Identification	BERT	_	Pre-train	https://github.com/xwang1427/miPTP/tree-/main/SCPseDNC/data

(continued on next page)

Table 5 (*continued*)

Author, Year [ref]	Task Name	Language Model	Classifier	Pre-train/ Self-train	Code link
Mai et al., 2022 [124]	Promoter Identification	BERT	_	Self-train	https://github.com/hanepira/TSSnote-CyaPromBert
Akiyama et al., 2022 [8]	RNA Cluster Analysis	BERT	_	Self-train	https://github.com/mana438/RNABERT.git
Linder et al., 2023 [27]	RNA-Seq Coverage Prediction	Transformer	_	Self-train	https://github.com/calico/borzoi
Cui et al., 2024 [242]	Single-Cell Multi-Omics Analysis	GPT	_	Self-train	https://github.com/bowang-lab/scGPT
Zhang et al., 2024 [222]	RNA Structure Prediction	BERT	_	Self-train	https://doi.org/10.5281/zenodo.8280831
Fei et al., 2022 [225]	RNA Structure Prediction	Transformer	_	Pre-train	https://github.com/jluF/LTPConstraint
Kalicki et al., [223]	RNA Structure Prediction	BERT	_	Pre-train	https://github.com/dhesin/RNABERT-2
Wang et al., 2024 [231]	Spatial Gene Expression Analysis	Transformer	_	Pre-train	https://zenodo.org/records/10646474
Wan et al., 2022 [276]	CRISPR/Cas9 single guide RNA Prediction	Transformer	_	Self-train	https://github.com/BioinfoApollo/TransCrispr
Liu et al., 2023 [313]	Micro RNA Target Prediction	Heterogeneous Graph Transformer	_	Self-train	https://github.com/Liangyushi/MiR-Graph/tree/main
Yamada et al., 2022 [9]	Coding RNA-Protein Interaction Prediction	BERT	_	Pre-train	https://github.com/kkyamada/bert-rbp
Chen et al., 2023 [314]	RNA Splicing Sites Prediction	BERT	_	Self-train	https://github.com/biomed-AI/SpliceBERT
Chen et al., 2022 [218]	RNA Structure Prediction RNA Function Prediction	BERT	CNN	Self-train	https://github.com/ml4bio/RNA-FM
Dai et al., 2023 [115]	Long non coding RNA Prediction	Transformer	Hybrid (CNN + BiLSTM + MLP)	Self-train	https://github.com/yatoka233/LncPNdeep
Zhang et al., 2024 [239]	Micro RNA Target Prediction	BERT	_	Self-train	https://github.com/mingziiz/miTDS
Cao et al., 2024 [175]	RNA-Protein Binding Sites Prediction	Transformer	_	Self-train	https://github.com/cc646201081/CircSI-SSL
Yan et al., 2024 [172]	RNA-Protein Binding Sites Prediction	ELMo, ESM-1b	XGBoost	Pre-train	https://github.com/yaoyao-11/Seq-RBPPred
Jin et al., 2023 [174]	RNA-Protein Binding Sites Prediction	BERT	BERT	Pre-train	https://github.com/YeoLab/HydRA
Du et al., 2022 [279]	RNA-Protein Binding Sites Prediction	BERT	BiLSTM	Pre-train	https://github.com/Xuezg/JLCRB
Soylu et al., 2023 [191]	2’-O-Methylation Modification Prediction	BERT	CNN	Pre-train	https://github.com/seferlab/bert2ome
Wang et al., 2024 [193]	Methylation Modification Prediction	BERT	CNN	Pre-train	https://github.com/abhhba999/MRM-BERT

Table 5 (*continued*)

Author, Year [ref]	Task Name	Language Model	Classifier	Pre-train/ Self-train	Code link
Chen et al., 2023 [209]	Methylation Modification Prediction	Transformer	Transformer	Pre-train	https://github.com/lennylv/TransRNAm
Wang et al., 2024 [192]	Methylation Modification Prediction	BERT	BERT	Pre-train	https://github.com/Moretta1/BERT-RNA
Danilevicz et al., 2023 [114]	Long non coding RNA Prediction	BERT	BERT	Pre-train	https://github.com/AppliedBioinformatics/lncRNAPrediction_Interpretation
Xu et al., 2024 [241]	siRNA Target Prediction	BERT	BERT	Pre-train	https://github.com/ChengkuiZhao/siRNABERT
Li et al., 2024 [188]	ac4C-Acetyl Cytidine Modification Prediction	BERT	BiLSTM	Pre-train	https://github.com/Marscolono/MetaAc4C

Beyond most common BERT, Transformer, and Word2vec approaches, several other representation approaches are used with different classifiers for various RNA sequence analysis tasks. Specifically, HIN2Vec with MLP classifier is used for circular RNA identification [112] and GloVe with hybrid (CNN+BiLSTM) classifier is used for long non-coding RNA identification [116]. Apart from word embedding and language models based predictive pipelines, k-mer composition along with hybrid (CNN+BiLSTM) classifier and physico-chemical properties and occurrence frequency based encoder with ensemble (SVM+LogR) classifier is used for circular RNA [13] identification and long non-coding RNA identification [286], respectively. Overall, among all representation learning approaches used for circular RNA identification, k-mer composition based representation learning approach with a hybrid (CNN+BiLSTM) classifier achieves state-of-the-art performance [13]. Similarly, among all representation learning approaches used for long non-coding RNA identification [286], physico-chemical properties and occurrence frequency based representation learning approach along with ensemble (SVM+LogR) classifier state-of-the-art performance. Among all 9 tasks, enhancer RNA and promoter identification have some room for improvement. Considering the performance trend of all predictive pipelines in this goal, potential of physico-chemical properties and occurrence frequency based representation learning approach with an ensemble classifier (CNN+BiLSTM or SVM+LogR) can enhance the performance figures for under-performing tasks.

Table 8 summarizes 54 existing studies related to 4 different RNA sequence analysis tasks classified under the biological goal of RNA target prediction. For this goal, 18 unique representation learning approaches are used that include Word2vec, DeepWalk, heterogeneous graph transformer, transformer, RWR, weisfeiler-leman algorithm, RotatE, Node2vec, Node2vec+GATNE, SDNE, BERT, sparse quality control, SVD, k-mer composition, stacked auto-encoder, Graph2vec, SVD+Node2vec, and HOPE. Using different representation learning approaches, predictive pipelines are developed by employing 27 classifiers including BiLSTM, GNN, CNN, LSTM, RF, MLP, GNN+MLP, GAT+MLP, GCN+MLP, hyper-graph convolutional network, rotation forest model, DF, transformer-self classifier, BERT-self classifier, heterogeneous graph transformer-self classifier, hybrid (CNN+GuassianNB), LogR, matrix multiplication+MLP, XGBoost, GBDT+LogR, XGBoost, ensemble (XGBoost + LightGBM + RF + ET + CatBoost), neural network regression model, GCN, ET, ensemble (AdaBoost-CNN+LightGBM), ensemble (SVM + GBDT + AdaBoost + XGBoost + RF + MLP), ensemble (CatBoost + ET + LightGBM + RF + XGBoost + LR), and adaptive subspace leaning method. Most commonly used representation learning approach is Node2Vec followed by Word2Vec and BERT. Node2vec representation learning approach is employed with 5 different classifiers for non-coding RNA disease association prediction tasks [145,146,143,248,262,261]. Specifically, Node2vec is used with three different classifiers namely RF [146], GCN [143] and hybrid (CNN+GuassianNB) [145] classifiers for 1 task namely miRNA-disease association prediction whereas Node2vec is employed with LogR [261] and neural network regression models [262] for lncRNA-disease association prediction. Moreover, combined potential of Node2vec and SVD is explored with XGBoost classifier for lncRNA-disease association [259] and Node2vec+GATNE representation learning is used with RF classifier for miRNA-disease association prediction [256]. Despite being the most common representation learning approach for this goal, Node2vec based any predictive pipeline does not achieve state-of-the-art performance on any task of this goal.

Word2vec is the second most commonly used representation learning approach which is employed with LSTM classifier for RNA-gene association prediction [127] and with GNN classifier for micro RNA target prediction [240]. Furthermore, potential of Word2vec is explored with BiLSTM classifier for 3 tasks namely RNA-gene association prediction [126], micro RNA target prediction [253], and RNA disease association prediction [248]. Word2vec is also used with ensemble (matrix factorization + MLP) classifier for lncRNA-disease association prediction [249]. Among all tasks, Word2Vec representation with LSTM classifier has achieved state-of-the-art performance for RNA-gene association prediction [127]. Apart from Node2vec and Word2vec, potential of BERT representation learning with a self classifier is explored for 3 tasks namely siRNA target prediction [241], miRNA target prediction [16] and lncRNA-disease association prediction [131,132,139]. BERT with a self-classifier manages to achieve state-of-the-art performance across 2 tasks namely siRNA target prediction [241], and miRNA target prediction [16]. Beyond Node2vec and Word2vec, transformer is used with hypergraph convolutional network for lncRNA-disease association prediction [129] and its potential is also explored with a self classifier for 2 tasks namely cirRNA-disease [133] and lncRNA-disease association prediction [137]. In addition, heterogeneous

Table 6
Summary of Uniquely Pre-trained Language Models along with pre-training Data for RNA Sequence Analysis Tasks.

Unique Language Model	Pre-trained Data	Unique Language Model	Pre-trained Data	Unique Language Model	Pre-trained Data
Shen et al., Transformer [182]	eCLIP-seq Data	Dai et al., Transformer [115]	48,876 LncRNAs, 99,187 Coding RNAs	Zhang et al., BERT [239]	miRAW Dataset
Zhao et al., Transformer [311]	34.39M Sequences from RNAcentral	Fei et al., Transformer [225]	Rfam Data (43,273 pieces of Data)	Kalicki et al., BERT [223]	410K sequences from 2 mRNA families virus and Humans for a total of 31 RNA families
Xu et al., Transformer [235]	scRNA-Seq Data	Zou et al., Heterogeneous Graph Transformer [128]	Trained on 35,547 Data from MDA Database	Devlin et al., BERT [266]	BooksCorpus (800M words), English Wikipedia (2,500M words)
He at al., Transformer [26]	OpenVaccine challenge Dataset	Liu et al., Heterogeneous Graph Transformer [313]	miRAW train-validation dataset	Ji et al., BERT [315]	human genome 78 mouse ENCODE ChIP-seq datasets
Li et al., Transformer [136]	2797 lncRNA-disease relationships	Yang et al., BERT [284]	scRNA-Seq Data	Zhang et al., BERT [316]	Cora, Citeseer and Pubmed Datasets
Yao et al., Transformer [135]	LncRNADisease, Lnc2Cancer Datasets	Zhang et al., BERT [123]	eRNA Data from eRNA Database (HeRA)	Brandes et al., BERT [317]	~106 million UniRef90 protein sequences
Wu et al., Transformer [133]	nCRNA Dataset	Zhang et al., BERT [194]	427,760 Human m6A Sites	Lee et al., BERT [318]	single cell RNA sequence data and gene contextual information
Zhao et al., Transformer [134]	LncRNA Data	Mai et al., BERT [124]	dRNA-Seq Dataset	Sarzynska-Wawer et al., ELMo [319]	20-million-words data set sampled from Wikipedia and Common Crawl
Zeng et al., Transformer [19]	lncRNA subcellular localization Dataset	Akiyama et al., BERT [8]	76237 Human derived small ncRNAs with lengths ranging from 20 to 440 bases from RNAcentral	Rives et al., ESM 1 [271]	250 million protein sequences
Raad et al., Transformer [121]	Metazoan pre-miRNAs (23178)	Zhang et al., BERT [222]	TR0 Dataset	Cui et al., GPT [242]	Over 10.3M scRNA-Seq samples of Human blood and bone marrow
Linder et al., Transformer [27]	CAGE Dataset (Human and Mouse RNA-Seq)	Chen et al., BERT [314]	Over 2M precursor messenger RNA (pre-mRNA) Sequences from 72 vertebrates	_	_
Wan et al., Transformer [276]	Sniper-Cas9, SpCas9-NG, xCas9, HypaCas9	Chen et al., BERT [218]	23M cRNA Sequences from RNAcentral Database	_	_

graph transformer is used with a self classifier for 2 tasks including miRNA-disease association prediction [130] and circRNA-disease association prediction [285]. Furthermore, HOPE representation learning is used with a rotation forest classifier [258], Graph2vec is used with an ensemble (GBDT+LR) classifier [144], and SVD is employed with ensemble (AdaBoost-CNN+LightGBM) classifier [157] for lncRNA-disease association prediction. Moreover, two studies have explored the potential of DeepWalk representation leaning with MLP classifier for miRNA-disease association prediction [140,257]. In addition, sparse quality control based representation learning is used with MLP classifier for circRNA-disease association prediction [155]. Overall among all different predictive pipelines, DeepWalk with MLP classifier based predictive pipelines achieves state-of-the-art performance for miRNA-disease association prediction [140]. Similarly, Transformer with a self classifier based predictive pipeline shows state-of-the-art performance across 5 different benchmark datasets related to lncRNA-disease association prediction [135,136]. From all 4 tasks of this goal, siRNA and miRNA target prediction offer some room for improvement. Considering performance trend of different predictive pipelines developed for this goal, potential

Table 7
RNA categorization and identification related 9 distinct RNA sequence analysis tasks predictive pipelines performance.

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
Clustering	RNA Cluster Analysis	**Akiyama et al., 2022 [8]**	**1. Akiyama et al. Train set-A, 2. Akiyama et al. Train set-B**	**BERT**	_	**(TrainSet-A) Sn = 0.881, Positive Predictive Value = 0.947, F1-score = 0.913; (TrainSet-B) Sn = 0.851, Positive Predictive Value = 0.932, F1-score = 0.890**
Multi-label Classification	Small Non-coding RNA Classification	**Deng et al., 2023 [109]**	**Deng et al. Dataset 1, Deng et al. Dataset 2**	**Word2Vec**	**BiLSTM**	**Dataset 1: Acc = 79.43, Precision = 79.25, Sn = 81.36, Sp = 77.38, F1-score = 0.803, MCC = 0.588, AUROC = 0.885; Dataset 2: Acc = 98.61, Precision = 99.22, Sn = 98.84, Sp = 98.01, F1-score = 0.990, MCC = 0.966, AUROC = 0.997**
		Asim et al., 2020 [110]	Asim et al. Non-Coding RNA Classification Dataset	One-hot Encoding	DenseNet	Acc = 0.9538, Precision = 0.9539, Recall = 0.9538, F1-score = 0.9536
		Aoki et al., 2018 [108]	**Aoki et al. Dataset**	**Word2Vec**	**CNN**	**Acc = 0.980, F1-score = 0.931**
Binary Classification	mRNA Identification	**Li et al., 2023 [107]**	**MLOS Flu Vaccines Dataset, Nieuwkoop et al. Dataset, Wint et al. Dataset, lixiProtein Expression Dataset, Groher et al. Dataset, Diez et al. Dataset, SARS-CoV-2 Vaccine Degradation Dataset**	**BERT**	_	**MLOS Flu Vaccines: RMSE = 0.78, Nieuwkoop et al. Dataset: RMSE = 0.88, Wint et al. Dataset: RMSE = 0.89, lixiProtein Expression Dataset: RMSE = 0.57, Groher et al. Dataset: RMSE = 0.35, Diez et al. Dataset: RMSE = 0.48, SARS-CoV-2 Vaccine Degradation: RMSE = 0.78**
Binary Classification	Circular RNA Identification	**Niu et al., 2024 [13]**	**Niu et al. Dataset**	**k-mer Composition**	**CNN + BiLSTM**	**Acc = 0.8614, SN = 0.8381, Sp = 0.8165, MCC = 0.6774**
		Chaabane et al., 2020 [111]	**Chaabane et al. Dataset**	**Word2Vec**	**CNN + BiLSTM**	**Acc = 0.8056, MCC = 0.6113, F1-score = 0.810**
		Deng et al., 2020 [112]	**Deng et al. Dataset**	**HIN2Vec**	**MLP**	**F1-score = 0.412, Recall = 0.400, Acc = 0.425**

(*continued on next page*)

of shallow neural network based embeddings such as Word2vec, random walk based node embedding methods such as Node2vec, DeepWalk, and graph based transformers like heterogeneous graph transformer along with standalone classifier (MLP, GCN) or an ensemble (CNN+GuassianNB) classifier can be explored for enhancing the performance of under-performing tasks.

Table 9 provides a summary of 29 RNA sequence analysis studies related to 4 different tasks classified under the hood of RNA interaction prediction. Overall, 12 unique representation learning approaches namely nucleotides composition encoder, Word2vec, Node2vec, HIN2vec, VGAE+Word2vec, ELMo+ESM-1b, BERT, one hot encoding, nucleotide frequency and density encoder, transformer, Word2vec in conjunction with nucleotide composition encoder, and Struct2vec are used across 4 different tasks. These representation learning approaches are used with 18 different classifiers including GCN, MLP, GNN, SVM, RF, XGBoost, CNN, BERT-self classifier, Transformer-self classifier, hybrid (CNN+BiLSTM), hybrid (CNN+BiGRU), LogR, BiLSTM, BiLSTM+LSTM, AdaBoost, DNN, GBDT, and CatBoost to develop predictive pipelines across 4 distinct tasks.

For this goal, most commonly used representation learning approach is Word2vec followed by BERT. Word2vec is utilized with 8 different classifiers for 3 different tasks namely coding RNA-protein interaction prediction [167,170], protein-RNA binding sites prediction [178,250,251,179,7], and non-coding RNA interaction prediction [187,184]. Specifically, Word2vec is employed with MLP for 2 different tasks namely coding RNA-protein interaction prediction [167], and protein-RNA binding sites prediction [178] and it is employed with RF classifier for coding RNA-protein interaction prediction [170]. In addition, potential of Word2vec is explored with 3 different classifiers namely CNN [7], hybrid (LSTM+BiLSTM) [251], and hybrid (CNN+BiLSTM) [250,179] for protein-RNA binding sites prediction [250,251,179,7], whereas Word2vec is employed with GCN [184], and AdaBoost [187] classifiers for non-coding

Table 7 (*continued*)

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
Binary Classification	Long Non-coding RNA Identification	**Tian et al., 2024 [117]**	**Tian et al. Datasets (Amborella trichopoda, Ananas comosus, Arabidopsis thaliana, Brachypodium distachyon, Cucumis sativus, Glycine max, Manihot esculenta, Medicago truncatula, Musa acuminata, Oryza sativa, Populus trichocarpa, Solanum lycopersicum, Sorghum bicolor, Vitis vinifera, Zea mays, Chlamydomonas reinhardtii, Coccomyxa subellipsoidea, Micromonas pusilla, Volvox carteri, Physcomitrella patens)**	**ORFS + ORFC + Fickett Test code + Hexamer usage bias + Sequence Intrinsic Composition + Structural Information + EIIP based Physiochemical Properties**	**SVM + LogR**	**Amborella trichopoda: Precision = 94.20, Ananas comosus: Precision = 97.30, Arabidopsis thaliana: Precision = 0.96, Brachypodium distachyon: Precision = 0.94, Cucumis sativus: Precision = 0.94, Glycine max: Precision = 0.91, Manihot esculenta: Precision = 0.96, Medicago truncatula: Precision = 0.92, Musa acuminata: Precision = 0.96, Oryza sativa: Precision = 0.95, Populus trichocarpa: Precision = 0.91, Solanum lycopersicum: Precision = 0.96, Sorghum bicolor: Precision = 0.97, Vitis vinifera: Precision = 0.92, Zea mays: Precision = 0.94, Chlamydomonas reinhardtii: Precision = 0.94, Coccomyxa subellipsoidea: Precision = 0.95, Micromonas pusilla: Precision = 1.00, Volvox carteri: Precision = 0.98, Physcomitrella patens: Precision = 0.93**
		Dai et al., 2023 [115]	Dai et al. Dataset	Transformer + BigBird + Longformer	CNN + BiLSTM + MLP	Acc = 0.971, Sp = 0.967, Sn = 0.980
		Danilevicz el al., 2023 [114]	**Danilevicz et al. Datasets: 1. Arabidopsis thaliana Dataset, 2. Brassica napus Dataset, 3. Brassica oleracea Dataset, 4. Brassica rapa Dataset, 5. Glycine max Dataset, 6. Oryza sativa Dataset, 7. Zea mays Dataset**	**BERT**	–	**Arabidopsis thaliana Dataset: Acc = 65.39, AUROC = 0.72, F1-score = 0.65, MCC = 0.31, Precision = 0.65, Recall = 0.65; Glycine max Dataset: Acc = 72.77, AUROC = 0.79, F1-score = 0.73, MCC = 0.45, Precision = 0.73, Recall = 0.73; Brassica napus Dataset: Acc = 74.6, AUROC = 0.81, F1-score = 0.74, MCC = 0.49, Precision = 0.75, Recall = 0.74; Brassica oleracea Dataset: Acc = 74.15, AUROC = 0.81, F1-score = 0.74, MCC = 0.49, Precision = 0.75, Recall = 0.74; Brassica rapa Dataset: Acc = 57.86, AUROC = 0.61, F1-score = 0.58, MCC = 0.16, Precision = 0.58, Recall = 0.58; Oryza sativa Dataset: Acc = 61.65, AUROC = 0.65, F1-score = 0.62, MCC = 0.23, Precision = 0.62, Recall = 0.62; Zea mays Dataset: Acc = 83.42, AUROC = 0.90, F1-score = 0.83, MCC = 0.67, Precision = 0.84, Recall = 0.84**
		Nadir et al., 2021 [119]	Nadir et al. Dataset	k-mer Composition	RF	Acc = 0.9984, Precision = 0.9999, Recall = 0.9968, F1-score = 0.9983

Table 7 (*continued*)

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
		Musleh et al., 2021 [118]	**Musleh et al. Datasets (Human, Mouse)**	**k-mer Composition + Pseudo Nucleotide Composition**	**CatBoost**	**Human: Acc = 96.04, Mouse: Acc = 96.05**
		Liu et al., 2019 [116]	Liu et al. Dataset	GloVe	CNN + BiLSTM	F1-score = 97.9, Acc = 96.4, AUROC = 99.0
Binary Classification	Pre-micro RNA Identification	**Gupta et al., 2023 [120]**	**Gupta et al. Dataset**	**Transformer**	**XGBoost**	**Acc = 98**
		Raad et al., 2022 [121]	**Raad et al. Dataset**	**Transformer**	**CNN**	**AUPRC = 0.12313**
Binary Classification	CRISPR/Cas9 single guide RNA Identification	**Zhu et al., 2024 [122]**	**Hart et al. Datasets: (WT, ESP, HF, xCas, SpCas9, Snipe, HCT116, HELA, HL60)**	**One-hot Encoding**	**CNN + RNN**	**WT: SRCC = 0.867, PRCC = 0.891; ESP: SRCC = 0.852, PRCC = 0.846; HF: SRCC = 0.859, PRCC = 0.875; xCas: SRCC = 0.866, PRCC = 0.855; SpCas9: SRCC = 0.852, PRCC = 0.873; Snipe: SRCC = 0.939, PRCC = 0.959; HCT116: SRCC = 0.335, PRCC = 0.346; HELA: SRCC = 0.354, PRCC = 0.344; HL60: SRCC = 0.389, PRCC = 0.386**
		Wan et al., 2022 [276]	Wang et al. Datasets: 1. eSpCas9, 2. SpCas9-HF1, 3. WT-SpCas9; Kim et al. Datasets: 4. Sniper-Cas9, 5. SpCas9-NG, 6. xCas9, 7. HypaCas9	Transformer	_	WT-SpCas9 Dataset: SRCC = 0.861, PCC = 0.889; SpCas9-HF1 Dataset: SRCC = 0.852, PCC = 0.864; eSpCas9 Dataset: SRCC = 0.863, PCC = 0.856; Four Datasets: (Sniper-Cas9, SpCas9-NG, xCas9, HypaCas9): Average SRCC = 0.818, PCC = 0.783
Binary Classification	Enhancer RNA Identification	**Zhang et al., 2023 [123]**	**Zhang et al. Dataset (Stomach, Lung, Liver, Pancreas, LIHC, LUAD, PRAD, PAAD)**	**BERT**	_	**Normal tissues: Stomach Dataset: Acc = 86.25, Lung Dataset: Acc = 78.59, Liver Dataset: Acc = 70.74, Pancreas Dataset: Acc = 65.43; Cancer tissues: LIHC Dataset: Acc = 70.45, LUAD Dataset: Acc = 86.25, PRAD Dataset: Acc = 86.25, PAAD Dataset: Acc = 86.25**
Binary Classification	Promoter Identification	**Wang et al., 2023 [125]**	**Wang et al. Dataset 3**	**BERT**	_	**Precision = 78.13, Recall = 75.76**
		Mai et al., 2022 [124]	**Mai et al. Datasets 1. Synechococcus elongatus sp. UTEX 2773 (promoter, non-promoter), 2. Synechocystis sp. PCC 6803 (promoter, non-promoter), 3. Synechocystis sp. PCC 6714 (promoter, non-promoter)**	**BERT**	_	**Synechococcus elongatus sp. UTEX 2773: promoter: AUROC = 0.98, Precision = 0.92, F1-score = 0.93, Support = 1001, non-promoter: AUROC = 0.98, Precision = 0.95, F1-score = 0.93, Support = 1036; Synechocystis sp. PCC 6803: promoter: AUROC = 0.96, Precision = 0.88, F1-score = 0.91, Support = 1407, non-promoter: AUROC = 0.96, Precision = 0.94, F1-score = 0.91, Support = 1433; Synechocystis sp. PCC 6714: promoter: AUROC = 0.96, Precision = 0.91, F1-score = 0.89, Support = 330, non-promoter: AUROC = 0.96, Precision = 0.88, F1-score = 0.89, Support = 330**

Table 8

Non-coding RNA target prediction related 4 distinct RNA sequence analysis tasks predictive pipelines performance.

