RNA BIOLOGY
2023, VOL. 20, NO. 1, 384–397
https://doi.org/10.1080/15476286.2023.2223498

Taylor & Francis
Taylor & Francis Group

REVIEW

OPEN ACCESS Check for updates

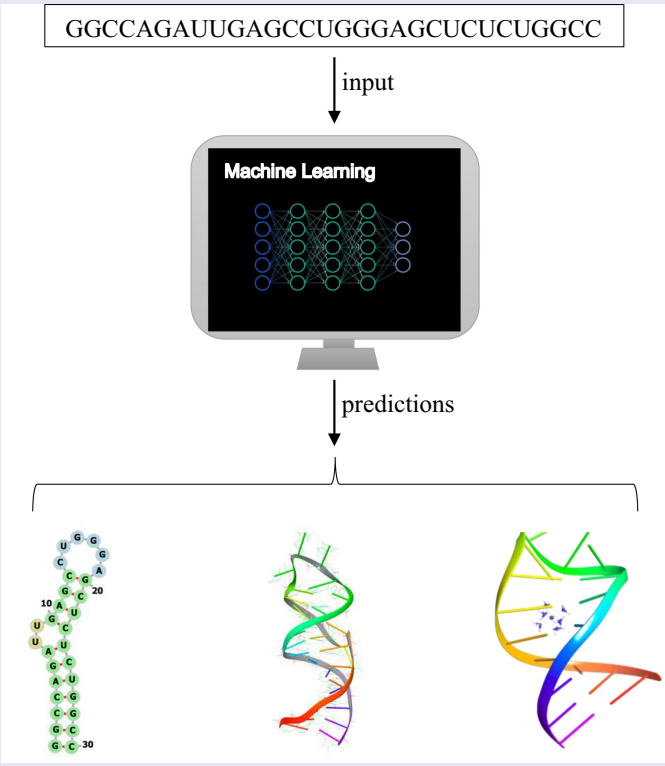
RNA-targeted small-molecule drug discoveries: a machine-learning perspective

Huan Xiao[a], Xin Yang[a], Yihao Zhang[a], Zongkang Zhang[a], Ge Zhang[b,c,d], and Bao-Ting Zhang[a]

[a]School of Chinese Medicine, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China; [b]Law Sau Fai Institute for Advancing Translational Medicine in Bone & Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, China; [c]Institute of Integrated Bioinformedicine and Translational Science, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, China; [d]Institute of Precision Medicine and Innovative Drug Discovery, HKBU Institute for Research and Continuing Education, Shenzhen, China

ABSTRACT

In the past two decades, machine learning (ML) has been extensively adopted in protein-targeted small molecule (SM) discovery. Once trained, ML models could exert their predicting abilities on large volumes of molecules within a short time. However, applying ML approaches to discover RNA-targeted SMs is still in its early stages. This is primarily because of the intrinsic structural instability of RNA molecules that impede the structure-based screening or designing of RNA-targeted SMs. Recently, with more studies revealing RNA structures and a growing number of RNA-targeted ligands being identified, it resulted in an increased interest in the field of drugging RNA. Undeniably, intracellular RNA is much more abundant than protein and, if successfully targeted, will be a major alternative target for therapeutics. Therefore, in this context, as well as under the premise of having RNA-related research data, ML-based methods can get involved in improving the speed of traditional experimental processes.



ARTICLE HISTORY
Revised 13 May 2023
Accepted 6 June 2023

KEYWORDS
RNA; microRNA; small molecule; machine learning; deep learning; drug discovery

Introduction

RNA molecules play important roles in biological processes ranging from genetic information transferring to gene expression regulating [1]. According to the Encyclopedia of DNA Elements (ENCODE) project [2], approximately 70–90% of the human genome is transcribed to RNA. Nevertheless, only

CONTACT Bao-Ting Zhang zhangbaoting@cuhk.edu.hk School of Chinese Medicine, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong 999077, China