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
Interaction	RNA-Gene Association Prediction	Yoon et al., 2023 [127]	Yoon et al. Dataset	Word2Vec	LSTM	AUROC = 0.9834
		Xie et al., 2021 [126]	**Xie et al. Dataset**	**Word2Vec**	**BiLSTM**	**AUROC = 0.94**
Interaction	RNA-Disease Association Prediction	**Lu et al., 2024 [140]**	**Lu et al. Dataset**	**DeepWalk**	**MLP**	**AUROC = 0.9478, AUPRC = 0.9464, Acc = 0.8908, Precision = 0.9237, Recall = 0.9096, F1-score = 0.8785**
		Zou et al., 2024 [128]	**Zou et al. Dataset**	**Heterogeneous Graph Transformer**	**MLP**	**Acc = 0.8927, Sn = 0.8838, Sp = 0.8881, Precision = 0.8926, MCC = 0.772, AUROC = 0.9551**
		Ouyang et al., 2024 [129]	**MDAv2.0 Dataset, MDAv3.2 Dataset**	**Transformer**	**Hypergraph Convolutional Network**	**MDAv2.0: AUROC = 0.945284, AUPRC = 0.945074, F1-score = 0.879973; MDAv3.2: AUROC = 0.962600, AUPRC = 0.959563, F1-score = 0.902512**
		Tian et al., 2024 [147]	**Tian et al. Dataset**	**RWR**	**GCN**	**AUROC = 0.9874 ± 0.0078, Acc = 0.9453 ± 0.0089, AUPRC = 0.9882 ± 0.0013**
		Ruan et al., 2024 [148]	**Ruan et al. Dataset**	**GCN**	**MLP**	**AUROC = 0.9484 ± 0.0002, AUPRC = 0.3526 ± 0.0038**
		Xu et al., 2024 [149]	**Xu et al. Dataset**	**GNN**	**MLP**	**AUROC = 96.76, AUPRC = 96.37, Acc = 86.95, F1-score = 88.32, Recall = 99.16, Precision = 79.99**
		Ji et al., 2024 [150]	**Ji et al. Dataset**	**Graph Attention Neural Network**	**MLP**	**Acc = 0.9292 ± 0.0287, Sn = 0.9331 ± 0.0244, Sp = 0.9254 ± 0.0343, Precision = 0.9261 ± 0.034, MCC = 0.8585 ± 0.0573, AUROC = 0.9738 ± 0.0135**
		Liang et al., 2024 [151]	**Li et al. Dataset**	**Weisfeiler-Leman Algorithm**	**CNN**	**AUROC = 0.9401 ± 0.0020, AUPRC = 0.2728 ± 0.0077, F1-score = 0.3212 ± 0.0078, Acc = 0.9937 ± 0.0004**
		Jindal et al., 2023 [141]	**Ding et al. Dataset**	**DeepWalk**	**DF**	**AUROC = 0.942**
		Liu et al., 2023 [130]	**Dai et al. Data2 Dataset**	**Heterogeneous Graph Transformer**	_	**Data2: AUROC = 0.9710, AUPRC = 0.9647, Acc = 0.9201, F1-score = 0.9221, Recall = 0.9457, Pecision = 0.8998**
		Wang et al., 2023 [152]	**Huang et al. Dataset**	**GCN**	**CNN**	**AUROC = 0.9032**
		Cao et al., 2023 [153]	**Cao et al. Dataset**	**RotatE**	**GCN**	**AUROC = 0.9892, AUPRC = 0.9898**
		Sun et al., 2022 [145]	Sun et al. Dataset	Node2Vec	CNN + GaussianNB	AUROC = 0.80, AUPRC = 0.87
		Pang et al., 2022 [321]	HMDD Dataset	Transformer	_	Average Precision = 92.735, F1-score = 84.430, Acc = 85.255, AUROC = 93.012
		Wang et al., 2021 [142]	**Wang et al. Dataset**	**DeepWalk**	**MLP**	**AUROC = 0.943, AUPRC = 0.937**
		Yu et al., 2021 [256]	HMDD v3.2 Dataset	Node2Vec + GATNE	RF	Precision = 0.6509, Recall = 0.4991, F1-score = 0.5649

Table 8 (*continued*)

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
		Zheng et al., 2020 [146]	Zheng et al. Datasets (miRNA-Disease Association baseline, Unknown Diseases and miRNAs)	Node2Vec	RF	miRNA-Disease Association baseline Dataset: AUROC = 0.9145, Acc = 84.49; Unknown Diseases and miRNAs Prediction: AUROC = 0.8765, Acc = 80.96
		Li et al., 2019 [143]	**Li et al. Datasets (Disease–Gene Interaction Data)**	**Node2Vec**	**GCN**	**AUROC = 0.9626, Precision = 0.9660**
		Gong et al., 2019 [154]	**Gong et al. Dataset**	**SDNE**	**RF**	**AUPRC = 0.6104 ±0.0012, AUROC = 0.9293 ± 0.0017, F1-score = 0.6147 ± 0.0025, Acc = 0.9956 ± 0.0001, Recall = 0.4893 ± 0.0060, Sp = 0.9993 ± 0.0001, Precision = 0.8289 ± 0.0164**
		Ning et al., 2023 [131]	**Ning et al. Dataset 1, Ning et al. Dataset 2**	**BERT**	_	**Ning et al. Dataset 1: AUROC = 0.998, AUPRC = 0.998; Ning et al. Dataset 2: AUROC = 0.987, AUPRC = 0.988**
		Yang et al., 2022 [132]	**HMDD Dataset, HMDAD Dataset, LncRNADisease v2017 Dataset**	**BERT**	_	**HMDD Dataset: AUROC = 0.9755 ± 0.0022; HMDAD Dataset: AUROC = 0.9654 ± 0.0160; LncRNADisease: AUROC = 0.9810 ± 0.0043**
		Wu et al., 2022 [137]	Wu et al. Disease-lncRNA Dataset, Wu et al. Disease-miRNA Dataset	Transformer	_	Disease-lncRNA Dataset: AUROC = 0.8748; Disease-miRNA Dataset: AUROC = 0.8797
		Li et al., 2024 [155]	**Lan et al. Dataset 1, Lan et al. Dataset 2, Lan et al. Dataset 3, Lan et al. Dataset 4, Lan et al. Dataset 5, Wu et al. Dataset 2, Li et al. Dataset 1, Li et al. Dataset 2**	**Sparse Quality Control (SQC)**	**MLP**	**Lan et al. Dataset 1: AUROC = 0.9569, AUPRC = 0.2451; Lan et al. Dataset 2: AUROC = 0.9057, AUPRC = 0.2027; Lan et al. Dataset 3: AUROC = 0.9495, AUPRC = 0.3217; Lan et al. Dataset 4: AUROC = 0.9409, AUPRC = 0.5360; Lan et al. Dataset 5: AUROC = 0.8644, AUPRC = 0.0062; Wu et al. Dataset 2: AUROC = 0.7543, AUPRC = 0.0130; Li et al. Dataset 1: AUROC = 0.9491, AUPRC = 0.0591; Li et al. Dataset 2: AUROC = 0.9384, AUPRC = 0.2759**
		Wu et al., 2023 [133]	**Wu et al. Dataset 1, Wu et al. Dataset 2, Wu et al. Dataset 3**	**Transformer**	_	**Wu et al. Dataset 1: AUROC = 0.9213, AUPRC = 0.0302; Wu et al. Dataset 2: AUROC = 0.7149, AUPRC = 0.0081; Wu et al. Dataset 3: AUROC = 0.8398, AUPRC = 0.0520**
		Ma et al., 2023 [138]	Ma et al. Dataset 2	Transformer	_	AUROC = 95.44
		Kang et al., 2023 [160]	Kang et al. Dataset 1, Kang et al. Dataset 2	GAT	MLP	Kang et al. Dataset 1: AUROC = 0.9461, Recall = 0.9475; Kange et al. Dataset 2: AUROC = 0.9415, Recall = 0.9423

(*continued on next page*)

Table 8 (*continued*)

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
		Liu et al., 2023 [159]	CircR2Disease Dataset, circRNADisease Dataset, Circ2Disease Dataset, circAtlas Dataset	GCN	MLP	CircR2Disease Dataset: AUROC = 0.9877, AUPRC = 0.9892, F1-score = 0.9878; circRNADisease Dataset: AUROC = 0.9743, AUPRC = 0.9851, F1-score = 0.9736; Circ2Disease Dataset: AUROC = 0.9799, AUPRC = 0.9830, F1-score = 0.9799; circAtlas Dataset: AUROC = 0.9587, AUPRC = 0.9787, F1-score = 0.9568
		Fu et al., 2023 [161]	Fu et al. Dataset 2, Fu et al. Dataset 3	Heterogenous GCN	CNN	Fu et al. Dataset 2: Acc = 0.9477, Precision = 0.9423, Sn = 0.9521, F1-score = 0.9470, MCC = 0.9018; Fu et al. Dataset 3: AUROC = 0.9032, AUPRC = 0.9123, Acc = 0.8582, Sn = 0.8335, F1-score = 0.8523, MCC = 0.7579
		Lu et al., 2022 [285]	Lu et al. Dataset 3	Heterogeneous Graph Transformer	–	AUROC = 0.886, AUPRC = 0.817, Acc = 0.824, Precision = 0.808, Recall = 0.814, F1-score = 0.804
		Xiao et al., 2021 [257]	Xiao et al. Dataset	DeepWalk	Adaptive Subspace Learning Model	AUROC = 0.926 ± 0.015, AUPRC = 0.284 ± 0.013, Precision = 0.381 ± 0.063, Recall = 0.285 ± 0.018, Acc = 0.997 ± 0.001, F1-score = 0.326 ± 0.040
		Yao et al., 2024 [135]	**Fu et al. Dataset 1, Zhou et al. Dataset, Li et al. Dataset 3**	**Transformer**	–	**Fu et al. Dataset 1: AUROC = 0.9739, AUPRC = 0.9812, Acc = 0.9726, F1-score = 0.9693, MCC = 0.9461; Zhou et al. Dataset: AUROC = 0.9642, AUPRC = 0.9616, Acc = 0.9196, F1-score = 0.9204, MCC = 0.8379; Li et al. Dataset: AUROC = 0.9681, AUPRC = 0.9623, Acc = 0.9203, F1-score = 0.9289, MCC = 0.8605**
		Li et al., 2024 [136]	**Li et al. D2 Dataset, Li et al. D3 Dataset**	**Transformer**	–	**Li et al. D2 Dataset: AUROC = 0.9630, AUPRC = 0.9624, F1-score = 0.9177, Acc = 0.9170, Recall = 0.9258, Sp = 0.9083, Precision = 0.9103; Li et al. D3 Dataset: UROC = 0.9419, AUPRC = 0.9437, F1-score = 0.8746, Acc = 0.8724, Recall = 0.8899, Sp = 0.8548, Precision = 0.8601**
		Yao et al., 2024 [162]	Yao et al. Dataset	GAT	CatBoost + ET + LightGBM + RF + XGBoost + LR	AUROC = 0.9907, AUPRC = 0.9927, MCC = 0.9249, F1-score = 0.9631, Acc = 0.9624
		Chen et al., 2024 [163]	Chen et al. Dataset 1, Chen et al. Dataset 2	GCN	SVM + GBDT + AdaBoost + XGBoost + RF + MLP	Chen et al. Dataset 1: AUROC = 0.8015; Chen et al. Dataset 2: AUROC = 0.8276

RNA interaction prediction. A combined potential of Word2vec and variational graph autoencoder based representation learning is explored with DNN classifier for coding RNA-protein interaction prediction [185]. Similarly, Word2vec is used in conjunction with nucleotide composition encoder along with LogR classifier for protein-RNA binding sites prediction [177]. Among all Word2vec based predictive pipelines, Word2vec and AdaBoost classifier based predictive pipeline demonstrates state-of-the-art performance for non-coding RNA interaction prediction [187]. Second most commonly used representation learning approach is BERT and its potential is explored with a self classifier [174], CNN [173] and BiLSTM [176] classifier for protein-RNA binding sites prediction. Moreover, BERT based representation learning is used with GBDT [280]and XGBoost [183] classifiers for non-coding RNA interaction prediction. However, BERT based any predictive pipeline does not achieve state-of-the-art performance across any of 4 tasks in

Table 8 (*continued*)

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
		Zhou et al., 2024 [157]	**lncRNADisease Dataset, MNDR Dataset**	**SVD**	**AdaBoost-CNN + LightGBM**	**lncRNADisease Dataset: Precision = 0.8980+0.0306, Recall = 0.7709+0.0622, Acc = 0.8444+0.0445, F1-score = 0.8278+0.0363, AUROC = 0.9328+0.0243, AUPRC = 0.9304+0.0252; MNDR Dataset: Precision = 0.9494+0.0172, Recall = 0.8436+0.0513, Acc = 0.8989+0.0317, F1-score = 0.8925+0.0307, AUROC = 0.9675+0.0147, AUPRC = 0.9709+0.0106**
		Wang et al., 2024 [164]	Wang et al. Dataset 1	GCN	ET	AUROC = 0.9916, AUPRC = 0.9951
		Lu et al., 2023 [156]	**Lu et al. Dataset 1, Lu et al., Dataset 2, Zhang et al. Dataset**	**k-mer Composition**	**GCN**	**Lu et al. Dataset 1: AUROC = 0.95919, AUPRC = 0.96059; Lu et al. Dataset 2: AUROC = 0.94037, AUPRC = 0.91658; Zhang et al. Dataset: AUROC = 0.9505, AUPRC = 0.94740**
		Zhang et al., 2023 [158]	**Li et al. Dataset 4, Ma et al. Dataset 1, Xia et al. Dataset**	**Stacked Auto Encoder**	**CNN**	**Li et al. Dataset 4: AUROC = 0.8863, AUPRC = 0.9079; Ma et al. Dataset: AUROC = 0.9013, AUPRC = 0.9182; Xia et al. Dataset: AUROC = 0.7629, AUPRC = 0.8027**
		Shi et al., 2022 [248]	Fu et al. Dataset	Word2Vec	BiLSTM	AUROC = 0.9786, AUPRC = 0.8891
		Madhavan et al., 2022 [262]	Madhavan et al. Dataset	Node2Vec	Neural Network Regression Model	AUROC = 0.96, AUPRC = 0.967
		Awn et al., 2022 [139]	Awn et al. Dataset	BERT	_	F1-score = 0.9072, Precision = 0.8410, Recall = 0.9848, AUROC = 0.9548
		Liang et al., 2022 [165]	Liang et al. Dataset	GCN	XGBoost + LightGBM + RF + ET + CatBoost	Acc = 0.9395, Sn = 0.9192, Sp = 0.9626, Precision = 0.9654, F1-score = 0.9417, MCC = 0.88
		Duan et al., 2021 [144]	Duan et al. DS1 Dataset, Duan et al. DS2 Dataset, Duan et al. DS3 Dataset	Graph2Vec	GBDT + LR	DS1 Dataset: Acc = 0.928, Recall = 0.920, F1-score = 0.927, MCC = 0.858, AUROC = 0.975; DS2 Dataset: Acc = 0.934, Recall = 0.928, F1-score = 0.934, MCC = 0.870, AUROC = 0.982; DS3 Dataset: Acc = 0.887, Recall = 0.871, F1-score = 0.885, MCC = 0.777, AUROC = 0.961
		Li et al., 2021 [259]	Li et al. Dataset	SVD + Node2Vec	XGBoost	Acc = 0.9460, MCC = 0.8922
		Xie et al., 2021 [261]	Fu et al. Dataset	Node2Vec	LogR	AUROC = 0.975
		Zhou et al., 2021 [258]	Zhou et al. Dataset	HOPE	Rotation Forest Model	AUROC = 0.8328 ± 0.0236
		Liu et al., 2020 [249]	Liu et al. Dataset	Word2Vec	Matrix. Factorization + MLP	5-fold Cross-Validation: AUROC = 0.904 ± 0.003; Leave-one-out Cross-validation: AUROC = 0.918 ± 0.002

(*continued on next page*)

Table 8 (*continued*)

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
Interaction	Micro RNA Target Prediction	Zhang et al., 2024 [239]	Pla et al. miRAW Dataset	BERT	_	F1-score = 0.81, Acc = 0.77
		Yang et al., 2024 [238]	**miRAW Dataset, DeepMirTar Dataset, deepTargetPro Dataset**	_	**CNN**	**miRAW: Acc = 95.71, Sn = 94.08, Sp = 97.42, PPV = 97.44, NPV = 94.03, F1-score = 95.73; DeepMirTar: Acc = 81.25, Sn = 81.25; deepTargetPro: Acc = 79.97, Sn = 78.56, Sp = 81.29, PPV = 79.67, NPV = 80.25, F1-score = 79.11**
		Przybyszewski et al., 2023 [240]	**miTAR Dataset: 1. miRAW, 2. DeepMirTar, 3. MirTarRaw**	**Word2Vec**	**GNN**	**DeepMirTar Dataset: Acc = 0.922, Precision = 0.923, Recall = 0.922; MiRAW Dataset: Acc = 0.948, Precision = 0.949, Recall = 0.948; MirTarRaw Dataset: Acc = 0.921, Precision = 0.921, Recall = 0.921**
		Sun et al., 2022 [253]	Mock Dataset, Experimental Data	Word2Vec	BiLSTM	Mock Dataset: Acc = 96.86, Sn = 96.97, Sp = 96.75, F1-score = 96.91; Experimental Data: Acc = 96.04, Sn = 95.65, Sp = 96.44, F1-score = 96.09
Interaction	Small Interfering RNA Target Prediction	**Xu et al., 2024 [241]**	**Huesken et al. Dataset, Reynold et al. Dataset + Katoh et al. Dataset, Xu et al. Dataset 1, Xu et al. Dataset 2, Xu et al. Dataset 3**	**BERT**	_	**Huesken Train Dataset: PCC = 0.636; Reynold et al. Dataset + Katoh et al. Dataset: PCC = 0.611, SRCC = 0.639; Xu et al. Dataset 1: PCC = 0.57; Xu et al. Dataset 2: PCC = 0.595; Xu et al. Dataset 3: PCC = 0.669**

this goal. Apart from Word2vec and BERT, Transformer with a self-classifier achieves state-of-the-art performance for protein-RNA binding affinity prediction [182]. Similarly, combined potential of ELMo and ESM-1b representation learning approach along with XGBoost classifier manages to achieve state-of-the-art performance for protein-RNA binding sites identification [172]. In addition, Struc2vec representation learning approach is employed with CatBoost classifier for non-coding RNA interaction prediction [186]. Furthermore, Node2vec is utilized with GNN classifier [166,260], HIN2vec is used with SVM classifier [169], and nucleotide frequency and density based representation learning is used with GCN classifier [171] for coding RNA-protein interaction. Overall, among all predictive pipelines for coding RNA-protein interaction prediction, nucleotide frequency and density based representation learning along with GCN classifier based predictive pipeline manages to achieve state-of-the-art performance [171]. From all 4 tasks, protein-RNA binding affinity prediction offers some room for improvement. Taking into account the performance trend of other tasks in this goal, shallow neural network based word embedding such as Word2vec and hybrid representation learning approach (ELMo+ ESM-1b) with boosting classifiers namely AdaBoost and XGBoost can raise the predictive performance of this task.

Table 10 provides a holistic overview of 16 different predictive pipelines developed for 8 different tasks classified under the hood of 3 distinct goals namely RNA Subcellular Localization Prediction, RNA Sites Prediction, and Gene Analysis.

For RNA subcellular localization prediction, 4 different predictive pipelines are developed that use 4 unique representation learning approaches namely BERT, Word2vec, GraRep and Transformer with 4 unique classifiers namely BERT-self classifier, ensemble (CNN+GRU) classifier, Transformer-self classifier and LSTM classifier. It is important to mention that for RNA subcellular localization prediction, overall, 3 different benchmark datasets are used for the development and validation of 4 different predictive pipelines. Specifically, 2 studies [16,17] use Liu et al. benchmark dataset [17], whereas, 1 study [19] uses Zeng et al. benchmark dataset. Liu et al. [17] and Zeng et al. [19] benchmark datasets contain sequences only related to long non-coding RNA subcellular localization prediction. In contrast, 1 study makes use of Asim et al., dataset [18] that has coding and non-coding sequences related to 4 different types of RNAs namely miRNA, mRNA, snoRNA, and lncRNA for the task of RNA subcellular localization prediction. Considering the direct impact of coding RNA subcellular localization on the production of proteins, and influence of non-coding RNAs in the regulation of protein synthesis, Asim et al. dataset [18] holds greater value as it identifies subcellular compartments of more diverse array of RNAs.

Furthermore, for another biological goal namely RNA sites prediction, 3 predictive pipelines are developed. In these predictive pipelines, 2 unique representation learning approaches namely BERT [213,214] and Word2vec [215]are used with self-classifier, and CNN classifier. BERT with a self classifier achieves state-of-the-art performance for RNA splicing sites prediction [213], and Word2vec with CNN classifier achieves state-of-the-art performance for alternative splicing sites prediction [215]. Both splicing sites prediction, and alternative splicing sites prediction tasks in this goal offer some room for improvements. Potential of nucleotide

Table 9
RNA Interaction Prediction related 4 distinct RNA sequence analysis tasks predictive pipelines performance.

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
Interaction	Coding RNA-Protein Interaction Prediction	**Wang et al., 2024 [171]**	**RPI369 Dataset, RPI488 Dataset, RPI1446 Dataset, RPI1807 Dataset, RPI2241 Dataset**	**k-mer Composition + DCC + KGap Descriptors + PseTNC + Conjoint Triad + GDPC + QSOrder Descriptors + DDE + ACC**	**GCN**	**RPI369: Acc = 97.27; RPI488: Acc = 97.32; RPI1446: Acc = 96.54; RPI1807: Acc = 95.76; RPI2241: Acc = 94.98**
		Li et al., 2024 [167]	**Li et al. Dataset (DB1, DB2, DB3, DB4)**	**Word2Vec**	**MLP**	**DB1: AUROC = 95.51 ± 0.36, AUPRC = 94.24 ± 0.61, Acc = 89.95 ± 0.67, Precision = 87.44 ± 1.00, Recall = 93.31 ± 0.64, F1-score = 90.28 ± 0.61; DB2: AUROC = 97.31 ± 0.31, AUPRC = 96.80 ± 0.47, Acc = 92.30 ± 0.47, Precision = 92.12 ± 0.44, Recall = 92.51 ± 0.94, F1-score = 92.31 ± 0.49; DB3: AUROC = 95.47 ± 0.32, AUPRC = 93.87 ± 0.74, Acc = 91.02 ± 0.24, Precision = 87.67 ± 0.66, Recall = 95.49 ± 0.83, F1-score = 91.41 ± 0.23; DB4: AUROC = 96.46 ± 0.34, AUPRC = 94.91 ± 0.76, Acc = 92.83 ± 0.28, Precision = 90.10 ± 0.59, Recall = 96.23 ± 0.38, F1-score = 93.06 ± 0.25**
		Han et al., 2023 [166]	**NPInter2.0 Dataset, RPI7317 Dataset**	**Node2Vec**	**GNN**	**NPInter2.0: Sn = 98.2 ± 0.2, Sp = 95.0 ± 0.2, Precision = 95.1 ± 0.2, Acc = 96.6 ± 0.1, MCC = 0.932 ± 0.002; RPI7317: Sn = 94.5 ± 0.4, Sp = 91.3 ± 0.8, Precision = 92.0 ± 0.3, Acc = 93.1 ± 0.1, MCC = 0.863 ± 0.002**
		Wei et al., 2023 [169]	Wei et al. Dataset	HIN2Vec	SVM	AUROC = 0.97, Acc = 0.95, Precision = 0.932, Recall = 0.981, Sp = 0.928, MCC = 0.9102, F1-score = 0.956
		Zhao et al., 2023 [168]	**Zhao et al. Dataset 1, Zhao et al. Dataset 2**	**VGAE + Word2Vec**	**GAE**	**Dataset 1: AUROC = 0.974, AUPRC = 0.7688, Acc = 0.9851, F1-score = 0.6397, Precision = 0.4238; Dataset 2: AUROC = 0.9734, AUPRC = 0.9421, Acc = 0.9305, F1-score = 0.8534, Precision = 0.7871**
		Shen et al., 2021 [260]	NPInter2.0 Dataset, RPI7317 Dataset, RPI2241 Dataset, RPI38318 Dataset	Node2Vec	GNN	NPInter2.0: Acc = 93.3, Sn = 95.6, Sp = 91.1, Precision = 91.5, MCC = 0.868; RPI7317: Acc = 91.5, Sn = 92.7, Sp = 90.7, Precision = 90.7, MCC = 0.830; RPI2241: Acc = 62.6, Sn = 49.8, Sp = 74.8, Precision = 67.2, MCC = 0.270; RPI369: Acc = 60.2, Sn = 61.5, Sp = 58.9, Precision = 60.0, MCC = 0.212

(*continued on next page*)

Table 9 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
		Yi et al., 2020 [170]	RPI369 Dataset, RPI1807 Dataset, RPI488 Dataset	Word2Vec	RF	RPI369 Dataset: Acc = 73.06, Sn = 75.32, Sp = 71.14, Precision = 72.64, MCC = 46.67; RPI488 Dataset: Acc = 89.92, Sn = 82.75, Sp = 96.72, Precision = 96.32, MCC = 80.59; RPI1807 Dataset: Acc = 97.10, Sn = 97.89, Sp = 96.14, Precision = 96.91, MCC = 94.13;
Interaction	Protein-RNA Binding Sites Prediction	**Yan et al., 2024 [172]**	**Yan et al. Dataset**	**ELMo + ESM-1b**	**XGBoost**	**Acc = 0.922, Sn = 0.926, MCC = 0.757**
		Lasantha et al., 2024 [173]	**circRNA fragment Dataset 2**	**BERT**	**CNN**	**circRNA fragment Dataset 1: AUROC = 0.957 ± 0.031**
		Qiao et al., 2024 [180]	**Qiao et al. Dataset 1. RBP-120 Dataset, Maticzka et al. Dataset 2. RBP-24 Dataset**	**One-hot Encoding**	**CNN + BiLSTM**	**RBP-24 Dataset: AUROC = 0.952; RBP-120 Dataset: AUROC = 0.874**
		Liu et al., 2024 [181]	Liu et al. Dataset	Hybrid Nucleotide Frequencies + Nucleotide Density + Nucleotide Chemical Property + diNucleotide Physiochemical Properties	CNN + BiGRU	AUROC = 0.9135, Acc = 0.8407, Precision = 0.8398, Recall-0.8444, F1-score = 0.8407
		Cao et al., 2024 [175]	**Cao et al. Dataset**	**Transformer**	_	**AUROC = 0.977**
		Liu et al., 2023 [177]	**Liu et al. Dataset**	**Word2Vec + PseTNC + PSTNP + TNC**	**BiLSTM + LogR**	**AUROC = 0.9362**
		Ma et al., 2023 [178]	**Wang et al. Dataset**	**Word2Vec**	**MLP**	**LIN28A Dataset: AUROC = 0.9911 ± 0.0016, Acc = 0.9699 ± 0.0026, Precision = 0.9715 ± 0.0043, Recall = 0.9684 ± 0.0044, F1-score = 0.9699 ± 0.0021**
		Jin el al., 2023 [174]	**Jin et al. Protein Dataset**	**BERT**	_	**AUROC = 0.842, AUPRC = 0.643**
		Li el al., 2023 [176]	**Jia et al. Dataset 1. 37 circRNA Datasets, Zhang et al. Dataset 2. 31 linear RNA Dataset**	**BERT**	**BiLSTM**	**37 CircRNA Dataset: AUROC = 0.9385; 31 Linear RNA Dataset: AUROC = 0.9393**
		Cao et al., 2023 [278]	Jia et al. Dataset 1. 37 circRNA Datasets, Zhang et al. Dataset 2. 31 linear RNA Dataset	BERT	_	37 CircRNA Dataset: Average AUROC = 0.931 ± 0.054; 31 Linear RNA Dataset: Average AUROC = 0.931
		Du et al., 2022 [279]	Jia et al. Dataset 1. 37 circRNA Datasets	BERT	BiLSTM	AUROC = 93.68, AUPRC = 90.28, Acc = 86.72, Precision = 86.47, Recall = 87.53, F1-score = 86.90

Table 9 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
		Wang et al., 2021 [250]	Wang et al. Dataset (RBP Datasets: 1. IGF2BP1, 2. IGF2BP3, 3. LIN28A, 4. LIN28B)	Word2Vec	BiLSTM + LSTM	LIN28B Dataset: Precision = 0.9174, Recall = 0.8999, F1-score = 0.9086, Acc = 0.9095, AUROC = 0.9570
		Deng et al., 2020 [179]	**Strazar et al. Dataset**	**Word2Vec**	**CNN + BiLSTM**	**AUROC = 0.873**
		Deng et al., 2019 [251]	Maticzka et al. Dataset 1. RBP-24 Dataset, Stražar et al. Dataset 2. RBP-31 Dataset	Word2Vec	CNN + BiLSTM	RBP-24: AUROC = 0.943; RBP-31: AUROC = 0.873
		Pan et al., 2018 [7]	RBP-24 Dataset	Word2Vec	CNN	AUROC = 0.916
Regression	Protein-RNA Binding Affinity Prediction	**Shen et al., 2024 [182]**	**Shen et al. Benchmark Dataset**	**Transformer**	_	**PCC = 0.85**
Interaction	Non-coding RNA Interaction Prediction	**Sheng et al., 2023 [187]**	**1. Fu et al. Dataset, 2. Zhou et al. Dataset**	**Word2Vec**	**1. Adaboost, 2. RF**	**Fu et al. Dataset: AUROC = 0.967, AUPRC = 0.224; Zhou et al. Dataset: AUROC = 0.974, AUPRC = 0.132**
		Zhao et al., 2022 [184]	**Zhao et al. Dataset**	**Word2Vec**	**GCN**	**Acc = 87.09, Precision = 87.66, Sn = 87.03, Sp = 87.84, MCC = 74.18, F1-score = 86.99**
		Guo et al., 2024 [185]	**Wang et al. Dataset 1. CMI-9905, Liu et al. Datasets 2. CMI-9589, 3. CMI-20208**	**Word2Vec, CAE**	**DNN**	**CMI-9905: AUROC = 0.9138, AUPRC = 0.9088; CMI-9589: AUROC = 0.9156, AUPRC = 0.9086; CMI-20208: AUROC = 0.9170, AUPRC = 0.9131**
		Zhou et al., 2024 [280]	CMI-9905 Dataset	BERT	GBDT	AUROC = 0.9143
		Wang et al., 2023 [186]	CMI-753 Dataset	Struc2Vec	CatBoost	CMI-753 Dataset AUROC = 0.8187, AUPRC = 0.8081
		Wei et al., 2023 [183]	CircBank Dataset	BERT	XGBoost	CircBank Dataset: AUROC = 0.9463, AUPRC = 0.9405

physico-chemical properties and occurrence frequency based representation learning approaches along with ensemble machine or deep learning classifiers can be explored to enhance the predictive performance on these tasks.

For gene analysis goal, 4 unique representation learning approaches namely Transformer, BERT, k-mer composition and Word2vec are used with 6 classifiers namely CNN, RF, BERT-self classifier, NB, Transformer-self classifier and ensemble (SVM + Ridge Regression) for the development of 9 different predictive pipelines across 5 different tasks. Most commonly used representation learning approach is Transformer followed by Word2vec. Transformer is employed with a self classifier for 3 tasks namely spatial gene expression analysis [231], gene expression prediction [233,277], and cell-specific gene regulatory networks prediction [235]. Among all of Transformer based predictive pipelines, Transformer with a self classifier achieves state-of-the-art performance for 2 tasks: spatial gene expression analysis [231], and cell-specific gene regulatory networks prediction [235]. Second most commonly used representation learning approach Word2vec is used with RF [252], CNN [252] and NB [236] classifiers for 16S rRNA taxonomic classification and has achieved state-of-the-art performance with CNN classifier [252]. Apart from Transformer and Word2vec representation learning approaches, BERT is used with a self classifier for gene expression prediction task [234]and potential of k-mer composition representation learning is explored with ensemble (SVM+ Ridge Regression) classifier for 16s rRNA gene copy number prediction [237]. From 5 distinct tasks in gene analysis goal, spatial gene expression analysis, cell specific gene regulatory networks prediction and 16S rRNA gene copy number prediction offers room for improvements. Considering performance trends of all predictive pipelines developed in this goal, potential of shallow neural network word embeddings namely Word2vec and BERT representation with standalone or hybrid deep neural networks can be explored to improve the performance of these tasks.

Table 10
RNA Subcellular Localization Prediction, RNA Sites Prediction, and Gene Analysis related 8 distinct RNA sequence analysis tasks predictive pipelines performance.

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
Goal: RNA Subcellular Localization Prediction						
Multi-Class/ Multi-Label Classification	RNA Subcellular Localization Prediction	Zhang et al., 2024 [16]	Liu et al. Dataset	BERT	_	Micro AUROC = 0.791
		Liu et al., 2023 [17]	Liu et al. Dataset	Word2Vec	CNN + GRU	Acc = 0.6256, Macro F1-score = 0.6091, MCC = 0.2378, AUROC = 0.6599
		Zeng et al., 2023 [19]	**Zeng et al. Benchmark Dataset**	**Transformer**	_	**Average F1-score = 0.719, Micro Precision = 0.683, Micro Recall = 0.721, Micro F1-score = 0.701**
		Asim et al. 2022 [18]	Asim et al. Dataset 1. Homo Sapien a. miRNA b. mRNA c. snoRNA d. lncRNA 2. Mus Musculus a. miRNA b. mRNA c. snoRNA d. lncRNA	GraRep	LSTM	Human miRNA: Average Precision = 0.86, Acc = 0.63, Coverage = 0.70, Ranking Loss = 0.11, One error = 0.26 mRNA: Average Precision = 0.77, Acc = 0.46, Coverage = 0.68, Ranking Loss = 0.23, One error = 0.35 snoRNA: Average Precision = 0.83, Acc = 0.55, Coverage = 0.45, Ranking Loss = 0.17, One error = 0.20 lncRNA: Average Precision = 0.85, Acc = 0.55, Coverage = 0.45, Ranking Loss = 0.17, One error = 0.20 Mouse miRNA: Average Precision = 0.87, Acc = 0.69, Coverage = 0.50, Ranking Loss = 0.10, One error = 0.28 mRNA: Average Precision = 0.71, Acc = 0.37, Coverage = 0.87, Ranking Loss = 0.13, One error = 0.40 snoRNA: Average Precision = 0.82, Acc = 0.56, Coverage = 0.29, Ranking Loss = 0.20, One error = 0.20 lncRNA: Average Precision = 0.77, Acc = 0.47, Coverage = 0.60, Ranking Loss = 0.18, One error = 0.36
Goal: RNA Sites Prediction						
Binary Classification	RNA-Splicing Sites Prediction	**Chen et al., 2024 [213]**	**Chen et al. Dataset**	**BERT**	_	**Zebrafish: F1-score = 0.9568 Fruit: F1-score = 0.9461 Worm: F1-score = 0.9343 Arabidopsis: F1-score = 0.9361**
		Mo et al., 2021 [214]	**Jaganathan et al. Datasets: 1.SpliceAI-2k**	**BERT**	_	**SpliceAI-2k: Acc = 0.97, AUPRC = 0.99**
Binary Classification	Alternative Splicing Prediction	Oubounyt et al., 2018 [215]	Brawand et al. Dataset (Brain, Heart, Kidney, Liver, Testis)	Word2Vec	CNN	Brain: Low AUROC = 93.0 ± 0.4, Medium AUROC = 73.9 ± 1.5, High AUROC = 92.8 ± 0.3; Heart: Low AUROC = 96.1 ± 0.2, Medium AUROC = 77.3 ± 1.0, High AUROC = 95.8 ± 0.1; Kidney: Low AUROC = 96.0 ± 0.9, Medium AUROC = 80.1 ± 1.3, High AUROC = 95.8 ± 0.3; Liver: Low AUROC = 97.1 ± 0.5, Medium AUROC = 90.9 ± 0.8, High AUROC = 97.0 ± 0.6; Testis: Low AUROC = 89.2 ± 0.3, Medium AUROC = 73.5 ± 0.9, High AUROC = 89.3 ± 0.5

Table 10 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
Goal: Gene Analysis						
Interaction	Spatial Gene Expression Analysis	**Wang et al., 2024 [231]**	**scRNA-seq Dataset**	**Transformer**	_	**AUROC = 91.30**
Binary Classification	Gene Expression Prediction	Babjac et al., 2023 [234]	Babjac et al. Dataset	BERT	_	AUROC = 0.81, Acc = 0.62, Precision = 0.62, Recall = 0.62
		Khan et al., 2023 [232]	Khan et al. Datasets: 1. LUAD Dataset 2. LUSC Dataset	_	CNN	1. Acc = 0.9984 2. Acc = 0.9585
		Zhang et al., 2022 [233]	PBMC scRNA-Seq Dataset	Transformer	_	TCGA RNA-Seq Dataset: Acc = 94.92, MCC = 0.9469, AUROC = 0.9987 PBMC scRNA-Seq Dataset: Acc = 90.73, MCC = 0.8971, AUROC = 0.9964
		Khan et al., 2021 [277]	**LUAD Dataset**	**Transformer**	_	**Acc = 0.9868, AUROC = 0.9966, Precision = 0.9883, Recall = 0.9883, F1-score = 0.9883, MCC = 0.9617**
Binary Classification	Cell-Specific Gene Regulatory Networks Prediction	**Xu et al., 2023 [235]**	**Yuan et al. Balanced Benchmark Datasets**	**Transformer**	_	**AUROC = 85.71, AUPRC = 85.71**
Multi-Class Classification	16S rRNA Gene Copy Number Prediction	**Miao et al., 2022 [237]**	**Miao et al. 16S rRNA gene Dataset**	**k-mer Composition**	**SVM + Ridge Regression**	**RMSE = 0.685, SD = 0.0379**
Multi-Class Classification	16S rRNA Taxonomic Classification	Ziemski et al., 2021 [252]	McDonald et al. Greengenes Dataset	Word2Vec	RF, CNN	_
		Woloszynek et al., 2019 [236]	**Woloszynek et al. Dataset**	**Word2Vec**	**NB**	**Acc = 0.977, Precision = 0.971, Recall = 0.964, F1-score = 0.968**

Table 11 provides performance analysis of 7 distinct sequence analysis tasks classified under the hood of RNA modification prediction goal. Overall, for this goal, predictive pipelines have used 12 unique representation learning approaches namely BERT, one-hot encoding, SocDim,+Node2vec+GraRep, Word2vec, ELMo, NCP+EIIP, tSNE, nucleotides composition encoders, transformer, BiPSTP, GloVe, and nucleotide and physico-chemical properties aware encoder. Along with these representation learning approaches, 16 unique classifiers including CNN, MLP, FGM, SVM, ElMo-self classifier, BERT-self classifier, LightGBM, Transformer-self classifier, ensemble (LSTM+CNN), ensemble (LightGBM+SVM+LR), ensemble (CNN+DNN), ensemble (BiGRU+CNN), RF, XGBoost, BiLSTM, and ElasticNet Regression model are used.

For this goal, most commonly used representation learning approach is BERT followed by Word2vec and Transformers. Specifically BERT is used with BiLSTM classifier for ac4C-Acetylcytidine Modification Prediction [188] and with CNN classifier for 2'-OmU Methyluridine Modification Prediction [191]. In addition, BERT is also employed with a self-classifier for two tasks namely 6mA-methyladenine modification prediction [281] and Methylation modification prediction [193,282,10]. BERT is also used with FGM classifier for Methylation modification prediction [193]. Moreover, BERT representation learning is used with ensemble (LightGBM+SVM+LogR) for methylguasnosine modification prediction [204]. Among all BERT based predictive pipelines, BERT with BiLSTM and CNN classifiers has achieved state-of-the-art performance for ac4C-Acetylcytidine Modification Prediction [188] and 2'-OmU Methyluridine Modification Prediction [191], respectively. Second most commonly used representation is Word2vec which is used with 4 different classifiers namely MLP, CNN, RF and ensemble (LSTM+CNN). Potential of Word2vec representation is explored with MLP classifier for RNA methylation modification prediction [210], and with CNN classifier for 5mC-methyl cytosine [207] and 6mA-methyl adenine modification prediction [199]. Also, Word2vec is used with RF classifier for 6mA-methyl adenine modification prediction [197] and with hybrid (LSTM+CNN) classifier for methyl guanosine modification prediction [205]. Overall, among all Word2vec based predictive pipelines, Word2vec representation learning along with RF classifier has achieved state-of-the-art performance for 6mA-methyl adenine modification prediction [197]. In addition, Transformer is used with a self classifier for 3 different tasks namely 6mA-methyl adenine modification prediction [195], RNA methylation modification [209], and methyl-

Table 11
RNA Modification Prediction related 7 distinct RNA sequence analysis tasks predictive pipelines performance.

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
Binary Classification	ac4C-Acetylcytidine Modification Prediction	**Li et al., 2024 [188]**	**Wang et al. Dataset (Balanced and Unbalanced)**	**BERT**	**BiLSTM**	**Balanced Dataset: Sn = 79.22, Sp = 84.36, Acc = 81.79, MCC = 0.6368, AUROC = 0.8749; Unbalanced Dataset: Sn = 80.94, Sp = 84.8, Acc = 82.87, MCC = 0.6579, AUROC = 0.8951**
Binary Classification	5mU-Methyluridine Modification Prediction	**Alam et al., 2024 [190]**	**GSE78040 Dataset, GSE63753 Dataset**	**One-hot Encoding**	**CNN**	**GSE78040 Dataset: Acc = 91.26; GSE63753 Dataset: Acc = 95.63**
		Xu et al., 2023 [189]	**Jiang et al. Dataset**	**SocDim + Node2Vec + GraRep**	**XGBoost**	**Sn = 93.56, Sp = 93.90, Acc = 93.73, MCC = 0.875, AUROC = 0.984**
Binary Classification	2'-OmU Methyluridine Modification Prediction	**Soylu et al., 2023 [191]**	**Human Dataset, S.cerevisiae Dataset, M.musculus Dataset**	**BERT**	**CNN**	**Human Dataset: Acc = 99.15, AUROC = 0.99; M. musculus Dataset: Acc = 94.35, AUROC = 0.94; S. cerevisiae Dataset: Acc = 97, AUROC = 0.98**
Binary Classification	6mA-Methyladenosine Modification Prediction	Ye et al., 2024 [197]	Zhang et al. Dataset	Word2Vec	RF	_
		Li et al., 2024 [254]	Human Dataset	ELMo	_	Acc = 0.872, MCC = 0.745, Sn = 0.873, Sp = 0.870
		Tu et al., 2024 [200]	**Tu et al. Dataset**	**NCP, EIIP**	**SVM**	**Cross-Validation: Sn = 0.795, Sp = 0.789, Acc = 0.792, MCC = 0.584, AUROC = 0.871; Independent Test: Sn = 0.806, Sp = 0.796, Acc = 0.801, MCC = 0.603, AUROC = 0.879**
		Jiang et al., 2024 [202]	Song et al. Dataset	tSNE	Elastic Net Regression Model	Average R2 = 0.49, Median R2 = 0.486
		Wang et al., 2024 [201]	**Wang et al. Dataset**	**One-hot Encoding**	**CNN**	**AUROC = 77.13**

guanosine modification prediction [203] and has achieved state-of-the-art performance for methylguanosine modification prediction [203]. In addition, ELMo is used with a self classifier and DiNucleotide based representation learning is employed with ensemble (Bi-GRU+CNN) classifier for 6mA-methyl adenine modification prediction [255], whereas GLoVe is used with CNN for RNA methylation modification prediction [211]. Beyond word embeddings and language models, BiPSTP representation learning is used with SVM for methylation modification prediction [212], one-hot encoding is used with ensemble (CNN+DNN) classifier for methylguanosine modification prediction [206], and combined potential of DNC and TNC representation learning is used with LightGBM for 5mC-methyl cytosine modification prediction [208], respectively. Overall, among all predictive pipelines, BIPSTP with SVM classifier has achieved state-of-the-art performance for RNA methyaltion modification prediction [212]. Similarly, combined potential of DNC and TNC with LightGBM classifier has achieved state-of-the-art performance for 5mC-methyl cytosine modification prediction [208]. From all 9 different tasks, ac4C-Acetylcytidine Modification Prediction, 5mU-Methyluridine Modification Prediction offer some room for improvements. Considering the performance trends for this goal, potential of shallow neural network embedding such as Word2vec and graph based transformers along with hybrid deep learning classifiers can enhance the predictive performance of under-performing tasks.

Table 12 provides the summary of 16 different predictive pipelines developed for 3 distinct tasks classified under the hood of RNA function and structure prediction goal. Overall, 7 unique representation learning approaches namely LINE, RNAformer, Transformer, BERT, one-hot encoding, BCM+encoder decoder network, and Word2vec are used for this goal. Along with these representation learning approaches, 8 unique classifier including MLP, CNN, RNAformer-self classifier, BERT-self classifier, BiLSTM, GNN, Transformer-self classifier and ensemble (CNN+RNN) are used for developing different predictive pipelines.