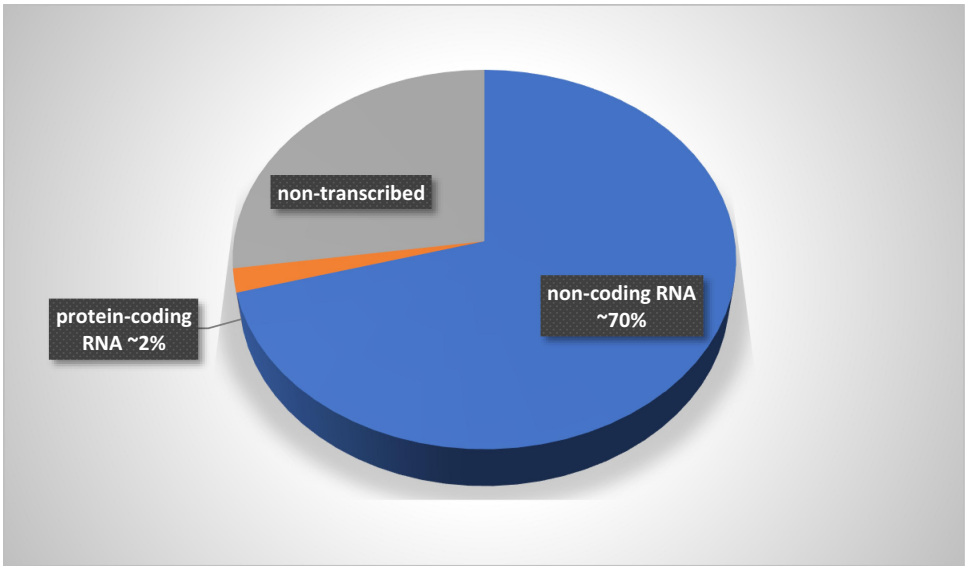


Figure 1. Only a small proportion (2%) of human genome is translated into proteins. More than 70% of human genome is transcribed into non-coding RNAs, which, if successfully targeted, would lead to an exponential increase in the number of drug discovery strategies.

Table 1. Therapeutic non-coding RNA targets.

Types	ncRNAs	Lengths
Housekeeping ncRNA	Transfer RNA (tRAN)	76–90
	Ribosomal RNA (rRNA)	>1500
Small ncRNA	MicroRNA (miRNA)	18–22
	Small interfering RNA (siRNA)	20–25
	Small nuclear RNA (snRNA)	100–300
	Small activating RNA (saRNA)	21
Long non-coding RNA	Long intergenic non-coding RNA (lincRNA)	~1000
	Circular RNA (circRNA)	100–999

2% of the genome encodes coding RNA (Figure 1), which means that the majority (~70%) of the genome yields non-coding RNA (Table 1) [3,4]. Screening or designing small molecules (SMs) targeting non-coding RNA is therapeutically promising, not only because the amount of non-coding RNAs in biological systems is considerable, but also because many of the disease-related proteins lack pocket-like motifs for SMs to bind to, leading to more than 85% of those proteins undruggable [5]. There would be a great potential to exponentially increase the number of therapeutical drug targets, even if a fraction of the tremendous RNAs were targetable [6]. For a long time in the past, RNA-targeted therapeutics were limited to oligonucleotide drugs that bind RNA by base-pairing [7–10]. For examples, the antisense oligonucleotides (ASOs)-based Nusinersen (marketed as Spinraza) was the first approved drug used in treating spinal muscular atrophy (SMA) [11]; Patisiran (marketed as Onpattro) [12] was the first small interfering RNA (siRNA)-based drug approved by U.S Food and Drug Administration (Table 2). However, the biggest obstacles to drugging RNAs with SMs were their structural flexibility and functional uncertainty [13], making it difficult to elucidate the structure of RNA molecules, let alone design SMs that can bind to them. In contrast, it is essentially always achievable to reach for a crystal structure of a protein, which enables structure-based drug design. Nowadays,

Table 2. FDA-approved oligonucleotide drugs.