For this goal, most commonly used representation learning approaches are BERT and transformer followed by Word2vec, LINE and RNAformer. Specifically, BERT is used with a self-classifier in 3 predictive pipelines developed for RNA structure prediction [226,222,223]. Similarly, Transformer is used with a self classifier in 3 predictive pipelines developed for RNA function prediction and structure prediction [217,224,221]. Second most commonly used representation learning approach: Word2vec [228], LINE [219], and RNAformer [220], are used with hybrid (CNN+RNN) [228], MLP [219], and a self classifier [220], respectively. In addition, one

Table 11 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
		Huang et al. 2024 [196]	**Dao et al. Dataset (Human Brain Dataset, Human Liver Dataset, Human Kidney Dataset, Mouse Brain Dataset, Mouse Liver Dataset, Mouse Kidney Dataset, Rat Brain Dataset, Rat Liver Dataset, Rat Kidney Dataset, Rat Heart Dataset, Rat Testis Dataset)**	**diNucleotides One-hot Encoding + NPC**	**BiGRU + CNN**	**5-fold Cross-Validation h-b: Acc = 0.7499, Sn = 0.8108, Sp = 0.6882, MCC = 0.503, AUROC = 0.8252; h-k: Acc = 0.8131, Sn = 0.8585, Sp = 0.7678, MCC = 0.6287, AUROC = 0.8889; h-l: Acc = 0.8238, Sn = 0.8414, Sp = 0.8065, MCC = 0.6488, AUROC = 0.893; m-b: Acc = 0.7998, Sn = 0.8451, Sp = 0.7543, MCC = 0.6021, AUROC = 0.8821; m-h: Acc = 0.7647, Sn = 0.8366, Sp = 0.692, MCC = 0.5364, AUROC = 0.8335; m-k: Acc = 0.8263, Sn = 0.8555, Sp = 0.7967, MCC = 0.6543, AUROC = 0.9001; m-l: Acc = 0.7361, Sn = 0.8132, Sp = 0.6577, MCC = 0.4783, AUROC = 0.8087; m-t: Acc = 0.7719, Sn = 0.8373, Sp = 0.7053, MCC = 0.5492, AUROC = 0.8498; r-b: Acc = 0.79, Sn = 0.8145, Sp = 0.7652, MCC = 0.5811, AUROC = 0.8607; r-k: Acc = 0.8388, Sn = 0.8448, Sp = 0.8331, MCC = 0.678, AUROC = 0.9107; r-l: Acc = 0.8246, Sn = 0.855, Sp = 0.7935, MCC = 0.6503, AUROC = 0.8888; Independent Test h-b: Acc = 0.7510, Sn = 0.8082, Sp = 0.6937, MCC = 0.5053, AUROC = 0.8300; h-k: Acc = 0.8093, Sn = 0.8231, Sp = 0.7955, MCC = 0.6189, AUROC = 0.8885; h-l: Acc = 0.8153, Sn = 0.8383, Sp = 0.7923, MCC = 0.6313, AUROC = 0.8907; m-b: Acc = 0.7989, Sn = 0.8510, Sp = 0.7468, MCC = 0.6010, AUROC = 0.8835; m-h: Acc = 0.7605, Sn = 0.8623, Sp = 0.6586, MCC = 0.5321, AUROC = 0.8384; m-k: Acc = 0.8163, Sn = 0.8016, Sp = 0.8310, MCC = 0.6329, AUROC = 0.8961; m-l: Acc = 0.7334, Sn = 0.8224, Sp = 0.6443, MCC = 0.4743, AUROC = 0.8094; m-t: Acc = 0.7850, Sn = 0.8198, Sp = 0.7501, MCC = 0.5713, AUROC = 0.8629; r-b: Acc = 0.7850, Sn = 0.7733, Sp = 0.7967, MCC = 0.5701, AUROC = 0.8647; r-k: Acc = 0.8403, Sn = 0.8622, Sp = 0.8185, MCC = 0.6813, AUROC = 0.9154; r-l: Acc = 0.8184, Sn = 0.8314, Sp = 0.8053, MCC = 0.6370, AUROC = 0.8956**

(*continued on next page*)

Table 11 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
		Li et al., 2023 [281]	Human Brain Dataset, Human Liver Dataset, Human Kidney Dataset, Mouse Brain Dataset, Mouse Liver Dataset, Mouse Kidney Dataset, Rat Brain Dataset, Rat Liver Dataset, Rat Kidney Dataset, Rat Heart Dataset, Rat Testis Dataset	BERT	_	H_b Dataset: Acc = 0.747, Sn = 0.812, Sp = 0.681, MCC = 0.498, AUROC = 0.827; *H_k* Dataset: Acc = 0.806, Sn = 0.838, Sp = 0.775, MCC = 0.614, AUROC = 0.888; *H_l* Dataset: Acc = 0.815, Sn = 0.857, Sp = 0.773, MCC = 0.632, AUROC = 0.89; *M_b* Dataset: Acc = 0.792, Sn = 0.806, Sp = 0.775, MCC = 0.582, AUROC = 0.876; *M_h* Dataset: Acc = 0.757, Sn = 0.831, Sp = 0.684, MCC = 0.521, AUROC = 0.835; *M_k* Dataset: Acc = 0.819, Sn = 0.814, Sp = 0.824, MCC = 0.638, AUROC = 0.898; *M_l* Dataset: Acc = 0.736, Sn = 0.786, Sp = 0.686, MCC = 0.474, AUROC = 0.816; *M_t* Dataset: Acc = 0.78, Sn = 0.772, Sp = 0.789, MCC = 0.561, AUROC = 0.867; *R_b* Dataset: Acc = 0.783, Sn = 0.773, Sp = 0.793, MCC = 0.566, AUROC = 0.866; *R_k* Dataset: Acc = 0.838, Sn = 0.848, Sp = 0.828, MCC = 0.676, AUROC = 0.914; *R_l* Dataset: Acc = 0.82, Sn = 0.844, Sp = 0.796, MCC = 0.64, AUROC = 0.903
		Xiang et al., 2023 [195]	**Wan et al. A101 Dataset**	**Transformer**	_	**Acc = 0.8434, MCC = 0.6867, Sn = 0.8488, Sp = 0.8377**
		Fan et al., 2022 [255]	Human Dataset	ELMo	_	Sn = 0.8876, Sp = 0.8779, Acc = 0.8828, MCC = 0.7663, AUROC = 0.9541
		Nazari et al., 2019 [198]	**Chen et al. Dataset (S51, H41), Dominissini et al. Dataset (M41)**	**Word2Vec**	**CNN**	**S51 Dataset: Acc = 75.38, Sn = 76.15, Sp = 74.62, MCC = 0.5078; M4 Dataset: Acc = 89.51, Sn = 78.87, Sp = 100.0, MCC = 0.8079; H41 Dataset: Acc = 91.11, Sn = 82.14, Sp = 100.0, MCC = 0.8354**
		Zou et al., 2019 [199]	Zhou et al. Dataset	Word2Vec	CNN	AUROC = 0.841, AUPRC = 0.980
Binary Classification	7mG-Methyl-guanosine Modification Prediction	Zhang el al., 2024 [203]	Zhang et al. Dataset (Benchmark, Independent)	Transformer	_	Benchmark Dataset: Acc = 98.70; Independent Dataset: Acc = 92.92
		Zhang et al., 2023 [206]	**Chen et al. Dataset**	**One-hot Encoding**	**CNN + DNN**	**Acc = 92.6, F1-score = 91.1, Recall = 92.8, Precision = 91.4, MCC = 0.852, AUROC = 0.968, AUPRC = 0.969**
		Tahir et al., 2022 [205]	Chen et al. Dataset	Word2Vec	LSTM + CNN	Acc = 95.95, Sp = 95.94, Sn = 95.97, MCC = 0.919
		Zhang el al., 2021 [204]	Dai et al. Dataset	BERT	LightGBM + SVM + LR	Sn = 95.8, Sp = 95.1, Acc = 95.5, MCC = 0.910
Binary Classification	5mC-Methyl-cytosine Modification Prediction	**Kurata et al., 2024 [208]**	**Kurata et al. Dataset**	**DNC + TNC + RCk-mer + CKSNAP + PseEIIP**	**LightGBM**	**MCC = 0.841, Acc = 0.92, AUROC = 0.971**
		Hasan et al., 2022 [207]	**Hasan et al. Dataset**	**Word2Vec**	**CNN**	**MCC = 0.691, Acc = 0.852**

Table 11 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
Binary Classification	Methylation Modification Prediction	**Human Dataset**	**Human Dataset**	Bidirectional position-specific trinucleotide propensities (BiPSTP)	SVM	NmH2: Acc = 0.981, Sn = 1.000, Sp = 0.974, MCC = 0.956, AUROC = 1.000; AID: Acc = 0.960, Sn = 0.937, Sp = 0.983, MCC = 0.921, AUROC = 0.986; m5CA1: Acc = 1.000, Sn = 1.000, Sp = 1.000, MCC = 1.000, AUROC = 1.000; m5CA2: Acc = 0.920, Sn = 0.912, Sp = 0.928, MCC = 0.840, AUROC = 0.976; m5UH1: Acc = 0.938, Sn = 0.944, Sp = 0.932, MCC = 0.877, AUROC = 0.983; m5UH2: Acc = 0.982, Sn = 0.988, Sp = 0.976, MCC = 0.963, AUROC = 0.996; ΨS: Acc = 1.000, Sn = 1.000, Sp = 1.000, MCC = 1.000, AUROC = 1.000; ΨH: Acc = 0.995, Sn = 1.000, Sp = 0.990, MCC = 0.990, AUROC = 0.999; m6AmH1: Acc = 0.977, Sn = 0.983, Sp = 0.972, MCC = 0.955, AUROC = 0.997; m7GH1: Acc = 0.965, Sn = 0.620, Sp = 1.000, MCC = 0.773, AUROC = 0.993; m7GH2: Acc = 0.928, Sn = 0.919, Sp = 0.937, MCC = 0.857, AUROC = 0.980; m6AA: Acc = 0.986, Sn = 0.990, Sp = 0.982, MCC = 0.973, AUROC = 0.998; m6AS1: Acc = 0.806, Sn = 0.669, Sp = 0.820, MCC = 0.337, AUROC = 0.845; m6AH: Acc = 0.826, Sn = 0.842, Sp = 0.809, MCC = 0.652, AUROC = 0.901

(*continued on next page*)

hot encoding representation learning is used with MLP classifier [229] and combine potential of BCM and encoder decoder network is explored with CNN classifier [230] for RNA structure prediction. Overall, LINE representation learning with MLP classifier has achieved state-of-the-art performance for non-coding RNA function prediction while RNAformer with a self classifier has achieved state-of-the-art performance for RNA structure prediction. From all tasks of this goal, non-coding RNA function prediction has some room for improvement. Considering the performance trends in this goal, potential of large language models namely RNAformer, Transformer, and BERT with their own classifiers or separate deep learning classifiers can enhance the performance of under-performing task.

Table 13 provides a high-level overview of 6 different predictive pipelines developed for 3 distinct tasks classified under the hood of 2 different categories namely RNA special characteristics analysis and RNA single cell analysis.

For RNA special characteristics analysis goal, RNA sequence analysis tasks mostly belongs to regression. 2 transformer based predictive pipelines with a self-classier are used for mRNA degradation prediction [26], and RNA-Seq coverage prediction [27] and have achieved state-of-the-art performance. Considering the room for performance improvement in both tasks, potential of word embedding, physico-chemical properties and occurrence frequencies aware representation learning approaches can be explored with deep learning predictors.

Furthermore, for single-cell analysis goal, researchers have developed 4 different predictive pipelines using 3 unique representation learning approaches namely BERT, GPT, and non-negative matrix factorization (NNMF) and 4 unique classifiers namely kNN, BERT-self classifier, GPT-self classifier and 1 clustering algorithm for 3 different tasks. Most commonly used representation learning approach is BERT which is employed with kNN [283] and a self classifier [284] for single-cell RNA-Seq cell type detection. Both BERT based predictive pipelines achieved state-of-the-art performance across 3 benchmark datasets. Apart from BERT, potential of non-negative matrix factorization is explored with clustering algorithm for single-cell multi-omics cell type detection across 5 different benchmark datasets [180]. In addition, Cui et al., [242] explored the potential of GPT representation with a self classifier for 3 tasks namely cell type detection, scRNA-seq cell type detection, and scMultiomic cell type detection. Moreover, GPT based predictive pipeline manages to achieve top performing values across 9 different dataset for all 3 task. It is imperative to understand that single-cell RNA-seq and single-cell multi-omics analysis encompasses a multitude of diverse tasks. Consequently, this domain inherently possesses substantial room for enhancement. An in-depth analysis of all these studies uncovers that physico-chemical properties and occurrence frequencies based representation learning approaches along with ensemble classifiers, can enhance performance in this domain.

Table 11 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
		Wang et al., 2024 [193]	**DS_song Dataset (m5C, Am, Cm, Gm, Um, m6Am, 37G, AtoI, Psi)**	**BERT**	_	**m5C: Acc = 0.9440+0.0331, Sn = 0.9245+0.0416, Sp = 0.9632+0.0253, MCC = 0.8887+0.0657, AUROC = 0.9827+0.0122, F1-score = 0.9422+0.0346; Am: Acc = 0.9566+0.0294, Sn = 0.9324+0.0497, Sp = 0.9789+0.0121, MCC = 0.9141+0.0575, AUROC = 0.9815+0.0152, F1-score = 0.9532+0.0329; Cm Acc = 0.9567+0.0129, Sn = 0.9517+0.0145, Sp = 0.9602+0.0166, MCC = 0.9107+0.0264, AUROC = 0.9794+0.0072, F1-score = 0.973+0.0153; Gm: Acc = 0.9784+0.0102, Sn = 0.9669+0.0155, Sp = 0.9873+0.0101, MCC = 0.9562+0.0207, AUROC = 0.9933+0.0076, F1-score = 0.9750+0.0118; Um: Acc = 0.9429+0.0260, Sn = 0.9340+0.0272, Sp = 0.9511+0.0316, MCC = 0.8859+0.0520, AUROC = 0.9789+0.0119, F1-score = 0.9404+0.0268; m6Am: Acc = 0.8923+0.0266, Sn = 0.8339+0.0526, Sp = 0.9550+0.0139, MCC = 0.7927+0.0475, AUROC = 0.9544+0.0054, F1-score = 0.8884+0.0304; m7G: Acc = 0.8859+0.0579, Sn = 0.8589+0.0607, Sp = 0.9107+0.0684, MCC = 0.7729+0.1167, AUROC = 0.9373+0.0398, F1-score = 0.8786+0.0614; AtoI: Acc = 0.9230+0.0297, Sn = 0.8693+0.0608, Sp = 0.9631+0.0142, MCC = 0.8437+0.0597, AUROC = 0.9715+0.0163, F1-score = 0.9053+0.0394; Psi: Acc = 0.7522+0.0331, Sn = 0.6498+0.1207, Sp = 0.8369+0.0940, MCC = 0.5071+0.0604, AUROC = 0.8492+0.0238, F1-score = 0.6992+0.0655**
		Wang et al., 2024 [192]	**Wang et al. Dataset**	**BERT**	**FGM**	**Sn = 0.97, Sp = 0.98, AUROC = 0.99, MCC = 0.94**
		Wang et al., 2024 [210]	Chen et al. Dataset (m6A, m1A, m5C, m5U, m6Am, m7G, Ψ, I, Am, Cm, Gm, Um)	Word2Vec	MLP	m6A: AUROC = 98.34; m1A: AUROC = 85.41; m5C: AUROC = 97.29; m5U: AUROC = 96.74; m6Am: AUROC = 99.04; m7G: AUROC = 79.94; Ψ: AUROC = 76.22; I: AUROC = 65.69; Am: AUROC = 92.92; Cm: AUROC = 92.03; Gm: AUROC = 95.77; Um: AUROC = 89.66

Table 11 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
		Chen et al., 2023 [209]	**DS_song Dataset (I)**	**Transformer**	_	**Sn = 0.8000, Sp = 0.6200, Acc = 0.7100, MCC = 0.4270**
		Liang et al., 2023 [282]	DS_song Dataset (m1A, m5U, m6Am, Ψ)	BERT	_	m1A: Acc = 0.9376, MCC = 0.8752, Sn = 0.9406, Sp = 0.9345; m5U: Acc = 0.9662, MCC = 0.9323, Sn = 0.9648, Sp = 0.9676; m6A: Acc = 0.9246, MCC = 0.8492, Sn = 0.9264, Sp = 0.9228; Ψ: Acc = 0.8320, MCC = 0.6655, Sn = 0.7902, Sp = 0.8726
		Zhang et al., 2023 [10]	1. Zhang et al. M. musculus Dataset (m1A, m6A, m5C, Ψ); 2. Zhang et al. A. thaliana Dataset (m6A, m5C, Ψ); 3. Zhang et al. S. cerevisiae Dataset (m6A, m5C, Ψ)	BERT	_	1. M. musculus Dataset: m1A: AUROC = 1.000, Average Precision = 1.000; m6A: AUROC = 0.988, Average Precision = 0.983; m5C: AUROC = 0.997, Average Precision = 0.996; Ψ: AUROC = 0.840, Average Precision = 0.832; 2. A. thaliana Dataset: m6A: AUROC = 0.977, Average Precision = 0.956; m5C: AUROC = 0.949, Average Precision = 0.942; Ψ: AUROC = 0.830, Average Precision = 0.825; 3. S. cerevisiae Dataset: m6A: AUROC = 0.998, Average Precision = 0.997; m5C: AUROC = 1.000, Average Precision = 1.000; Ψ: AUROC = 0.732, Average Precision = 0.775
		Wang et al., 2022 [211]	Chen et al. Dataset (m1A site), Zou et al. Dataset (m6A site)	GloVe	CNN	m1A: AUROC = 95.56; m6A: AUROC = 85.24

In conclusion, comprehensive analysis of advanced predictive pipelines based on word embeddings, language models, and domain-specific representation learning methods reveals interesting trends. Among 47 RNA sequence analysis tasks classified into 10 main biological goals, 26 tasks belong to binary classification tasks, 8 tasks belong to interaction prediction, 5 tasks belong to multi-class classification, 1 task belongs to multi-label classification, 4 tasks belong to regression, 1 task belongs to clustering, 1 task namely RNA Subcellular localization is performed as multi-class classification task as well as multi-label classification task, and 1 task namely Cell Type detection is performed as a clustering task as well as multi-class classification task. In total, 38 distinct representation learning methods and 56 predictive algorithms are used to develop robust predictive pipelines for these tasks. Language models-based representation learning strategies and deep learning classifiers consistently achieve superior performance across majority of tasks within these 10 biological goals. Researchers should consider exploring potential of latest transformer-based language models, such as hierarchical and heterogeneous Graph transformer, GPT-4, and hybrid representation learning techniques along with advanced ensemble machine learning or deep learning predictors for various classification, regression, and clustering tasks.

11. Publisher and journal-wise distribution of research articles

This section provides a comprehensive overview of distribution of 47 RNA sequence analysis studies across various conferences, journals, and publishers. Selection of appropriate journals for submission of a study in interdisciplinary field of AI-based RNA sequence analysis is a critical step. There are primarily three types of journals relevant to this field: 1) journals dedicated to core AI algorithms, 2) journals focusing on biological findings, and 3) hybrid journals that integrate both biology and AI algorithms. Researchers often encounter desk rejections when submitting to core AI or biological journals due to narrow disciplinary focus of journal. To avoid this, researchers should target hybrid journals that bridge both domains. Numerous tools exist for identifying suitable journals. However, this comprehensive analysis provides in-depth information to assist researchers in identifying journals that have published applications of word embeddings and large language models for RNA sequence analysis.

Fig. 7 illustrates publication landscape of 172 RNA sequence analysis studies across 60 journals, 4 conferences, 2 transactions, and 2 pre-print repositories. Among all journals, most number of studies are published in Briefing in Bioinformatics followed by BMC Bioinformatics, Bioinformatics, Computational and Structural Biotechnology Journal. Similarly, among all conferences, more studies are published in IEEE International Conference on Bioinformatics and Biomedicine (IEEE-BIBM) followed by 2022 IEEE 24^{th} Inter-

Table 12
RNA Function and Structure Prediction related 3 distinct RNA sequence analysis tasks predictive pipelines performance.

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
Multi-label Classification	Non-coding RNA Functions Prediction	**Wang et al., 2019 [219]**	**Wang et al. Dataset miRNA2GO-337**	**LINE**	**MLP**	**AUROC = 0.8696, AUPRC = 0.4110, F1-score = 0.2693**
Multi-class Classification	RNA Structure Prediction	**Franke et al., 2024 [220]**	**Franke et al. Dataset (Rfam Dataset)**	**RNAformer**	_	**F1-score = 0.725, Precision = 0.765, Recall = 0.707**
		Penic et al., 2024 [221]	**Szikszai et al. Dataset**	**Transformer**	_	**Mean F1-score = 0.72**
		Gong et al., 2024 [226]	bpRNA-1m Dataset (TR0)	BERT	_	Acc = 0.460
		Zhang et al., 2024 [222]	**Zhang et al. Dataset 1**	**BERT**	_	**F1-score = 0.74**
		Kalicki et al., [223]	**Kalicki et al. Dataset**	**BERT**	_	**Acc = 0.70**
		Wang et al., 2023 [224]	**Wang et al. Dataset (30 Independent RNAs, CASP15)**	**Transformer**	_	**30 Independent RNAs: Average RMSD = 8.5+5.7; CASP15: Average RMSD = 7.4**
		Qiu et al., 2023 [229]	Tan et al. Dataset (Stralign), Solma et al. Dataset (ArchiveII)	One-hot Encoding, k-mer	MLP	_
		Chen et al., 2023 [230]	**RNAStralign Dataset, Chen et al. Dataset (ncRNA Benchmark)**	**BCM + Encoder Decoder Network**	**CNN**	**RNAStralign: Acc = 0.970, Sn = 0.974, PPV = 0.971, F1-score = 0.973; ncRNA Benchmark: Acc = 0.950, Sn = 0.952, PPV = 0.939, F1-score = 0.946**
		Fei et al., 2022 [225]	**Rfam, 5SrRNA, tRNA, PDB, SPR**	_	**BiLSTM**	**Rfam: Precision = 0.8599, Recall = 0.7897, F1-score = 0.8233; 5SrRNA: Precision = 0.9857, Recall = 0.9804, F1-score = 0.9831; tRNA: Precision = 0.9985, Recall = 0.9992, F1-score = 0.9988; PDB: Precision = 0.6695, Recall = 0.3050, F1-score = 0.4190; SPR: Precision = 0.9929, Recall = 0.9971, F1-score = 0.9950**
		Wang et al., 2020 [227]	RNA Straligh Datasets (1. tRNA, 2. 5S_rRNA, 3. Telomerase, 4. tmRNA)	_	CNN	tRNA: F1-score = 0.966, Positive Predictive Value = 0.972, Sn = 0.961; 5S_rRNA: F1-score = 0.927, Positive Predictive Value = 0.933, Sn = 0.923; Telomerase: F1-score = 0.816, Positive Predictive Value = 0.846, Sn = 0.791; tmRNA: F1-score = 0.66, Positive Predictive Value = 0.686, Sn = 0.64
		Zhao et al., 2021 [228]	SILVA 16S rRNA Dataset	Word2Vec	CNN + RNN	Acc = 0.742 ± 0.001
Multi-Class Classification	RNA Function and Structure Prediction	Shulgina et al., 2024 [216]	23S rRNA Sequence Dataset, 228 RNA Sequence Dataset	_	GNN	_
		Yin et al., 2024 [11]	bpRNA-1m Dataset	_	CNN	Average Binary F1-score = 0.748, Macro Average F1-score = 0.873, Recall = 0.867, Precision = 0.887

Table 12 (*continued*)

Task Type	Task Name	Author, Year	Dataset	Representation Learning	Classifier	Performance Evaluation
		Boyd et al., 2023 [217]	Boyd et al. Dataset (PDB, ArchiveII), Sato et al. Dataset (bpRNA-1m TS0)	Transformer	_	PDB: F1-score = 0.879, PPV = 0.891, Sn = 0.856; bpRNA-1m TS0: F1-score = 0.564, PPV = 0.524, Sn = 0.653; ArchiveII: F1-score = 0.636, PPV = 0.653, Sn = 0.628
		Chen el al., 2022 [218]	ArchiveII600 Dataset, bpRNA TS0 Dataset	_	CNN	ArchiveII600 Dataset: Precision = 0.936, Recall = 0.951, F1-score = 0.941; bpRNA TS0 Dataset: Precision = 0.718, Recall = 0.713, F1-score = 0.704

Table 13
RNA special characteristics analysis and single-cell analysis related 3 distinct RNA sequence analysis tasks predictive pipelines performance.

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
Goal: RNA special characteristics analysis						
Regression	mRNA Degradation Prediction	**He et al., 2023 [26]**	**NLuc Eterna PCC, eGFP, MEV**	**Transformer**	_	**NLuc Eterna: PCC = -0.655; eGFP: PCC = -0.499; MEV: PCC = -0.578**
Regression	RNA-Seq Coverage Prediction	**Linder et al., 2023 [27]**	**Human Samples Gene-level**	**Transformer**	_	**Human Samples: PCC = 0.83; Gene-level: PCC = 0.89**
Goal: RNA single-cell analysis						
Multi-class Classification	Cell Type Detection	Wan et al., 2024 [283]	Large cell type Alpha, Small cell type Delta	BERT	KNN	Large cell type Alpha: H-score = 0.838, Acc = 0.845; Small cell type Delta: H-score = 0.826, Acc = 0.837
		Yang et al., 2022 [284]	Human Cell Atlas Dataset	BERT	_	F1-score = 0.826, Acc = 0.840
		Qiu et al., 2024 [243]	**Single-Cell Multi-omics Dataset: Specter Dataset, 10X_10K Dataset, SMAGE Dataset, Spleen Dataset, BMNC Dataset**	**Non-negative Matrix Factorization**	**Clustering Algorithm**	**Specter: ACC = 0.70, AMI = 0.72, NMI = 0.62, ARI = 0.68; 10X_10K: ACC = 0.82, AMI = 0.78, NMI = 0.76, ARI = 0.78; Spleen: ACC = 0.69, AMI = 0.72, NMI = 0.71, ARI = 0.68; BMNC: ACC = 0.78, AMI = 0.79, NMI = 0.82, ARI = 0.80; SMAGE: ACC = 0.77, AMI = 0.66, NMI = 0.67, ARI = 0.68**
		Cui et al., 2024 [242]	**1. Cell Type Discovery a) Myeloid Dataset, b) Multiple Sclerosis Dataset, c) hPancreas Dataset**	**GPT**	_	**Myeloid Dataset: Acc = 0.642, Precision = 0.366, Recall = 0.347, Macro F1-score = 0.346; Multiple Sclerosis Dataset: Acc = 0.856, Precision = 0.729, Recall = 0.720, Macro F1-score = 0.703; hPancreas Dataset: Acc = 0.968, Precision = 0.735, Recall = 0.725, Macro F1-score = 0.718**

(*continued on next page*)

national Conference on High Performance Computing & Communications (ICHPC), the 2021 International Conference on Innovative Computing (ICIC), and the 14^{th} ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCBHI). Among all Transactions, more studies are published in ACM Transaction on Computational Biology. Considering the fast pace of research findings, researchers have also published 15 studies in BioRxiv, and arXIv platforms. Overall, researchers are more inclined towards journals publications due to broader dissemination, and lasting impact of their work. Furthermore, Fig. 8 illustrates distribution of 172 RNA sequence analysis studies across 21 different publishers including Oxford University Press,[8] Springer,[9] El-

[8] https://academic.oup.com/.

[9] https://www.springer.com/in.

Table 13 (*continued*)

Task Type	Task Name	Author, Year [ref]	Dataset	Representation Learning	Classifier	Performance Evaluation
Clustering	Cell Type Detection	**Cui et al., 2024 [242]**	**1. scRNA-seq cell type clustering a) COVID-19 Dataset, b) PBMC 10K Dataset, c) Perirhinal Cortex; 2. scMultiomic cell type clustering a) 10X Multiome PBMC Dataset, b) BMMC Dataset, c) ASAP PBMC Dataset**	**GPT**	_	**COVID-19 Dataset: Biological Conservation (Average BIO = 0.504, NMI = 0.659, ARI = 0.400, ASW = 0.452), Batch Correction (Average BATCH = 0.850, ASW = 0.826, GraphCon = 0.874) Overall = 0.642; PBMC 10K Dataset: Biological Conservation (Average BIO = 0.821, NMI = 0.850, ARI = 0.873, ASW = 0.740), Batch Correction (Average BATCH = 0.923, ASW = 0.950, GraphCon = 0.895) Overall = 0.862; Perirhinal Cortex Dataset: Biological Conservation (Average BIO = 0.899, NMI = 0.930, ARI = 0.919, ASW = 0.848), Batch Correction (Average BATCH = 0.930, ASW = 0.898, GraphCon = 0.964) Overall = 0.911; BMMC Dataset: Biological Conservation (Average BIO = 0.697, NMI = 0.783, ARI = 0.725, ASW = 0.582), Batch Correction (Average BATCH = 0.871, ASW = 0.834, GraphCon = 0.908) Overall = 0.766; ASAP PBMC Dataset: Biological Conservation (Average BIO = 0.587, NMI = 0.645, ARI = 0.469, ASW = 0.648), Batch Correction (Average BATCH = 0.951, ASW = 0.909, GraphCon = 0.992) Overall = 0.732; 10X Multiome PBMC Dataset: Biological Conservation (Average BIO = 0.758, NMI = 0.807, ARI = 0.822, ASW = 0.645)**

sevier,[10] IEEE,[11] MDPI,[12] ACS Publications,[13] Frontiers Media SA,[14] Frontiers,[15] Public Library of Science San Francisco CA USA,[16] Nature Publishing Group US New York,[17] Nature Publishing Group UK London,[18] Taylor & Francis,[19] Cold Spring Harbor Lab,[20] Hindawi,[21] Hindawi Limited,[22] PeerJ Inc.,[23] ASBMB,[24] Pre-print,[25] AIMS,[26] Stanford Project,[27] and ACM.[28] Notably, 113 out of 172 RNA sequence analysis articles have been published by Oxford University Press, Springer, Elsevier, and IEEE. Additionally, MDPI, Frontiers Media SA, Nature Publishing Group US New York, and Cold Spring Harbor Lab have collectively contributed 32 relevant

[10] https://www.elsevier.com/.

[11] https://www.ieee.org/.

[12] https://www.mdpi.com/.

[13] https://pubs.acs.org/.

[14] https://research.monash.edu/en/activities/frontiers-media-sa-publisher.

[15] https://www.frontiersin.org/.

[16] https://plos.org/.

[17] https://www.nature.com/.

[18] https://www.iabuk.com/member-directory/nature-publishing-group.

[19] https://taylorandfrancis.com/.

[20] https://www.cshlpress.com/.

[21] https://www.hindawi.com/.

[22] https://hindawi.editage.com/.

[23] https://peerj.com/.

[24] https://www.asbmb.org/.

[25] https://arxiv.org/.

[26] https://www.aimspress.com/.

[27] https://www.sup.org/.

[28] https://www.acm.org/publications.

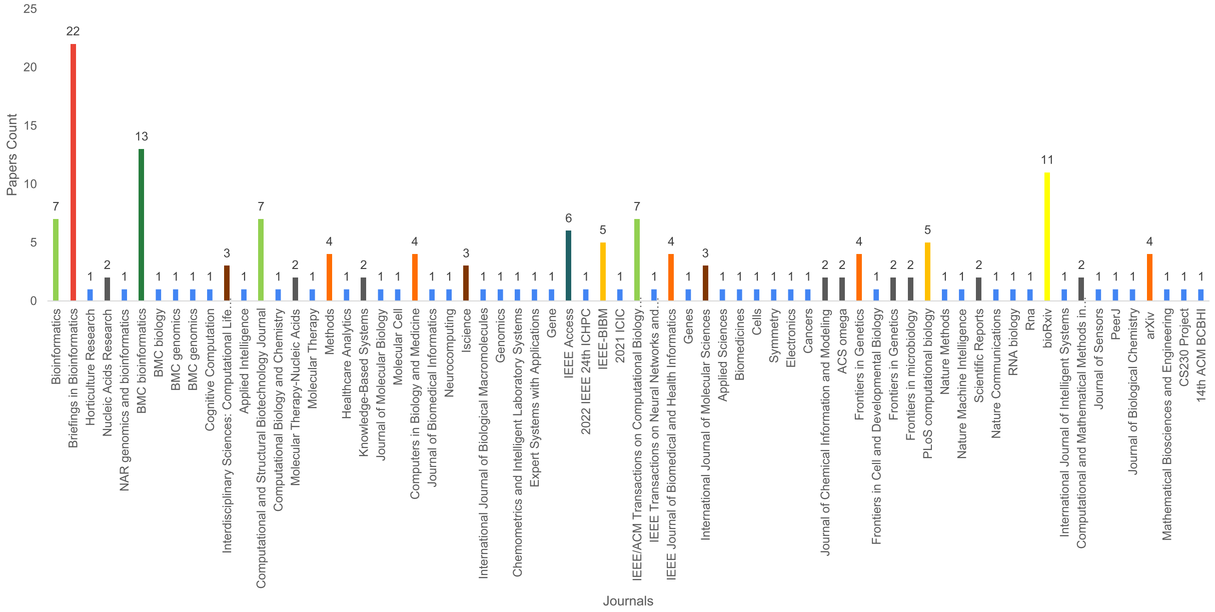

Fig. 7. Publication Distribution of RNA Sequence Analysis Literature Across Diverse Journals and Conferences.

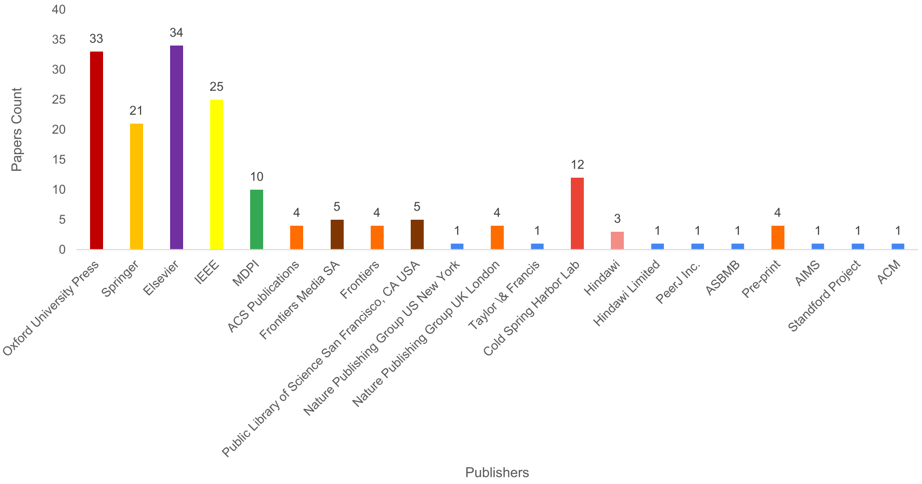

Fig. 8. Distribution of Publishers Involved in the Publication of RNA Sequence Analysis Literature.

articles. Also, 27 RNA sequence analysis articles have appeared in journals published by ACS Publications, Frontiers, Public Library of Science San Francisco, CA USA, Nature Publishing Group UK London, Taylor & Francis, Hindawi, Hindawi Limited, PeerJ Inc., ASBMB, Pre-print, AIMS, Stanford Project, and ACM. In summary, from 172 RNA sequence analysis studies, 141 are journal studies, 8 are conference studies, 8 are transaction studies, and 15 are pre-print studies, published by 21 different publishers. This comprehensive analysis across different journals, conferences, transactions, pre-print repositories and published underscores diverse and extensive research landscape in RNA sequence analysis.

12. Discussion

The paper in hand performs comprehensive analysis of existing literature having focus on AI applications across 47 distinct RNA sequence analysis tasks to provide a detailed overview of benchmark datasets, innovative representation learning methods (word embeddings and large language models), machine and deep learning predictors. A thorough analysis of the existing AI-driven RNA sequence analysis literature identifies a total of 90 potential databases that have been utilized to create benchmark datasets for 47 unique RNA sequence analysis tasks. Among these databases, only 64 are currently accessible, while 26 are either unavailable or no longer exist. Furthermore, 172 AI-driven RNA sequence analysis studies have generated 310 unique datasets to support development of AI predictors for 47 diverse RNA sequence analysis tasks. Among these 310 datasets, 236 are publicly available, while 74 remain proprietary or in-house.

Despite the availability of numerous public datasets, a notable inconsistency remains in the evaluation of predictors across the same datasets for each RNA sequence analysis task. In the process of new predictors development, most of the researchers are evaluating their predictors solely on their newly developed datasets and are overlooking the vast array of existing datasets available in the field. Development of new benchmark datasets is a valuable effort since sequences in public databases are updated daily, weekly, or monthly. The new datasets contains up-to-date and newly discovered sequences. In addition, most of existing datasets are relatively small and deep learning models demonstrate better performance with larger datasets. To address this, there is an urgent need for standardized dataset utilization. For a more objective and transparent performance comparison, two approaches can be used: 1) Evaluation of new predictors on both existing and new datasets, 2) Benchmark existing predictors performance on new datasets. Unfortunately, limited availability of open-source code of existing predictors intensifies this issue and hinders direct performance comparison of predictive pipelines. To ensure methodological advancement, it is important to develop task-specific standardized datasets and foster open-source practices for predictive pipeline implementations.

Besides datasets standardization, a robust and precise predictor development relies heavily on sophisticated representation learning methods along with appropriate machine or deep learning algorithms. The role of representation learning methods is key in AI-driven RNA sequence analysis predictive pipelines, as raw RNA sequences cannot be directly processed by machine and deep learning algorithms. In the realm of AI-driven RNA sequence analysis, researchers have explored potential of various advance representation learning methods including 16 word embedding methods and 8 large language models. In addition to these methods, potential of other 15 word embedding methods and 12 large language models has been explored in DNA and protein sequence analysis fields. However, the potential of these word embedding methods and language models has not been explored yet in RNA sequence analysis predictive pipelines. The unexplored word embedding methods include DANE [322], FastText [323,324], GEMSEC [325], MetaGraph2Vec [144], HAKE [326], Laplacian eigen maps [327], Locally linear embedding [327], Mashup [328,329], OPA2Vec [330], Random Watcher-Walker (RW2) [331], RWR [147], SVD [157,259], Topo2Vec [332], TransE [333], and Graph2vec [334]. Moreover, unexplored language models are ALBERT [335–337], AlphaFold [220,338–341], AlphaFold2 [342,343], ELECTRA [335,344], ESM-2 [342,345,172,346], Graph Transformer Network [347], IgFold [348], RoBERTa [337,349], T5 [346,350–352], Transformer-XL [353], ULMFiT [354,355], Vision Transformer [356], and XLNet [335]. The utilization of additional word embeddings methods and large language models for DNA and protein sequences can offer new insights and improved accuracy for AI-driven RNA sequence analysis tasks.

In the current landscape of AI-driven RNA sequence analysis predictive pipelines, an analysis at the predictor level algorithms indicates that researchers have investigated the potential of 13 machine learning and 9 deep learning algorithms. Overall in 58 different word embedding based predictive pipelines, 13 predictive pipelines have utilized standalone machine learning algorithms, and 33 have employed standalone deep learning algorithms. In addition, 2 predictive pipelines are designed by using machine learning based algorithms meta predictors, 5 are developed using deep learning based meta predictors and 4 predictive pipelines have employed both machine and deep learning based meta predictors are utilized. On the other hand, within 70 large language models based predictive pipelines, 53 predictive pipelines have utilized self classifier, 12 are developed by using standalone machine learning based algorithm, and 5 are designed by employing deep learning based algorithms. Moreover, 2 language models based predictive pipelines have utilized deep learning based meta predictors and 1 has leveraged machine learning based meta predictor.

CRediT authorship contribution statement

Muhammad Nabeel Asim: Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Conceptualization. **Muhammad Ali Ibrahim:** Writing – original draft, Formal analysis, Data curation. **Tayyaba Asif:** Writing – original draft, Formal analysis, Data curation. **Andreas Dengel:** Supervision, Formal analysis.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Grammarly and Paperpal tool in order to fix language and grammar issues and ChatGpt for outlining, better understanding different studies, and expansion of concepts. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

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

Data and code availability

No new data was generated for the research described in the article.

References

[1] Jay Shendure, Hanlee Ji, Next-generation dna sequencing, Nat. Biotechnol. 26 (10) (2008) 1135–1145.

[2] Kimberly R. Kukurba, Stephen B. Montgomery, Rna sequencing and analysis, Cold Spring Harb. Protoc. 2015 (11) (2015) pdb-top084970.

[3] Diksha Pandey, P. Onkara Perumal, A scoping review on deep learning for next-generation rna-seq. data analysis, Funct. Integr. Genomics 23 (2) (2023) 134.

[4] Muhammad Nabeel Asim, An Efficient Automated Machine Learning Framework for Genomics and Proteomics Sequence Analysis, PhD thesis, Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, 2023.

[5] Zhen Chen, Pei Zhao, Chen Li, Fuyi Li, Dongxu Xiang, Yong-Zi Chen, Tatsuya Akutsu, Roger J. Daly, Geoffrey I. Webb, Quanzhi Zhao, et al., ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization, Nucleic Acids Res. 49 (10) (2021) e60.

[6] Hong-Liang Li, Yi-He Pang, Bin Liu, Bioseq-blm: a platform for analyzing dna, rna and protein sequences based on biological language models, Nucleic Acids Res. 49 (22) (2021) e129.

[7] Xiaoyong Pan, Hong-Bin Shen, Learning distributed representations of rna sequences and its application for predicting rna-protein binding sites with a convolutional neural network, Neurocomputing 305 (2018) 51–58.

[8] Manato Akiyama, Yasubumi Sakakibara, Informative rna base embedding for rna structural alignment and clustering by deep representation learning, NAR Genomics Bioinform. 4 (1) (2022) lqac012.

[9] Keisuke Yamada, Michiaki Hamada, Prediction of rna–protein interactions using a nucleotide language model, Bioinform. Adv. 2 (1) (2022) vbac023.

[10] Ying Zhang, Fang Ge, Fuyi Li, Xibei Yang, Jiangning Song, Dong-Jun Yu, Prediction of multiple types of rna modifications via biological language model, IEEE/ACM Trans. Comput. Biol. Bioinform. (2023).

[11] Weijie Yin, Zhaoyu Zhang, Liang He, Rui Jiang, Shuo Zhang, Gan Liu, Xuegong Zhang, Tao Qin, Zhen Xie, Ernie-rna: an rna language model with structure-enhanced representations, bioRxiv (2024), pages 2024–03.

[12] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel, Sheraz Ahmed, Advances in computational methodologies for classification and sub-cellular locality prediction of non-coding rnas, Int. J. Mol. Sci. 22 (16) (2021) 8719.

[13] Mengting Niu, Chunyu Wang, Yaojia Chen, Quan Zou, Ren Qi, Lei Xu, Circrna identification and feature interpretability analysis, BMC Biol. 22 (1) (2024) 44.

[14] Marco Stricker, Muhammad Nabeel Asim, Andreas Dengel, Sheraz Ahmed, Circnet: an encoder–decoder-based convolution neural network (cnn) for circular rna identification, Neural Comput. Appl. (2022) 1–12.

[15] Sagar Gupta, Ravi Shankar, miwords: transformer-based composite deep learning for highly accurate discovery of pre-mirna regions across plant genomes, Brief. Bioinform. 24 (2) (2023) bbad088.

[16] Zhao-Yue Zhang, Zheng Zhang, Xiucai Ye, Tetsuya Sakurai, Hao Lin, A bert-based model for the prediction of lncrna subcellular localization in homo sapiens, Int. J. Biol. Macromol. 265 (2024) 130659.

[17] Haibin Liu, Dianguo Li, Hao Wu, Lnclocator-imb: an imbalance-tolerant ensemble deep learning framework for predicting long non-coding rna subcellular localization, IEEE J. Biomed. Health Inform. (2023).

[18] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Christoph Zehe, Olivier Cloarec, Johan Trygg, Andreas Dengel, Sheraz Ahmed, El-rmlocnet: an explainable lstm network for rna-associated multi-compartment localization prediction, Comput. Struct. Biotechnol. J. 20 (2022) 3986–4002.

[19] Min Zeng, Yifan Wu, Yiming Li, Rui Yin, Chengqian Lu, Junwen Duan, Min Li, Lnclocformer: a transformer-based deep learning model for multi-label lncrna subcellular localization prediction by using localization-specific attention mechanism, Bioinformatics 39 (12) (2023) btad752.

[20] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel, Sheraz Ahmed, Circ-locnet: a computational framework for circular rna sub-cellular localization prediction, Int. J. Mol. Sci. 23 (15) (2022) 8221.

[21] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Christoph Zehe, Olivier Cloarec, Rickard Sjogren, Johan Trygg, Andreas Dengel, Sheraz Ahmed, L2s-mirloc: a lightweight two stage mirna sub-cellular localization prediction framework, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.

[22] Zheng-Yang Zhao, Jie Lin, Zhen Wang, Jian-Xin Guo, Xin-Ke Zhan, Yu-An Huang, Chuan Shi, Wen-Zhun Huang, Sebglma: semantic embedded bipartite graph network for predicting lncrna-mirna associations, Int. J. Intell. Syst. 2023 (1) (2023) 2785436.

[23] Yong Han, Shao-Wu Zhang, ncrpi-lgat: prediction of ncrna-protein interactions with line graph attention network framework, Comput. Struct. Biotechnol. J. 21 (2023) 2286–2295.

[24] F. Ding, S. Sharma, P. Chalasani, V. Demidov, N. Broude, N. Dokholyan, Ab initio rna folding by discrete molecular dynamics: from structure prediction to folding mechanisms, RNA 14 (2008) 1164–1173.

[25] Iain M. Dykes, Costanza Emanueli, Transcriptional and post-transcriptional gene regulation by long non-coding rna, Genomics Proteomics Bioinform. 15 (3) (2017) 177–186.

[26] Shujun He, Baizhen Gao, Rushant Sabnis, Qing Sun, Rnadegformer: accurate prediction of mrna degradation at nucleotide resolution with deep learning, Brief. Bioinform. 24 (1) (2023) bbac581.

[27] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, David R. Kelley, Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation, bioRxiv (2023), pages 2023–08.

[28] Jialin Zhang, Haoran Zhu, Yin Liu, Xiangtao Li, mitds: uncovering mirna-mrna interactions with deep learning for functional target prediction, Methods 223 (2024) 65–74.

[29] Jiayu Xu, Nan Xu, Weixin Xie, Chengkui Zhao, Lei Yu, Weixing Feng, Bert-sirna: sirna target prediction based on bert pre-trained interpretable model, Gene (2024) 148330.

[30] Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, Yuedong Yang, Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction, bioRxiv (2023), pages 2023–01.

[31] Mhaned Oubounyt, Zakaria Louadi, Hilal Tayara, Kil ToChong, Deep learning models based on distributed feature representations for alternative splicing prediction, IEEE Access 6 (2018) 58826–58834.

[32] A. Sonawane, J. Platig, M. Fagny, C. Chen, J. Paulson, C. Lopes-Ramos, D. DeMeo, J. Quackenbush, K. Glass, M. Kuijjer, Understanding Tissue-Specific Gene Regulation, 2017.

[33] Y. Qiu, Scmnmf: a novel method for single-cell multi-omics clustering based on matrix factorization, Brief. Bioinform. 25 (2024).

[34] D. Cui, W. Li, J. Wu, J. Xie, Y. Wu, Advances in multi-omics applications in hbv-associated hepatocellular carcinoma, Front. Med. 8 (2021).

[35] Megha Patel, Nimish Magre, Himanshi Motwani, Nik Bear Brown, Advances in machine learning, statistical methods, and ai for single-cell rna annotation using raw count matrices in scrna-seq data, arXiv preprint arXiv:2406.05258, 2024.

[36] Enrique Pola-Sánchez, Karen Magdalena Hernández-Martínez, Rafael Pérez-Estrada, Nelly Sélem-Mójica, June Simpson, María Jazmín Abraham-Juárez, Alfredo Herrera-Estrella, José Manuel Villalobos-Escobedo, Rna-seq data analysis: a practical guide for model and non-model organisms, Curr. Protoc. 4 (5) (2024) e1054.

[37] Kengo Sato, Michiaki Hamada, Recent trends in rna informatics: a review of machine learning and deep learning for rna secondary structure prediction and rna drug discovery, Brief. Bioinform. 24 (4) (2023) bbad186.

[38] K. Chandrashekar, Vidya Niranjan, Adarsh Vishal, Anagha S. Setlur, Integration of artificial intelligence, machine learning and deep learning techniques in genomics: review on computational perspectives for ngs analysis of dna and rna seq data, Curr. Bioinform. 19 (9) (2024) 825–844.

[39] Hyeonseo Hwang, Hyeonseong Jeon, Nagyeong Yeo, Daehyun Baek, Big data and deep learning for rna biology, Exp. Mol. Med. (2024) 1–29.

[40] Kathi Zarnack, Eduardo Eyras, Artificial Intelligence and Machine Learning in Rna Biology, 2023.

[41] Zhanmin Liang, Haokai Ye, Jiongming Ma, Zhen Wei, Yue Wang, Yuxin Zhang, Daiyun Huang, Bowen Song, Jia Meng, Daniel J. Rigden, et al., m6a-atlas v2. 0: updated resources for unraveling the n 6-methyladenosine (m6a) epitranscriptome among multiple species, Nucleic Acids Res. 52 (D1) (2024) D194–D202.

[42] Jia Chen, Jiahao Lin, Yongfei Hu, Meijun Ye, Linhui Yao, Le Wu, Wenhai Zhang, Meiyi Wang, Tingting Deng, Feng Guo, et al., Rnadisease v4. 0: an updated resource of rna-associated diseases, providing rna-disease analysis, enrichment and prediction, Nucleic Acids Res. 51 (D1) (2023) D1397–D1404.

[43] MingLiu, Qian Wang, Jian Shen, Burton B. Yang, Xiangming Ding, Circbank: a comprehensive database for circrna with standard nomenclature, RNA Biol. 16 (7) (2019) 899–905.

[44] Jia-Jia Xuan, Wen-Ju Sun, Peng-Hui Lin, Ke-Ren Zhou, Shun Liu, Ling-Ling Zheng, Liang-Hu Qu, Jian-Hua Yang, Rmbase v2. 0: deciphering the map of rna modifications from epitranscriptome sequencing data, Nucleic Acids Res. 46 (D1) (2018) D327–D334.

[45] Bharat Panwar, Gilbert S. Omenn, Yuanfang Guan, mirmine: a database of human mirna expression profiles, Bioinformatics 33 (10) (2017) 1554–1560.

[46] ZhenYang, Fei Ren, Changning Liu, Shunmin He, Gang Sun, Qian Gao, Lei Yao, Yangde Zhang, Ruoyu Miao, Ying Cao, et al., dbdemc: a Database of Differentially Expressed Mirnas in Human Cancers, BMC Genomics, vol. 11, Springer, 2010, pp. 1–8.

[47] Boya Xie, Qin Ding, Hongjin Han, Di Wu, mircancer: a microrna–cancer association database constructed by text mining on literature, Bioinformatics 29 (5) (2013) 638–644.

[48] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, Jian-Hua Yang, starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data, Nucleic Acids Res. 42 (D1) (2014) D92–D97.

[49] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, Yunlong Liu, mir2disease: a manually curated database for microrna deregulation in human disease, Nucleic Acids Res. 37 (suppl_1) (2009) D98–D104.

[50] Yang Li, Chengxiang Qiu, Jian Tu, Bin Geng, Jichun Yang, Tianzi Jiang, Qinghua Cui, Hmdd v2. 0: a database for experimentally supported human microrna and disease associations, Nucleic Acids Res. 42 (D1) (2014) D1070–D1074.

[51] Praveen Sethupathy, Benoit Corda, Artemis G. Hatzigeorgiou, Tarbase: a comprehensive database of experimentally supported animal microrna targets, RNA 12 (2) (2006) 192–197.

[52] Xueyi Teng, Xiaomin Chen, Hua Xue, Yiheng Tang, Peng Zhang, Quan Kang, Yajing Hao, Runsheng Chen, Yi Zhao, Shunmin He, Npinter v4. 0: an integrated database of ncrna interactions, Nucleic Acids Res. 48 (D1) (2020) D160–D165.

[53] Sam Griffiths-Jones, mirbase: the microrna sequence database, MicroRNA Protoc. (2006) 129–138.

[54] Carrie A. Davis, Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, Kriti Jain, Ulugbek K. Baymuradov, Aditi K. Narayanan, et al., The encyclopedia of dna elements (encode): data portal update, Nucleic Acids Res. 46 (D1) (2018) D794–D801.

[55] Shuhei Noguchi, Takahiro Arakawa, Shiro Fukuda, Masaaki Furuno, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, Kaoru Kaida, Ai Kaiho, Mutsumi Kanamori-Katayama, et al., Fantom5 cage profiles of human and mouse samples, Sci. Data 4 (1) (2017) 1–10.

[56] Michał Wojciech Szcześniak, Oleksii Bryzghalov, Joanna Ciomborowska-Basheer, Izabela Makałowska, Cantatadb 2.0: expanding the collection of plant long noncoding rnas, in: Plant Long Non-Coding RNAs: Methods and Protocols, 2019, pp. 415–429.

[57] Geng Chen, Ziyun Wang, Dongqing Wang, Chengxiang Qiu, Mingxi Liu, Xing Chen, Qipeng Zhang, Guiying Yan, Qinghua Cui, Lncrnadisease: a database for long-non-coding rna-associated diseases, Nucleic Acids Res. 41 (D1) (2012) D983–D986.

[58] ShuangSang Fang, LiLi Zhang, JinCheng Guo, YiWei Niu, Yang Wu, Hui Li, LianHe Zhao, XiYuan Li, XueYi Teng, XianHui Sun, et al., Noncodev5: a comprehensive annotation database for long non-coding rnas, Nucleic Acids Res. 46 (D1) (2018) D308–D314.

[59] Pieter-Jan Volders, Kenny Helsens, Xiaowei Wang, Björn Menten, Lennart Martens, Kris Gevaert, Jo Vandesompele, Pieter Mestdagh, Lncipedia: a database for annotated human lncrna transcript sequences and structures, Nucleic Acids Res. 41 (D1) (2013) D246–D251.

[60] Mercy Rophina, Disha Sharma, Mukta Poojary, Vinod Scaria, Circad: a comprehensive manually curated resource of circular rna associated with diseases, Database 2020 (2020) baaa019.

[61] Siyu Xia, Jing Feng, Ke Chen, Yanbing Ma, Jing Gong, Fangfang Cai, Yuxuan Jin, Yang Gao, Linjian Xia, Hong Chang, et al., Cscd: a database for cancer-specific circular rnas, Nucleic Acids Res. 46 (D1) (2018) D925–D929.

[62] Zhi-Yan Sun, Chang-Lin Yang, Li-Jie Huang, Zong-Chao Mo, Ke-Nan Zhang, Wen-Hua Fan, Kuan-Yu Wang, Fan Wu, Ji-Guang Wang, Fan-Lin Meng, et al., < ? mode longmeta? > circrnadisease v2. 0: an updated resource for high-quality experimentally supported circrna-disease associations, Nucleic Acids Res. 52 (D1) (2024) D1193–D1200.

[63] Xiaoping Chen, Ping Han, Tao Zhou, Xuejiang Guo, Xiaofeng Song, Yan Li, circrnadb: a comprehensive database for human circular rnas with protein-coding annotations, Sci. Rep. 6 (1) (2016) 34985.

[64] Glaž, Circbase: a Database for Circular Rnas, 2014.

[65] Allan Peter Davis, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C. Wiegers, Carolyn J. Mattingly, Comparative toxicogenomics database (ctd): update 2021, Nucleic Acids Res. 49 (D1) (2021) D1138–D1143.

[66] Patricia P. Chan, Todd M. Lowe, Gtrnadb: a database of transfer rna genes detected in genomic sequence, Nucleic Acids Res. 37 (suppl_1) (2009) D93–D97.

[67] Jiajia Wang, Peng Zhang, Yiping Lu, Yanyan Li, Yu Zheng, Yunchao Kan, Runsheng Chen, Shunmin He, pirbase: a comprehensive database of pirna sequences, Nucleic Acids Res. 47 (D1) (2019) D175–D180.

[68] Xiaotong Luo, Yuantai Huang, Huiqin Li, Yihai Luo, Zhixiang Zuo, Jian Ren, Yubin Xie, Spencer: a comprehensive database for small peptides encoded by noncoding rnas in cancer patients, Nucleic Acids Res. 50 (D1) (2022) D1373–D1381.

[69] Tianyu Cui, Yiying Dou, Puwen Tan, Zhen Ni, Tianyuan Liu, DuoLin Wang, Yan Huang, Kaican Cai, Xiaoyang Zhao, Dong Xu, et al., Rnalocate v2. 0: an updated resource for rna subcellular localization with increased coverage and annotation, Nucleic Acids Res. 50 (D1) (2022) D333–D339.

[70] Yue Gao, Shipeng Shang, Shuang Guo, Xin Li, Hanxiao Zhou, Hongjia Liu, Yue Sun, Junwei Wang, Peng Wang, Hui Zhi, et al., Lnc2cancer 3.0: an updated resource for experimentally supported lncrna/circrna cancer associations and web tools based on rna-seq and scrna-seq data, Nucleic Acids Res. 49 (D1) (2021) D1251–D1258.

[71] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al., Gencode reference annotation for the human and mouse genomes, Nucleic Acids Res. 47 (D1) (2019) D766–D773.

[72] WeiLan, Mingrui Zhu, Qingfeng Chen, Baoshan Chen, Jin Liu, Min Li, Yi-Ping Phoebe Chen, Circr2cancer: a manually curated database of associations between circrnas and cancers, Database (2020) 2020:baaa085.

[73] Bailing Zhou, Baohua Ji, Kui Liu, Guodong Hu, Fei Wang, Qingshuai Chen, Ru Yu, Pingping Huang, Jing Ren, Chengang Guo, et al., Evlncrnas 2.0: an updated database of manually curated functional long non-coding rnas validated by low-throughput experiments, Nucleic Acids Res. 49 (D1) (2021) D86–D91.

[74] Oscar Franzén, Li-Ming Gan, Johan L.M. Björkegren, Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data, Database (2019) 2019:baz046.

[75] Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E. Loveland, Jonathan M. Mudge, Cristina Sisu, James C. Wright, Joel Armstrong, If Barnes et al. Gencode 2021, Nucleic Acids Res. 49 (D1) (2021) D916–D923.

[76] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al., Gencode: the reference human genome annotation for the encode project, Genome Res. 22 (9) (2012) 1760–1774.

[77] Dawood B. Dudekula, Amaresh C. Panda, Ioannis Grammatikakis, Supriyo De, Kotb Abdelmohsen, Myriam Gorospe, Circinteractome: a web tool for exploring circular rnas and their interacting proteins and micrornas, RNA Biol. 13 (1) (2016) 34–42.

[78] Girolamo Giudice, Fátima Sánchez-Cabo, Carlos Torroja, Enrique Lara-Pezzi, Attract—a database of rna-binding proteins and associated motifs, Database (2016) 2016:baw035.

[79] Wei Ma, Lu Zhang, Pan Zeng, Chuanbo Huang, Jianwei Li, Bin Geng, Jichun Yang, Wei Kong, Xuezhong Zhou, Qinghua Cui, An analysis of human microbe–disease associations, Brief. Bioinform. 18 (1) (2017) 85–97.

[80] Buvaneswari Coimbatore Narayanan, John Westbrook, Saheli Ghosh, Anton I. Petrov, Blake Sweeney, Craig L. Zirbel, Neocles B. Leontis, Helen M. Berman, The nucleic acid database: new features and capabilities, Nucleic Acids Res. 42 (D1) (2014) D114–D122.

[81] S. Searle, A. Frankish, A. Bignell, B. Aken, T. Derrien, M. Diekhans, R. Harte, C. Howald, F. Kokocinski, M. Lin, et al., The gencode human gene set, Genome Biol. 11 (2010) 1.

[82] Rnacentral: a Comprehensive Database of Non-coding Rna Sequences, Nucleic Acids Res. 45 (D1) (2017) D128–D134.

[83] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, Jame sG.R. Gilbert, Roy Storey, David Swarbreck, et al., Gencode: producing a reference annotation for encode, Genome Biol. 7 (2006) 1–9.

[84] Scott Federhen, The ncbi taxonomy database, Nucleic Acids Res. 40 (D1) (2012) D136–D143.

[85] ENCODE Project Consortium, A user's guide to the encyclopedia of dna elements (encode), PLoS Biol. 9 (4) (2011) e1001046.

[86] Andrew D. Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Andrey G. Azov, Ruth Bennett, et al., Ensembl 2020, Nucleic Acids Res. 48 (D1) (2020) D682–D688.

[87] Ada Hamosh, Alan F. Scott, Joanna Amberger, David Valle, Victor A. McKusick, Online Mendelian inheritance in man (omim), Human Mutat. 15 (1) (2000) 57–61.

[88] Jian-You Liao, Bing Yang, Yu-Chan Zhang, Xiao-Juan Wang, Yushan Ye, Jing-Wen Peng, Zhi-Zhi Yang, Jie-Hua He, Yin Zhang, KaiShun Hu, et al., Eurbpdb: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic rna binding proteins (rbps), Nucleic Acids Res. 48 (D1) (2020) D307–D313.

[89] Qinghua Jiang, Jixuan Wang, Xiaoliang Wu, Rui Ma, Tianjiao Zhang, Shuilin Jin, Zhijie Han, Renjie Tan, Jiajie Peng, Guiyou Liu, et al., Lncrna2target: a database for differentially expressed genes after lncrna knockdown or overexpression, Nucleic Acids Res. 43 (D1) (2015) 193–196.

[90] René Dreos, Giovanna Ambrosini, Rouayda Cavin Périer, Philipp Bucher, Epd and epdnew, high-quality promoter resources in the next-generation sequencing era, Nucleic Acids Res. 41 (D1) (2013) D157–D164.

[91] Peter D. Stenson, Edward V. Ball, Matthew Mort, Andrew D. Phillips, Jacqueline A. Shiel, Nick S.T. Thomas, Shaun Abeysinghe, Michael Krawczak, David N. Cooper, Human gene mutation database (hgmd®): 2003 update, Human Mutat. 21 (6) (2003) 577–581.

[92] Kim Pruitt, Garth Brown, Tatiana Tatusova, Donna Maglott, The reference sequence (refseq) database, NCBI Handb. 2 (2012).

[93] Ya-Ru Miao, Wei Liu, Qiong Zhang, An-Yuan Guo, lncrnasnp2: an updated database of functional snps and mutations in human and mouse lncrnas, Nucleic Acids Res. 46 (D1) (2018) D276–D280.

[94] Padideh Danaee, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, David Hendrix, bprna: large-scale automated annotation and analysis of rna secondary structure, Nucleic Acids Res. 46 (11) (2018) 5381–5394.

[95] Wen-Ju Sun, Jun-Hao Li, Shun Liu, Jie Wu, Hui Zhou, Liang-Hu Qu, Jian-Hua Yang, Rmbase: a resource for decoding the landscape of rna modifications from high-throughput sequencing data, Nucleic Acids Res. 44 (D1) (2016) D259–D265.

[96] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, Laura I. Furlong, Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res. (2016) gkw943.

[97] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al., Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res. 44 (D1) (2016) D733–D745.

[98] Jingjing Jin, Peng Lu, Yalong Xu, Zefeng Li, Shizhou Yu, Jun Liu, Huan Wang, Nam-Hai Chua, Peijian Cao, Plncdb v2. 0: a comprehensive encyclopedia of plant long noncoding rnas, Nucleic Acids Res. 49 (D1) (2021) D1489–D1495.

[99] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al., Clinvar: public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 44 (D1) (2016) D862–D868.

[100] Matthew D. Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, et al., The ncbi dbgap database of genotypes and phenotypes, Nat. Genet. 39 (10) (2007) 1181–1186.

[101] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, Sean R. Eddy, Rfam: an rna family database, Nucleic Acids Res. 31 (1) (2003) 439–441.

[102] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, et al., Ncbi geo: archive for functional genomics data sets—update, Nucleic Acids Res. 41 (D1) (2012) D991–D995.

[103] Minoru Kanehisa, The kegg database, in: 'In Silico' Simulation of Biological Processes: Novartis Foundation Symposium 247, vol. 247, Wiley Online Library, 2002, pp. 91–103.

[104] Weizhong Li, Andrew Cowley, Mahmut Uludag, Tamer Gur, Hamish McWilliam, Silvano Squizzato, Young Mi Park, Nicola Buso, Rodrigo Lopez, The embl-ebi bioinformatics web and programmatic tools framework, Nucleic Acids Res. 43 (W1) (2015) W580–W584.

[105] Gerd Anders, Sebastian D. Mackowiak, Marvin Jens, Jonas Maaskola, Andreas Kuntzagk, Nikolaus Rajewsky, Markus Landthaler, Christoph Dieterich, dorina: a database of rna interactions in post-transcriptional regulation, Nucleic Acids Res. 40 (D1) (2012) D180–D186.

[106] Daniel J. Nasko, Sergey Koren, Adam M. Phillippy, Todd J. Treangen, Refseq database growth influences the accuracy of k-mer-based lowest common ancestor species identification, Genome Biol. 19 (2018) 1–10.

[107] Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Lorenzo Kogler-Anele, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, et al., Codonbert: large language models for mrna design and optimization, bioRxiv (2023), pages 2023–09.

[108] Genta Aoki, Yasubumi Sakakibara, Convolutional neural networks for classification of alignments of non-coding rna sequences, Bioinformatics 34 (13) (2018) i237–i244.

[109] Lei Deng, Ying Jiang, Xiaowen Hu, Rongtao Zheng, Zhijian Huang, Jingpu Zhang, Ablncpp: attention mechanism-based bidirectional long short-term memory for noncoding rna coding potential prediction, J. Chem. Inf. Model. 63 (12) (2023) 3955–3966.

[110] Muhammad Nabeel Asim, Muhammad Imran Malik, Christoph Zehe, Johan Trygg, Andreas Dengel, Sheraz Ahmed, A robust and precise convnet for small non-coding rna classification (rpc-snrc), IEEE Access 9 (2020) 19379–19390.

[111] Mohamed Chaabane, Robert M. Williams, Austin T. Stephens, Juw Won Park, circdeep: deep learning approach for circular rna classification from other long non-coding rna, Bioinformatics 36 (1) (2020) 73–80.

[112] Lei Deng, Wei Lin, Jiacheng Wang, Jingpu Zhang, Deepcirgo: functional prediction of circular rnas through hierarchical deep neural networks using heterogeneous network features, BMC Bioinform. 21 (2020) 1–18.

[113] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel, Sheraz Ahmed, Adh-ppi: an attention-based deep hybrid model for protein-protein interaction prediction, iScience 25 (10) (2022).

[114] Monica F. Danilevicz, Mitchell Gill, Cassandria G. Tay Fernandez, Jakob Petereit, Shriprabha R. Upadhyaya, Jacqueline Batley, Mohammed Bennamoun, David Edwards, Philipp E. Bayer, Dnabert-based explainable lncrna identification in plant genome assemblies, Comput. Struct. Biotechnol. J. 21 (2023) 5676–5685.

[115] Zongrui Dai, Feiyang Deng, Lncpndeep: a long non-coding rna classifier based on large language model with peptide and nucleotide embedding, bioRxiv (2023), pages 2023–11.

[116] Xiu-Qin Liu, Bing-Xiu Li, Guan-Rong Zeng, Qiao-Yue Liu, Dong-Mei Ai, Prediction of long non-coding rnas based on deep learning, Genes 10 (4) (2019) 273.

[117] Xue-Chan Tian, Zhao-Yang Chen, Shuai Nie, Tian-Le Shi, Xue-Mei Yan, Yu-Tao Bao, Zhi-Chao Li, Hai-Yao Ma, Kai-Hua Jia, Wei Zhao, et al., Plant-lncpipe: a computational pipeline providing significant improvement in plant lncrna identification, in: Horticulture Research, 2024, uhae041.

[118] Saleh Musleh, Mohammad Tariqul Islam, Tanvir Alam, Lncri: long non-coding rna identifier in multiple species, IEEE Access 9 (2021) 167219–167228.

[119] Rana M. Nadir, Hafsa Mateen, Saif U. Din, Classification of incrna and mrna using k-mers and random forest, in: 2021 International Conference on Innovative Computing (ICIC), IEEE, 2021, pp. 1–8.

[120] Sagar Gupta, Ravi Shankar, miwords: transformer-based composite deep learning for highly accurate discovery of pre-mirna regions across plant genomes, Brief. Bioinform. 24 (2) (2023) bbad088.

[121] Jonathan Raad, Leandro A. Bugnon, Diego H. Milone, Georgina Stegmayer, mire2e: a full end-to-end deep model based on transformers for prediction of pre-mirnas, Bioinformatics 38 (5) (2022) 1191–1197.

[122] Wentao Zhu, Huanzeng Xie, Yaowen Chen, Guishan Zhang, Crnncrispr: an interpretable deep learning method for crispr/cas9 sgrna on-target activity prediction, Int. J. Mol. Sci. 25 (8) (2024) 4429.

[123] Tianjiao Zhang, Liangyu Li, Hailong Sun, Guohua Wang, Deepiteh: a deep learning framework for identifying tissue-specific ernas from the human genome, Bioinformatics 39 (6) (2023) btad375.

[124] Dung Hoang, Anh Mai, Linh Thanh Nguyen, Eun Yeol Lee, Tssnote-cyaprombert: development of an integrated platform for highly accurate promoter prediction and visualization of synechococcus sp. and synechocystis sp. through a state-of-the-art natural language processing model bert, Front. Genet. 13 (2022) 1067562.

[125] Xin Wang, Xin Gao, Guohua Wang, Dan Li, miprobert: identification of microrna promoters based on the pre-trained model bert, Brief. Bioinform. 24 (3) (2023) bbad093.

[126] Weidun Xie, Jiawei Luo, Chu Pan, Ying Liu, Sg-lstm-frame: a computational frame using sequence and geometrical information via lstm to predict mirna–gene associations, Brief. Bioinform. 22 (2) (2021) 2032–2042.

[127] Seungwon Yoon, Inwoo Hwang, Jaeeun Cho, Hyewon Yoon, Kyuchul Lee, migap: mirna–gene association prediction method based on deep learning model, Appl. Sci. 13 (22) (2023) 12349.

[128] Haitao Zou, Boya Ji, Meng Zhang, Fen Liu, Xiaolan Xie, Shaoliang Peng, Mhgtmda: molecular heterogeneous graph transformer based on biological entity graph for mirna-disease associations prediction, in: Molecular Therapy-Nucleic Acids, 2024.

[129] Dong Ouyang, Yong Liang, Jinfeng Wang, Le Li, Ning Ai, Junning Feng, Shanghui Lu, Shuilin Liao, Xiaoying Liu, Shengli Xie, Hgclamir: hypergraph contrastive learning with attention mechanism and integrated multi-view representation for predicting mirna-disease associations, PLoS Comput. Biol. 20 (4) (2024) e1011927.

[130] Yinbo Liu, Xiaodi Yan, Jun Li, Xinxin Ren, Qi Wu, Gang-Ao Wang, Yuqing Chen, Xiaolei Zhu, mirna-disease association prediction based on heterogeneous graph transformer with multi-view similarity and random auto-encoder, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 885–888.

[131] Zhiwei Ning, Jinyang Wu, Yidong Ding, Ying Wang, Qinke Peng, Laiyi Fu, Bertnda: a model based on graph-bert and multi-scale information fusion for ncrna-disease association prediction, IEEE J. Biomed. Health Inform. (2023).

[132] Minghao Yang, Zhi-An Huang, Wenhao Gu, Kun Han, Wenying Pan, Xiao Yang, Zexuan Zhu, Prediction of biomarker–disease associations based on graph attention network and text representation, Brief. Bioinform. 23 (5) (2022) bbac298.

[133] Jinyang Wu, Zhiwei Ning, Yidong Ding, Ying Wang, Qinke Peng, Laiyi Fu, Kgetcda: an efficient representation learning framework based on knowledge graph encoder from transformer for predicting circrna-disease associations, Brief. Bioinform. 24 (5) (2023) bbad292.

[134] Yi Zhou, Xinyi Wang, Lin Yao, Min Zhu, Ldaformer: predicting lncrna-disease associations based on topological feature extraction and transformer encoder, Brief. Bioinform. 23 (6) (2022) bbac370.

[135] Dengju Yao, Bailin Li, Xiaojuan Zhan, Xiaorong Zhan, Liyang Yu, Gcnformer: graph convolutional network and transformer for predicting lncrna-disease associations, BMC Bioinform. 25 (1) (2024) 5.

[136] Guanghui Li, Peihao Bai, Cheng Liang, Jiawei Luo, Node-adaptive graph transformer with structural encoding for accurate and robust lncrna-disease association prediction, BMC Genomics 25 (1) (2024) 73.

[137] Xiaoyu Wu, Jinli Zhang, Zongli Jiang, Jianqiang Li, Contrastive self-supervised learning for predicting disease-rna associations, in: 2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), IEEE, 2022, pp. 118–125.

[138] Chen Ma, Yuhong Chi, Donglai Hao, Xiongfei Ji, A new approach based on feature selection of light gradient boosting machine and transformer to predict circrna-disease associations, IEEE Access (2023).

[139] Norah Saeed Awn, Yiming Li, Baoying Zhao, Min Zeng, Min Li, Ldagso: predicting incrna-disease associations from graph sequences and disease ontology via deep learning techniques, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2022, pp. 398–403.

[140] Pengli Lu, Jicheng Jiang, Ae-rw: predicting mirna-disease associations by using autoencoder and random walk on mirna-gene-disease heterogeneous network, Comput. Biol. Chem. (2024) 108085.

[141] Latika Jindal, Aditi Sharma, K.D.V. Prasad, Azeem Irshad, Richard Rivera, Abdurakhimova Dilora Karimovna, A machine learning method for predicting disease-associated microrna connections using network internal topology data, Healthcare Anal. 4 (2023) 100215.

[142] Xun Wang, Fuyu Wang, Xinzeng Wang, Sibo Qiao, Yu Zhuang, Demlp: deepwalk embedding in mlp for mirna-disease association prediction, J. Sens. 2021 (2021) 1–8.

[143] Chunyan Li, Hongju Liu, Qian Hu, Jinlong Que, Junfeng Yao, A novel computational model for predicting microrna–disease associations based on heterogeneous graph convolutional networks, Cells 8 (9) (2019) 977.

[144] Tao Duan, Zhufang Kuang, Jiaqi Wang, Zhihao Ma, Gbdtlrl2d predicts lncrna–disease associations using metagraph2vec and k-means based on heterogeneous network, Front. Cell Dev. Biol. 9 (2021) 753027.

[145] Xiaoping Sun, Xingshuai Ren, Jie Zhang, Yunzhi Nie, Shan Hu, Xiao Yang, Shoufeng Jiang, Discovering mirnas associated with multiple sclerosis based on network representation learning and deep learning methods, Front. Genet. 13 (2022) 899340.

[146] Kai Zheng, Zhu-Hong You, Lei Wang, Zhen-Hao Guo, imda-bn: identification of mirna-disease associations based on the biological network and graph embedding algorithm, Comput. Struct. Biotechnol. J. 18 (2020) 2391–2400.

[147] Zhen Tian, Chenguang Han, Lewen Xu, Zhixia Teng, Wei Song, Mgcnss: mirna–disease association prediction with multi-layer graph convolution and distance-based negative sample selection strategy, Brief. Bioinform. 25 (3) (2024) bbae168.

[148] Xinru Ruan, Changzhi Jiang, Peixuan Lin, Yuan Lin, Juan Liu, Shaohui Huang, Xiangrong Liu, Msgcl: inferring mirna–disease associations based on multi-view self-supervised graph structure contrastive learning, Brief. Bioinform. 24 (2) (2023) bbac623.

[149] Lei Xu, Xiangzheng Fu, Linlin Zhuo, Zhecheng Zhou, Xuefeng Liao, Sha Tian, Ruofei Kang, Yifan Chen, Sgae-mda: exploring the mirna-disease associations in herbal medicines based on semi-supervised graph autoencoder, Methods 221 (2024) 73–81.

[150] Boya Ji, Haitao Zou, Liwen Xu, Xiaolan Xie, Shaoliang Peng, Muscle: multi-view and multi-scale attentional feature fusion for microrna–disease associations prediction, Brief. Bioinform. 25 (3) (2024) bbae167.

[151] Xujun Liang, Ming Guo, Longying Jiang, Ying Fu, Pengfei Zhang, Yongheng Chen, Predicting mirna–disease associations by combining graph and hypergraph convolutional network, in: Interdisciplinary Sciences: Computational Life Sciences, 2024, pp. 1–15.

[152] Wengang Wang, Hailin Chen, Predicting mirna-disease associations based on lncrna–mirna interactions and graph convolution networks, Brief. Bioinform. 24 (1) (2023) bbac495.

[153] Wen Cao, Yang Chen, Jian-Ye Yang, Fei-Yang Xue, Zhan-Hui Yu, Jing Feng, Ze-Jun Wu, Jing Gong, Xiao-Hui Niu, Metapath-aggregated multilevel graph embedding for mirna–disease association prediction, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 468–473.

[154] Yuchong Gong, Yanqing Niu, Wen Zhang, Xiaohong Li, A network embedding-based multiple information integration method for the mirna-disease association prediction, BMC Bioinform. 20 (2019) 1–13.

[155] Shiyuan Li, Qingfeng Chen, Zhixian Liu, Shirui Pan, Shichao Zhang, Bi-sgtar: a simple yet efficient model for circrna-disease association prediction based on known association pair only, Knowl.-Based Syst. 291 (2024) 111622.

[156] Zhonghao Lu, Hua Zhong, Lin Tang, Jing Luo, Wei Zhou, Lin Liu, Predicting lncrna-disease associations based on heterogeneous graph convolutional generative adversarial network, PLoS Comput. Biol. 19 (11) (2023) e1011634.

[157] Liqian Zhou, Xinhuai Peng, Lijun Zeng, Lihong Peng, Finding potential lncrna–disease associations using a boosting-based ensemble learning model, Front. Genet. 15 (2024) 1356205.

[158] Zequn Zhang, Junlin Xu, Yanan Wu, Niannian Liu, Yinglong Wang, Ying Liang, Capsnet-lda: predicting lncrna-disease associations using attention mechanism and capsule network based on multi-view data, Brief. Bioinform. 24 (1) (2023) bbac531.

[159] Wei Liu, Ting Tang, Xu Lu, Xiangzheng Fu, Yu Yang, Li Peng, Mpclcda: predicting circrna–disease associations by using automatically selected meta-path and contrastive learning, Brief. Bioinform. 24 (4) (2023) bbad227.

[160] Wen-Yue Kang, Chun-Hou Zheng, Ying-Lian Gao, Juan Wang, Junliang Shang, Jin-Xing Liu, Grpgat: predicting circrna-disease associations based on graph random propagation network and graph attention network, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2023, pp. 233–236.

[161] Yao Fu, Runtao Yang, Lina Zhang, Xu Fu, Hgecda: a heterogeneous graph embedding model for circrna-disease association prediction, IEEE J. Biomed. Health Inform. (2023).

[162] Dengju Yao, Bo Zhang, Xiangkui Li, Xiaojuan Zhan, Xiaorong Zhan, Binbin Zhang, Applying negative sample denoising and multi-view feature for lncrna-disease association prediction, Front. Genet. 14 (2024) 1332273.

[163] Qingfeng Chen, Junlai Qiu, Wei Lan, Junyue Cao, Similarity-guided graph contrastive learning for lncrna-disease association prediction, J. Mol. Biol. (2024) 168609.

[164] Shengchang Wang, Jiaqing Qiao, Shou Feng, Prediction of lncrna and disease associations based on residual graph convolutional networks with attention mechanism, Sci. Rep. 14 (1) (2024) 5185.

[165] Ying Liang, Ze-Qun Zhang, Nian-Nian Liu, Ya-Nan Wu, Chang-Long Gu, Ying-Long Wang, Magcnse: predicting lncrna-disease associations using multi-view attention graph convolutional network and stacking ensemble model, BMC Bioinform. 23 (1) (2022) 189.

[166] Yong Han, Shao-Wu Zhang, ncrpi-lgat: prediction of ncrna-protein interactions with line graph attention network framework, Comput. Struct. Biotechnol. J. 21 (2023) 2286–2295.

[167] Li Hui, Bin Wu, Miaomiao Sun, Zhenfeng Zhu, Kuisheng Chen, Hong Ge, Cross-domain contrastive graph neural network for lncrna-protein interaction prediction, Knowl.-Based Syst. (2024) 111901.

[168] Jingxuan Zhao, Jianqiang Sun, Stella C. Shuai, Qi Zhao, Jianwei Shuai, Predicting potential interactions between lncrnas and proteins via combined graph auto-encoder methods, Brief. Bioinform. 24 (1) (2023) bbac527.

[169] Meng-Meng Wei, Chang-Qing Yu, Li-Ping Li, Zhu-Hong You, Zhong-Hao Ren, Yong-Jian Guan, Xin-Fei Wang, Yue-Chao Li, Lpih2v: lncrna-protein interactions prediction using hin2vec based on heterogeneous networks model, Front. Genet. 14 (2023) 1122909.

[170] Hai-Cheng Yi, Zhu-Hong You, Li Cheng, Xi Zhou, Tong-Hai Jiang, Xiao Li, Yan-Bin Wang, Learning distributed representations of rna and protein sequences and its application for predicting lncrna-protein interactions, Comput. Struct. Biotechnol. J. 18 (2020) 20–26.

[171] Yifei Wang, Pengju Ding, Congjing Wang, Shiyue He, Xin Gao, Bin Yu, Rpi-ggcn: prediction of rna–protein interaction based on interpretability gated graph convolution neural network and co-regularized variational autoencoders, IEEE Trans. Neural Netw. Learn. Syst. (2024).

[172] Yuyao Yan, Wenran Li, Sijia Wang, Tao Huang, Seq-rbppred: predicting rna-binding proteins from sequence, ACS Omega 9 (11) (2024) 12734–12742.

[173] Dilan Lasantha, Sugandima Vidanagamachchi, Sam Nallaperuma, Criecnn: ensemble convolutional neural network and advanced feature extraction methods for the precise forecasting of circrna-rbp binding sites, Comput. Biol. Med. 174 (2024) 108466.

[174] Wenhao Jin, Kristopher W. Brannan, Katannya Kapeli, Samuel S. Park, Hui Qing Tan, Maya L. Gosztyla, Mayuresh Mujumdar, Joshua Ahdout, Bryce Henroid, Katherine Rothamel, et al., Hydra: deep-learning models for predicting rna-binding capacity from protein interaction association context and protein sequence, Mol. Cell 83 (14) (2023) 2595–2611.

[175] Chao Cao, Chunyu Wang, Shuhong Yang, Quan Zou, Circsi-ssl: circrna-binding site identification based on self-supervised learning, Bioinformatics 40 (1) (2024) btae004.

[176] Li Lei, Zhigang Xue, Xiuquan Du, Ascrb: multi-view based attentional feature selection for circrna-binding site prediction, Comput. Biol. Med. 162 (2023) 107077.

[177] Niannian Liu, Zequn Zhang, Yanan Wu, Yinglong Wang, Ying Liang, Crbsp: prediction of circrna-rbp binding sites based on multimodal intermediate fusion, IEEE/ACM Trans. Comput. Biol. Bioinform. (2023).

[178] Zheng Ma, Zhan-Li Sun, Mengya Liu, Crbp-hfef: prediction of rbp-binding sites on circrnas based on hierarchical feature expansion and fusion, Interdiscip. Sci. Comput. Life Sci. 15 (3) (2023) 465–479.

[179] Lei Deng, Youzhi Liu, Yechuan Shi, Wenhao Zhang, Chun Yang, Hui Liu, Deep neural networks for inferring binding sites of rna-binding proteins by using distributed representations of rna primary sequence and secondary structure, BMC Genomics 21 (2020) 1–10.

[180] Yixuan Qiao, Rui Yang, Yang Liu, Jiaxin Chen, Lianhe Zhao, Peipei Huo, Zhihao Wang, Dechao Bu, Yang Wu, Yi Zhao, Deepfusion: a deep bimodal information fusion network for unraveling protein-rna interactions using in vivo rna structures, Comput. Struct. Biotechnol. J. 23 (2024) 617–625.

[181] Liwei Liu, Yixin Wei, Zhebin Tan, Qi Zhang, Jianqiang Sun, Qi Zhao, Predicting circrna-rbp binding sites using a hybrid deep neural network, Interdiscip. Sci. Comput. Life Sci. (2024) 1–14.

[182] Xilin Shen, Xiangchun Li, Reformer: deep learning model for characterizing protein-rna interactions from sequence at single-base resolution, bioRxiv (2024), pages 2024–01.

[183] Meng-Meng Wei, Chang-Qing Yu, Li-Ping Li, Zhu-Hong You, Lei Wang, Bcmcmi: a fusion model for predicting circrna-mirna interactions combining semantic and meta-path, J. Chem. Inf. Model. 63 (16) (2023) 5384–5394.

[184] Zheng-Yang Zhao, Jie Lin, Zhen Wang, Jian-Xin Guo, Xin-Ke Zhan, Yu-An Huang, Chuan Shi, Wen-Zhun Huang, et al., Sebglma: semantic embedded bipartite graph network for predicting lncrna-mirna associations, Int. J. Intell. Syst. (2023) 2023.

[185] Lu-Xiang Guo, Lei Wang, Zhu-Hong You, Chang-Qing Yu, Meng-Lei Hu, Bo-Wei Zhao, Yang Li, Likelihood-based feature representation learning combined with neighborhood information for predicting circrna–mirna associations, Brief. Bioinform. 25 (2) (2024) bbae020.

[186] Xin-Fei Wang, Chang-Qing Yu, Zhu-Hong You, Yan Qiao, Zheng-Wei Li, Wen-Zhun Huang, Ji-Ren Zhou, Hai-Yan Jin, Ks-cmi: a circrna-mirna interaction prediction method based on the signed graph neural network and denoising autoencoder, iScience 26 (8) (2023).

[187] Nan Sheng, Yan Wang, Lan Huang, Ling Gao, Yangkun Cao, Xuping Xie, Yuan Fu, Multi-task prediction-based graph contrastive learning for inferring the relationship among lncrnas, mirnas and diseases, Brief. Bioinform. 24 (5) (2023) bbad276.

[188] Zutan Li, Bingbing Jin, Jingya Fang, Metaac4c: a multi-module deep learning framework for accurate prediction of n4-acetylcytidine sites based on pre-trained bidirectional encoder representation and generative adversarial networks, Genomics 116 (1) (2024) 110749.

[189] Zhongxing Xu, Xuan Wang, Jia Meng, Lin Zhang, Bowen Song, m5u-gepred: prediction of rna 5-methyluridine sites based on sequence-derived and graph embedding features, Front. Microbiol. 14 (2023) 1277099.

[190] Waleed Alam, Muhammad Tahir, Shahid Hussain, Sarah Gul, Maqsood Hayat, Reyazur Rashid Irshad, Fabiano Pallonetto, Unveiling the potential pattern representation of rna 5-methyluridine modification sites through a novel feature fusion model leveraging convolutional neural network and tetranucleotide composition, IEEE Access (2024).

[191] Necla Nisa Soylu, Emre Sefer, Bert2ome: prediction of 2'-o-methylation modifications from rna sequence by transformer architecture based on bert, IEEE/ACM Trans. Comput. Biol. Bioinform. (2023).

[192] Rulan Wang, Chia-Ru Chung, Tzong-Yi Lee, Interpretable multi-scale deep learning for rna methylation analysis across multiple species, Int. J. Mol. Sci. 25 (5) (2024) 2869.

[193] Linshu Wang, Yuan Zhou, Mrm-bert: a novel deep neural network predictor of multiple rna modifications by fusing bert representation and sequence features, RNA Biol. 21 (1) (2024) 1–10.

[194] Ting-He Zhang, Sumin Jo, Michelle Zhang, Kai Wang, Shou-Jiang Gao, Yufei Huang, Understanding ythdf2-mediated mrna degradation by m6a-bert-deg, arXiv preprint arXiv:2401.08004, 2024.

[195] Shuang Xiang, Te Zhang, Minghao Wu, M6atmr: identifying n6-methyladenosine sites through rna sequence similarity matrix reconstruction guided by transformer, PeerJ 11 (2023) e15899.

[196] Guohua Huang, Xiaohong Huang, Jinyun Jiang, Deepm6a-mt: a deep learning-based method for identifying rna n6-methyladenosine sites in multiple tissues, Methods 226 (2024) 1–8.

[197] Haokai Ye, Tenglong Li, Daniel J. Rigden, Zhen Wei, m6acali: machine learning-powered calibration for accurate m6a detection in merip-seq, Nucleic Acids Res. 52 (9) (2024) 4830–4842.

[198] Iman Nazari, Muhammad Tahir, Hilal Tayara, Kil To Chong, in6-methyl (5-step): identifying rna n6-methyladenosine sites using deep learning mode via chou's 5-step rules and chou's general pseknc, Chemom. Intell. Lab. Syst. 193 (2019) 103811.

[199] Quan Zou, Pengwei Xing, Leyi Wei, Bin Liu, Gene2vec: gene subsequence embedding for prediction of mammalian n6-methyladenosine sites from mrna, RNA 25 (2) (2019) 205–218.

[200] Gang Tu, Xuan Wang, Rong Xia, Bowen Song, m6a-tcpred: a web server to predict tissue-conserved human m6a sites using machine learning approach, BMC Bioinform. 25 (1) (2024) 127.

[201] Honglei Wang, Wenliang Zeng, Xiaoling Huang, Zhaoyang Liu, Yanjing Sun, Lin Zhang, Mttlm 6 a: a multi-task transfer learning approach for base-resolution mrna m 6 a site prediction based on an improved transformer, Math. Biosci. Eng. 21 (1) (2024) 272–299.

[202] Jie Jiang, Bowen Song, Jia Meng, Jingxian Zhou, Tissue-specific rna methylation prediction from gene expression data using sparse regression models, Comput. Biol. Med. 169 (2024) 107892.

[203] Shengli Zhang, Yujie Xu, Yunyun Liang, Tmsc-m7g: a transformer architecture based on multi-sense-scaled embedding features and convolutional neural network to identify rna n7-methylguanosine sites, Comput. Struct. Biotechnol. J. 23 (2024) 129–139.

[204] Lu Zhang, Xinyi Qin, Min Liu, Guangzhong Liu, Yuxiao Ren, et al., Bert-m7g: a transformer architecture based on bert and stacking ensemble to identify rna n7-methylguanosine sites from sequence information, Comput. Math. Methods Med. (2021) 2021.

[205] Muhammad Tahir, Maqsood Hayat, Rahim Khan, Kil To Chong, An effective deep learning-based architecture for prediction of n7-methylguanosine sites in health systems, Electronics 11 (12) (2022) 1917.

[206] Yonglin Zhang, Lezheng Yu, Runyu Jing, Bin Han, Jiesi Luo, Fast and efficient design of deep neural networks for predicting n7-methylguanosine sites using autobioseqpy, ACS Omega 8 (22) (2023) 19728–19740.

[207] Md Mehedi Hasan, Sho Tsukiyama, Jae Youl Cho, Hiroyuki Kurata, Md Ashad Alam, Xiaowen Liu, Balachandran Manavalan, Hong-Wen Deng, Deepm5c: a deep-learning-based hybrid framework for identifying human rna n5-methylcytosine sites using a stacking strategy, Mol. Ther. 30 (8) (2022) 2856–2867.

[208] Hiroyuki Kurata, Md Harun-Or-Roshid, Md Mehedi Hasan, Sho Tsukiyama, Kazuhiro Maeda, Balachandran Manavalan, Mlm5c: a high-precision human rna 5-methylcytosine sites predictor based on a combination of hybrid machine learning models, Methods (2024).

[209] Taoning Chen, Tingfang Wu, Deng Pan, Jinxing Xie, Jin Zhi, Xuejiao Wang, Lijun Quan, Qiang Lyu, Transrnam: identifying twelve types of rna modifications by an interpretable multi-label deep learning model based on transformer, IEEE/ACM Trans. Comput. Biol. Bioinform. (2023).

[210] Honglei Wang, Tao Huang, Dong Wang, Wenliang Zeng, Yanjing Sun, Lin Zhang, Mscan: multi-scale self- and cross-attention network for rna methylation site prediction, BMC Bioinform. 25 (1) (2024) 32.

[211] Honglei Wang, Hui Liu, Tao Huang, Gangshen Li, Lin Zhang, Yanjing Sun, Emdlp: ensemble multiscale deep learning model for rna methylation site prediction, BMC Bioinform. 23 (1) (2022) 221.

[212] Mingzhao Wang, Haider Ali, Yandi Xu, Juanying Xie, Shengquan Xu, Bipstp: sequence feature encoding method for identifying different rna modifications with bidirectional position-specific trinucleotides propensities, J. Biol. Chem. 300 (4) (2024).

[213] Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, Yuedong Yang, Self-supervised learning on millions of primary rna sequences from 72 vertebrates improves sequence-based rna splicing prediction, Brief. Bioinform. 25 (3) (2024) bbae163.

[214] Shentong Mo, Xi Fu, Chenyang Hong, Yizhen Chen, Yuxuan Zheng, Xiangru Tang, Zhiqiang Shen, Eric P. Xing, Yanyan Lan, Multi-modal self-supervised pre-training for regulatory genome across cell types, arXiv preprint arXiv:2110.05231, 2021.

[215] Mhaned Oubounyt, Zakaria Louadi, Hilal Tayara, Kil To Chong, Deep learning models based on distributed feature representations for alternative splicing prediction, IEEE Access 6 (2018) 58826–58834.

[216] Yekaterina Shulgina, Marena I. Trinidad, Conner J. Langeberg, Hunter Nisonoff, Seyone Chithrananda, Petr Skopintsev, Amos J. Nissley, Jaymin Patel, Ron S. Boger, Honglue Shi, et al., Rna language models predict mutations that improve rna function, bioRxiv (2024).

[217] Nicholas Boyd, Brandon M. Anderson, Brent Townshend, Ryan Chow, Connor J. Stephens, Ramya Rangan, Matias Kaplan, Meredith Corley, Akshay Tambe, Yuzu Ido, et al., Atom-1: a foundation model for rna structure and function built on chemical mapping data, bioRxiv (2023), pages 2023–13.

[218] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al., Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions, arXiv preprint arXiv:2204.00300, 2022.

[219] Jiacheng Wang, Jingpu Zhang, Yideng Cai, Lei Deng, Deepmir2go: inferring functions of human micrornas using a deep multi-label classification model, Int. J. Mol. Sci. 20 (23) (2019) 6046.

[220] K.H. Joerg Franke, Frederic Runge, Ryan Koeksal, Rolf Backofen, Frank Hutter, Rnaformer: a simple yet effective deep learning model for rna secondary structure prediction, bioRxiv (2024), pages 2024–02.

[221] Rafael Josip Penić, Vlaš, Rinalmo: General-Purpose Rna Language Models Can Generalize Well on Structure Prediction Tasks, 2024.

[222] Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, et al., Multiple sequence alignment-based rna language model and its application to structural inference, Nucleic Acids Res. 52 (1) (2024) e3.

[223] Colin Hall Kalicki, Esin Darici Haritaoglu, Rnabert: Rna family classification and secondary structure prediction with bert pretrained on rna sequences, https://cs230.stanford.edu/projects_fall_2022/reports/88.pdf.

[224] Wenkai Wang, Chenjie Feng, Renmin Han, Ziyi Wang, Lisha Ye, Zongyang Du, Hong Wei, Fa Zhang, Zhenling Peng, Jianyi Yang, trrosettarna: automated prediction of rna 3d structure with transformer network, Nat. Commun. 14 (1) (2023) 7266.

[225] Yinchao Fei, Hao Zhang, Yili Wang, Zhen Liu, Yuanning Liu, Ltpconstraint: a transfer learning based end-to-end method for rna secondary structure prediction, BMC Bioinform. 23 (1) (2022) 354.

[226] Tiansu Gong, Dongbo Bu, Language models enable zero-shot prediction of rna secondary structures including pseudoknots, bioRxiv (2024), pages 2024–01.

[227] Yili Wang, Yuanning Liu, Shuo Wang, Zhen Liu, Yubing Gao, Hao Zhang, Liyan Dong, Attfold: rna secondary structure prediction with pseudoknots based on attention mechanism, Front. Genet. 11 (2020) 612086.

[228] Zhengqiao Zhao, Stephen Woloszynek, Felix Agbavor, Joshua Chang Mell, Bahrad A. Sokhansanj, Gail L. Rosen, Learning, visualizing and exploring 16s rrna structure using an attention-based deep neural network, PLoS Comput. Biol. 17 (9) (2021) e1009345.

[229] Xiangyun Qiu, Sequence similarity governs generalizability of de novo deep learning models for rna secondary structure prediction, PLoS Comput. Biol. 19 (4) (2023) e1011047.

[230] Chun-Chi Chen, Yi-Ming Chan, Redfold: accurate rna secondary structure prediction using residual encoder-decoder network, BMC Bioinform. 24 (1) (2023) 122.

[231] Yuchen Wang, Xingjian Chen, Zetian Zheng, Lei Huang, Weidun Xie, Fuzhou Wang, Zhaolei Zhang, Ka-Chun Wong, scgreat: transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics, iScience (2024).

[232] Anwar Khan, Boreom Lee, Deepgene transformer: transformer for the gene expression-based classification of cancer subtypes, Expert Syst. Appl. 226 (2023) 120047.

[233] Ting-He Zhang, Md Musaddaqul Hasib, Yu-Chiao Chiu, Zhi-Feng Han, Yu-Fang Jin, Mario Flores, Yidong Chen, Yufei Huang, Transformer for gene expression modeling (t-gem): an interpretable deep learning model for gene expression-based phenotype predictions, Cancers 14 (19) (2022) 4763.

[234] Ashley Nicole Babjac, Zhixiu Lu, Scott J. Emrich, Codonbert: using bert for sentiment analysis to better predict genes with low expression, in: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2023, pp. 1–6.

[235] Jing Xu, Aidi Zhang, Fang Liu, Xiujun Zhang, Stgrns: an interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data, Bioinformatics 39 (4) (2023) btad165.

[236] Stephen Woloszynek, Zhengqiao Zhao, Jian Chen, Gail L. Rosen, 16s rrna sequence embeddings: meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses, PLoS Comput. Biol. 15 (2) (2019) e1006721.

[237] Jiazheng Miao, Tianlai Chen, Mustafa Misir, Yajuan Lin, Deep learning for predicting 16s rrna gene copy number, bioRxiv (2022), pages 2022–11.

[238] Tingpeng Yang, Yu Wang, Yonghong He, Tec-mitarget: enhancing microrna target prediction based on deep learning of ribonucleic acid sequences, BMC Bioinform. 25 (1) (2024) 159.

[239] Jialin Zhang, Haoran Zhu, Yin Liu, Xiangtao Li, mitds: uncovering mirna-mrna interactions with deep learning for functional target prediction, Methods 223 (2024) 65–74.

[240] Jan Przybyszewski, Maciej Malawski, Sabina Lichołai, Graphtar: applying word2vec and graph neural networks to mirna target prediction, BMC Bioinform. 24 (1) (2023) 436.

[241] Jiayu Xu, Nan Xu, Weixin Xie, Chengkui Zhao, Lei Yu, Weixing Feng, Bert-sirna: sirna target prediction based on bert pre-trained interpretable model, Gene (2024) 148330.

[242] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, Bo Wang, scgpt: toward building a foundation model for single-cell multi-omics using generative ai, Nat. Methods (2024) 1–11.

[243] Yushan Qiu, Dong Guo, Pu Zhao, Quan Zou, scmnmf: a novel method for single-cell multi-omics clustering based on matrix factorization, Brief. Bioinform. 25 (3) (2024) bbae228.

[244] Joseph Lilleberg, Yun Zhu, Yanqing Zhang, Support vector machines and word2vec for text classification with semantic features, in: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2015, pp. 136–140.

[245] Will Y. Zou, Richard Socher, Daniel Cer, Christopher D. Manning, Bilingual word embeddings for phrase-based machine translation, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1393–1398.

[246] Md Al-Amin, Md Saiful Islam, Shapan Das Uzzal, Sentiment analysis of bengali comments with word2vec and sentiment information of words, in: 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2017, pp. 186–190.

[247] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, Radu Prodan, Welfake: word embedding over linguistic features for fake news detection, IEEE Trans. Comput. Soc. Syst. 8 (4) (2021) 881–893.

[248] Hong Shi, Xiaomeng Zhang, Lin Tang, Lin Liu, Heterogeneous graph neural network for lncrna-disease association prediction, Sci. Rep. 12 (1) (2022) 17519.

[249] Yue Liu, Shu-Lin Wang, Jun-Feng Zhang, Wei Zhang, Wen Li, Lncrna-disease associations prediction based on neural network-based matrix factorization, IEEE Access 11 (2020) 59071–59080.

[250] Zhengfeng Wang, Xiujuan Lei, Prediction of rbp binding sites on circrnas using an lstm-based deep sequence learning architecture, Brief. Bioinform. 22 (6) (2021) bbab342.

[251] Lei Deng, Youzhi Liu, Yechuan Shi, Hui Liu, A deep neural network approach using distributed representations of rna sequence and structure for identifying binding site of rna-binding proteins, in: 2019 Ieee International Conference on Bioinformatics and Biomedicine (Bibm), IEEE, 2019, pp. 12–17.

[252] Michal Ziemski, Treepop Wisanwanichthan, Nicholas A. Bokulich, Benjamin D. Kaehler, Beating naive Bayes at taxonomic classification of 16s rrna gene sequences, Front. Microbiol. 12 (2021) 644487.

[253] Yuzhuo Sun, Fei Xiong, Yongke Sun, Youjie Zhao, Yong Cao, et al., A mirna target prediction model based on distributed representation learning and deep learning, Comput. Math. Methods Med. (2022) 2022.

[254] Guodong Li, Bowei Zhao, Xiaorui Su, Yue Yang, Pengwei Hu, Xi Zhou, Lun Hu, Discovering consensus regions for interpretable identification of rna n6-methyladenosine modification sites via graph contrastive clustering, IEEE J. Biomed. Health Inform. (2024).

[255] Yongxian Fan, Guicong Sun, Xiaoyong Pan, Elmo4m6a: a contextual language embedding-based predictor for detecting rna n6-methyladenosine sites, IEEE/ACM Trans. Comput. Biol. Bioinform. 20 (2) (2022) 944–954.

[256] Dong-Ling Yu, Zu-Guo Yu, Guo-Sheng Han, Jinyan Li, Vo Anh, Heterogeneous types of mirna-disease associations stratified by multi-layer network embedding and prediction, Biomedicines 9 (9) (2021) 1152.

[257] Qiu Xiao, Yu Fu, Yide Yang, Jianhua Dai, Jiawei Luo, Nsl2cd: identifying potential circrna–disease associations based on network embedding and subspace learning, Brief. Bioinform. 22 (6) (2021) bbab177.

[258] Ji-Ren Zhou, Zhu-Hong You, Li Cheng, Bo-Ya Ji, Prediction of lncrna-disease associations via an embedding learning hope in heterogeneous information networks, Mol. Ther. Nucleic Acids 23 (2021) 277–285.

[259] Jianwei Li, Jianing Li, Mengfan Kong, Duanyang Wang, Kun Fu, Jiangcheng Shi, Svdnvlda: predicting lncrna-disease associations by singular value decomposition and node2vec, BMC Bioinform. 22 (2021) 1–18.

[260] Zi-Ang Shen, Tao Luo, Yuan-Ke Zhou, Han Yu, Pu-Feng Du, Npi-gnn: predicting ncrna–protein interactions with deep graph neural networks, Brief. Bioinform. 22 (5) (2021) bbab051.

[261] Fansen Xie, Ziqi Yang, Jinmiao Song, Qiguo Dai, Xiaodong Duan, Dhnlda: a novel deep hierarchical network based method for predicting lncrna-disease associations, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (6) (2021) 3395–3403.

[262] Manu Madhavan, G. Gopakumar, Dbnlda: deep belief network based representation learning for lncrna-disease association prediction, Appl. Intell. 52 (5) (2022) 5342–5352.

[263] Iz Beltagy, Matthew E. Peters, Arman Cohan, Longformer: the long-document transformer, arXiv preprint arXiv:2004.05150, 2020.

[264] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al., Big bird: transformers for longer sequences, Adv. Neural Inf. Process. Syst. 33 (2020) 17283–17297.

[265] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[266] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.

[267] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., Improving Language Understanding by Generative Pre-Training, 2018.

[268] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.

[269] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[270] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774, 2023.

[271] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118.

[272] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, Alex Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, Adv. Neural Inf. Process. Syst. 34 (2021) 29287–29303.

[273] Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, Alexander Rives, Msa transformer, in: International Conference on Machine Learning, PMLR, 2021, pp. 8844–8856.

[274] Jörg K.H. Franke, Frederic Runge, Frank Hutter, Scalable deep learning for rna secondary structure prediction, arXiv preprint arXiv:2307.10073, 2023.

[275] Ziniu Hu, Yuxiao Dong, Kuansan Wang, Yizhou Sun, Heterogeneous graph transformer, in: Proceedings of the Web Conference 2020, 2020, pp. 2704–2710.

[276] Yunqi Wan, Zhenran Jiang, Transcrispr: transformer based hybrid model for predicting crispr/cas9 single guide rna cleavage efficiency, IEEE/ACM Trans. Comput. Biol. Bioinform. 20 (2) (2022) 1518–1528.

[277] Anwar Khan, Boreom Lee, Gene transformer: transformers for the gene expression-based classification of lung cancer subtypes, arXiv preprint arXiv:2108.11833, 2021.

[278] Chao Cao, Shuhong Yang, Mengli Li, Chungui Li, Circssnn: circrna-binding site prediction via sequence self-attention neural networks with pre-normalization, BMC Bioinform. 24 (1) (2023) 220.

[279] Xiuquan Du, Zhigang Xue, Jlcrb: a unified multi-view-based joint representation learning for circrna binding sites prediction, J. Biomed. Inform. 136 (2022) 104231.

[280] Jiren Zhou, Xinfei Wang, Rui Niu, Xuequn Shang, Jiayu Wen, Predicting circrna-mirna interactions utilizing transformer-based rna sequential learning and high-order proximity preserved embedding, iScience 27 (1) (2024).

[281] Qianyue Li, Xin Cheng, Chen Song, Taigang Liu, M6a-bert-stacking: a tissue-specific predictor for identifying rna n6-methyladenosine sites based on bert and stacking strategy, Symmetry 15 (3) (2023) 731.

[282] Sirui Liang, Yanxi Zhao, Junru Jin, Jianbo Qiao, Ding Wang, Yu Wang, Leyi Wei, Rm-lr: a long-range-based deep learning model for predicting multiple types of rna modifications, Comput. Biol. Med. 164 (2023) 107238.

[283] Hui Wan, Musu Yuan, Yiwei Fu, Minghua Deng, Continually adapting pre-trained language model to universal annotation of single-cell rna-seq data, Brief. Bioinform. 25 (2) (2024) bbae047.

[284] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, Jianhua Yao, scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data, Nat. Mach. Intell. 4 (10) (2022) 852–866.

[285] Chengqian Lu, Lishen Zhang, Min Zeng, Wei Lan, Jianxin Wang, Identifying disease-associated circrnas based on edge-weighted graph attention and heterogeneous graph neural network, bioRxiv (2022), pages 2022–05.

[286] Xue-Chan Tian, Zhao-Yang Chen, Shuai Nie, Tian-Le Shi, Xue-Mei Yan, Yu-Tao Bao, Zhi-Chao Li, Hai-Yao Ma, Kai-Hua Jia, Wei Zhao, et al., Plant-lncpipe: a computational pipeline providing significant improvement in plant lncrna identification, in: Horticulture Research, 2024, uhae041.

[287] Sho Tsukiyama, Md Mehedi Hasan, Hong-Wen Deng, Hiroyuki Kurata, Bert6ma: prediction of dna n6-methyladenine site using deep learning-based approaches, Brief. Bioinform. 23 (2) (2022) bbac053.

[288] Yingwen Zhao, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, Predicting protein functions based on heterogeneous graph attention technique, IEEE J. Biomed. Health Inform. (2024).

[289] Sovan Saha, Piyali Chatterjee, Subhadip Basu, Mita Nasipuri, Epi-sf: essential protein identification in protein interaction networks using sequence features, PeerJ 12 (2024) e17010.

[290] Rufeng Lei, Jianhua Jia, Lulu Qin, Xin Wei, ipro2l-dg: hybrid network based on improved densenet and global attention mechanism for identifying promoter sequences, Heliyon 10 (6) (2024).

[291] Guang Yang, Jianing Li, Jinlu Hu, Jian-Yu Shi, Recognition of cyanobacteria promoters via Siamese network-based contrastive learning under novel non-promoter generation, Brief. Bioinform. 25 (3) (2024) bbae193.

[292] Faiza Mehmood, Muhammad Usman Ghani, Muhammad Nabeel Asim, Rehab Shahzadi, Aamir Mehmood, Waqar Mahmood, Mpf-net: A computational multi-regional solar power forecasting framework, Renew. Sustain. Energy Rev. 151 (2021) 111559.

[293] Ahtisham Fazeel Abbasi, Muhammad Nabeel Asim, Sheraz Ahmed, Sebastian Vollmer, Andreas Dengel, Survival prediction landscape: an in-depth systematic literature review on activities, methods, tools, diseases, and databases, medRxiv (2024), pages 2024–01.

[294] Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Imran Razzak, Andreas Dengel, Sheraz Ahmed, Histone-net: a multi-paradigm computational framework for histone occupancy and modification prediction, Complex Intell. Syst. 9 (1) (2023) 399–419.

[295] Piotr Grabowski, Juri Rappsilber, A primer on data analytics in functional genomics: how to move from data to insight?, Trends Biochem. Sci. 44 (1) (2019) 21–32.

[296] Aimin Yang, Wei Zhang, Jiahao Wang, Ke Yang, Yang Han, Limin Zhang, Review on the application of machine learning algorithms in the sequence data mining of dna, Front. Bioeng. Biotechnol. 8 (2020) 1032.

[297] Faiza Mehmood, Hina Ghafoor, Muhammad Nabeel Asim, Muhammad Usman Ghani, Waqar Mahmood, Andreas Dengel, Passion-net: a robust precise and explainable predictor for hate speech detection in roman Urdu text, Neural Comput. Appl. 36 (6) (2024) 3077–3100.

[298] S.W. Chong, Philip J. Peyton, A meta-analysis of the accuracy and precision of the ultrasonic cardiac output monitor (uscom), Anaesthesia 67 (11) (2012) 1266–1271.

[299] ChenRui Duan, Zelin Zang, Yongjie Xu, Hang He, Zihan Liu, Zijia Song, Ju-Sheng Zheng, Stan Z. Li, Fgbert: function-driven pre-trained gene language model for metagenomics, arXiv preprint arXiv:2402.16901, 2024.

[300] Lihua Chen, Zhenkang Hu, Yuzhi Rong, Bao Lou, Deep2pep: a deep learning method in multi-label classification of bioactive peptide, Comput. Biol. Chem. 109 (2024) 108021.

[301] Muhammad Wasim, Waqar Mahmood, Muhammad Nabeel Asim, Muhammad Usman Khan, Multi-label question classification for factoid and list type questions in biomedical question answering, IEEE Access 7 (2018) 3882–3896.

[302] Summra Saleem, Muhammad Nabeel Asim, Ludger Van Elst, Peter Schichtel, Andreas Dengel, Rprp-sap: a robust and precise resnet predictor for steering angle prediction of autonomous vehicles, IEEE Access (2024).

[303] Alexei Botchkarev, Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology, arXiv preprint arXiv:1809.03006, 2018.

[304] Sungil Kim, Heeyoung Kim, A new metric of absolute percentage error for intermittent demand forecasts, Int. J. Forecast. 32 (3) (2016) 669–679.

[305] Muhammad Azhar, Philippe Blanc, Muhammad Asim, Shahid Imran, Nasir Hayat, Hamza Shahid, Hasnain Ali, et al., The evaluation of reanalysis and analysis products of solar radiation for Sindh province, Pakistan, Renew. Energy 145 (2020) 347–362.

[306] Muhammad Nabeel Asim, Muhammad Imran Malik, Andreas Dengel, Sheraz Ahmed, K-mer neural embedding performance analysis using amino acid codons, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[307] Etienne Lord, Abdoulaye Baniré Diallo, Vladimir Makarenkov, Classification of bioinformatics workflows using weighted versions of partitioning and hierarchical clustering algorithms, BMC Bioinform. 16 (2015) 1–19.

[308] Mohsen Rahmanian, Eghbal G. Mansoori, An unsupervised gene selection method based on multivariate normalized mutual information of genes, Chemom. Intell. Lab. Syst. 222 (2022) 104512.

[309] M. Bipul Hossen, M. Rabiul Auwul, Comparative study of k-means, partitioning around medoids, agglomerative hierarchical, and Diana clustering algorithms by using cancer datasets, Biomed. Stat. Inf. 5 (1) (2020) 20.

[310] Akhilesh Kumar Singh, Shantanu Mittal, Prashant Malhotra, Yash Vardhan Srivastava, Clustering evaluation by Davies-Bouldin index (dbi) in cereal data using k-means, in: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2020, pp. 306–310.

[311] Yichong Zhao, Kenta Oono, Hiroki Takizawa, Masaaki Kotera, Generrna: a generative pre-trained language model for de novo rna design, bioRxiv (2024), pages 2024–02.

[312] Nguyen Quoc, Khanh Le, Quang-Thai Ho, Deep transformers and convolutional neural network in identifying dna n6-methyladenine sites in cross-species genomes, Methods 204 (2022) 199–206.

[313] Pei Liu, Ying Liu, Jiawei Luo, Yue Li, Mirgraph: a transformer-based feature learning approach to identify microrna-target interactions by integrating heterogeneous graph network and sequence information, bioRxiv (2023), pages 2023–11.

[314] Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, Yuedong Yang, Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction, bioRxiv (2023), pages 2023–01.

[315] Yanrong Ji, Zhihan Zhou, Han Liu, Ramana V. Davuluri, Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome, Bioinformatics 37 (15) (2021) 2112–2120.

[316] Jiawei Zhang, Haopeng Zhang, Congying Xia, Li Sun, Graph-bert: only attention is needed for learning graph representations, arXiv preprint arXiv:2001.05140, 2020.

[317] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, Proteinbert: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (8) (2022) 2102–2110.

[318] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[319] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, Lukasz Okruszek, Detecting formal thought disorder by deep contextualized word representations, Psychiatry Res. 304 (2021) 114135.

[320] Tianjiao Zhang, Liangyu Li, Hailong Sun, Guohua Wang, Deepiteh: a deep learning framework for identifying tissue-specific ernas from the human genome, Bioinformatics 39 (6) (2023) btad375.

[321] Shanchen Pang, Yu Zhuang, Sibo Qiao, Fuyu Wang, Shudong Wang, Zhihan Lv, Dctgm: a novel dual-channel transformer graph model for mirna-disease association prediction, Cogn. Comput. (2022) 1–10.

[322] Weihua Li, Wenyang Liu, Yanbu Guo, Bingyi Wang, Hua Qing, Deep contextual representation learning for identifying essential proteins via integrating multisource protein features, Chin. J. Electron. 32 (4) (2023) 868–881.

[323] The-Anh Tran, Dinh-Minh Pham, Yu-Yen Ou, et al., An extensive examination of discovering 5-methylcytosine sites in genome-wide dna promoters using machine learning based approaches, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (1) (2021) 87–94.

[324] Nguyen Quoc Khanh Le, Edward Kien Yee Yapp, Nagarajan Nagasundaram, Hui-Yuan Yeh, Classifying promoters by interpreting the hidden information of dna sequences via deep learning and combination of continuous fasttext n-grams, Front. Bioeng. Biotechnol. 7 (2019) 305.

[325] Ke Cai, Yuan Zhu, A method for identifying essential proteins based on deep convolutional neural network architecture with particle swarm optimization, in: 2022 Asia Conference on Advanced Robotics, Automation, and Control Engineering (ARACE), IEEE, 2022, pp. 7–12.

[326] Hongxiao Wang, Hao Zheng, Danny Z. Chen, Tango: a go-term embedding based method for protein semantic similarity prediction, IEEE/ACM Trans. Comput. Biol. Bioinform. 20 (1) (2022) 694–706.

[327] Nada Al Taweraqi, Ross D. King, Improved prediction of gene expression through integrating cell signalling models with machine learning, BMC Bioinform. 23 (1) (2022) 323.

[328] Kaiyi Wu, Di Zhou, Donna Slonim, Xiaozhe Hu, Lenore Cowen, Melissa: semi-supervised embedding for protein function prediction across multiple networks, bioRxiv (2023), pages 2023–08.

[329] Cen Wan, Domenico Cozzetto, Rui Fa, David T. Jones, Using deep maxout neural networks to improve the accuracy of function prediction from protein interaction networks, PLoS ONE 14 (7) (2019) e0209958.

[330] Susana Nunes, Rita T. Sousa, Catia Pesquita, Multi-domain knowledge graph embeddings for gene-disease association prediction, J. Biomed. Semant. 14 (1) (2023) 11.

[331] Lorenzo Madeddu, Giovanni Stilo, Paola Velardi, Network-based methods for disease-gene prediction, arXiv preprint arXiv:1902.10117, 2019.

[332] Koushik Mallick, Sanghamitra Bandyopadhyay, Subhasis Chakraborty, Rounaq Choudhuri, Sayan Bose, Topo2vec: a novel node embedding generation based on network topology for link prediction, IEEE Trans. Comput. Soc. Syst. 6 (6) (2019) 1306–1317.

[333] Joana Vilela, Muhammad Asif, Ana Rita Marques, Joao Xavier Santos, Célia Rasga, Astrid Vicente, Hugo Martiniano, Biomedical knowledge graph embeddings for personalized medicine: predicting disease-gene associations, Expert Syst. 40 (5) (2023) e13181.

[334] Sai Narayanan, Akhilesh Ramachandran, Sathyanarayanan N. Aakur, Arunkumar Bagavathi, Genome sequence classification for animal diagnostics with graph representations and deep neural networks, arXiv preprint arXiv:2007.12791, 2020.

[335] Wenhuan Zeng, Anupam Gautam, Daniel H. Huson, Mulan-methyl—multiple transformer-based language models for accurate dna methylation prediction, GigaScience 12 (2023) giad054.

[336] Meng Yang, Lichao Huang, Haiping Huang, Hui Tang, Nan Zhang, Huanming Yang, Jihong Wu, Feng Mu, Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution, Nucleic Acids Res. 50 (14) (2022) e81.

[337] M. Saadat, A. Behjati, F. Zare-Mirakabad, S. Gharaghani, Drug-Target Binding Affinity Prediction Using Transformers, 2021.

[338] Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, Qiuyang Ding, Exploring evolution-aware &-free protein language models as protein function predictors, Adv. Neural Inf. Process. Syst. 35 (2022) 38873–38884.

[339] Syed Muazzam Ali Shah, Yu-Yen Ou, Disto-trp: an approach for identifying transient receptor potential (trp) channels using structural information generated by alphafold, Gene 871 (2023) 147435.

[340] Jing Wang, Sheng Chen, Qianmu Yuan, Jianwen Chen, Danping Li, Lei Wang, Yuedong Yang, Predicting the effects of mutations on protein solubility using graph convolution network and protein language model representation, J. Comput. Chem. 45 (8) (2024) 436–445.

[341] Xiaoyang Hou, Yu Wang, Dongbo Bu, Yaojun Wang, Shiwei Sun, Emngly: predicting n-linked glycosylation sites using the language models for feature extraction, Bioinformatics 39 (11) (2023) btad650.

[342] Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Sumit Tarafder, Debswapna Bhattacharya, Equipnas: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks, Nucleic Acids Res. 52 (5) (2024) e27.

[343] Jun Ma, Zhili Zhao, Tongfeng Li, Yunwu Liu, Jun Ma, Ruisheng Zhang, Graphsformercpi: graph transformer for compound–protein interaction prediction, Interdiscip. Sci. Comput. Life Sci. (2024) 1–17.

[344] Weizhi An, Yuzhi Guo, Yatao Bian, Hehuan Ma, Jinyu Yang, Chunyuan Li, Junzhou Huang, Modna: motif-oriented pre-training for dna language model, in: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2022, pp. 1–5.

[345] Jiani Ma, Jiangning Song, Neil D. Young, Bill C.H. Chang, Pasi K. Korhonen, Tulio L. Campos, Hui Liu, Robin B. Gasser, 'bingo'—a large language model- and graph neural network-based workflow for the prediction of essential genes from protein data, Brief. Bioinform. 25 (1) (2024) bbad472.

[346] Shijie Xu, Akira Onoda, Accurate and fast prediction of intrinsically disordered protein by multiple protein language models and ensemble learning, J. Chem. Inf. Model. 64 (7) (2023) 2901–2911.

[347] Yang Li, Zihou Guo, Keqi Wang, Xin Gao, Guohua Wang, End-to-end interpretable disease–gene association prediction, Brief. Bioinform. 24 (3) (2023) bbad118.

[348] Igor Melnyk, Vijil Chenthamarakshan, Pin-Yu Chen, Payel Das, Amit Dhurandhar, Inkit Padhi, Devleena Das, Reprogramming pretrained language models for antibody sequence infilling, in: International Conference on Machine Learning, PMLR, 2023, pp. 24398–24419.

[349] Mark Lennox, Neil Robertson, Barry Devereux, Modelling drug-target binding affinity using a bert based graph neural network, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2021, pp. 4348–4353.

[350] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, Burkhard Rost, Ankh: optimized protein language model unlocks general-purpose modelling, arXiv preprint arXiv:2301.06568, 2023.

[351] Florian Haselbeck, Maura John, Yuqi Zhang, Jonathan Pirnay, Juan Pablo Fuenzalida-Werner, Rubén D. Costa, Dominik G. Grimm, Superior protein thermophilicity prediction with protein language model embeddings, NAR Genomics Bioinform. 5 (4) (2023) lqad087.

[352] Qianmu Yuan, Chong Tian, Yidong Song, Peihua Ou, Mingming Zhu, Huiying Zhao, Yuedong Yang, Gpsfun: geometry-aware protein sequence function predictions with language models, Nucleic Acids Res. (2024) gkae381.

[353] Jim Clauwaert, Willem Waegeman, Novel transformer networks for improved sequence labeling in genomics, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (1) (2020) 97–106.

[354] Faiza Mehmood, Shazia Arshad, Muhammad Shoaib, Adh-enhancer: an attention-based deep hybrid framework for enhancer identification and strength prediction, Brief. Bioinform. 25 (2) (2024) bbae030.

[355] Anthony Martin Navarez, Robert Roxas, An evaluation of multitask transfer learning methods in identifying 6ma and 5mc methylation sites of rice and maize, Available at SSRN 4178244, https://www.doi.10.2139/ssrn.4178244.

[356] Kanchan Jha, Sriparna Saha, Sourav Karmakar, Prediction of protein-protein interactions using vision transformer and language model, IEEE/ACM Trans. Comput. Biol. Bioinform. 20 (5) (2023) 3215–3225.