Types	Drug Names	Market Names	FDA Approval Dates	Indications
ASO	Fomivirsen	Vitravene	1998	Cytomegalovirus retinitis in immunocompromised patients
	Mipomersen	Kynamro	2013	Homozygous familial hypercholesterolemia
	Nusinersen	Spinraza	2016	Spinal muscular atrophy
	Eteplirsen	Exondys 51	2016	Duchenne muscular dystrophy
	Defibrotide	Defitelio	2016	Veno-occlusive disease
	Inotersen	Tegsedi	2018	Hereditary transthyretin-mediated amyloidosis
	Golodirsen	Vyondys 53	2019	Duchenne muscular dystrophy
	Viltolarsen	Viltepso	2020	Duchenne muscular dystrophy
	Casimersen	Amondys 45	2021	Duchenne muscular dystrophy
Aptamer	Pegaptanib	Macugen	2004	Neovascular, Age-Related Macular Degeneration
siRNA	Patisiran	Onpattro	2018	Hereditary transthyretin-mediated amyloidosis
	Givosiran	Givlaari	2019	Acute hepatic porphyria
	Lumasiran	Oxlumo	2020	Primary hyperoxaluria type 1
	Inclisiran	Leqvio	2021	Primary hypercholesterolemia
	Vutrisiran	Amvuttra	2022	Hereditary transthyretin-mediated amyloidosis

with accumulated research and technological progress in RNA structural studies, targeting RNAs to inhibit disease processes gradually becomes possible [14–19]. The accordingly emergence of large amounts of genomic, molecular, and structural data also provides development opportunities for computational methods, especially when facing the urgent need of reducing production time and labour cost required for regular experimental methods.

Computer-aided drug design emerged in early 1970s [20], 3D quantitative structure-activity relationship (QSAR) was one of the most early applications in that era [21]. Late in 1990s, the virtual screening approaches (e.g. molecular docking, pharmacophore searching, similarity searching, and fingerprint searching) began to be introduced for the identification of bioactive molecules from large chemical databases [22]. However, unlike these traditional computational algorithms used in the above techniques, the nowadays machine learning (ML) algorithms (Table 3) don't process data through fixed calculation logic, but dynamically learn the intrinsic relationships among sample data so that the generated logic can be used to make predictions on new data [23]. Deep learning [24] is a specialized subset of ML. It has the capability of handling larger amounts of data or higher dimensions of features more quickly and accurately. Therefore, when we mentioned machine learning methods in this article, it also included deep learning methods. The ML technique first introduced to RNA structure prediction was the system proposed by Takefuji et al. [25] in 1990, which composed of several interactional neurons. Since then, and with advances in ML over the past 30 years, methods like expectation maximization [26], linear regression [27], support vector machine [28], convolutional neural network (CNN) [29], long short-term memory (LSTM) [30], etc. were gradually applied to RNA-related research. Especially in the past five years, with the breakthrough and success of deep learning, new ML methods for RNA structure and targeting prediction have sprung up like mushrooms after rain [31]. Although still in its infancy, ML-based RNA drug discovery and development is off to a prosperous start.

In this review, we discussed several aspects of RNA-targeted SM drug discovery, in which ML was involved: from targetable RNA structure prediction to RNA-SM interactions identification, and then to *de novo* RNA-targeted SM design. ML methods don't always participate in the core part of the above tasks. They may be applied at some critical point but make substantial changes to the original methods. We also provided suggestions and expectations in the Future Perspectives section, where the potential use of other ML models was discussed.

Targetable RNA structural motifs and predicting RNA structures using ML

RNA molecules generally fold into secondary (2D) and tertiary (3D) structures for molecular stability and biological functions. Some of the appropriate characteristic structures

Table 3. Commonly used machine learning algorithms.

Methods	Algorithms	Learning Styles
Regression	Decision Tree	Supervised
	Elastic Net Regression	
	Gaussian Process Regression	
	LASSO Regression	
	Linear Regression	
	Logistic Regression	
	Partial Least Squares Regression	
	Principal Component Regression	
	Random Forest	
	Support Vector Regression	
Classification	Bayesian Classifier	Supervised
	Conditional Random Fields	
	Discriminant Analysis	
	Gradient Boosting	
	K-Nearest Neighbor	
	Markov Random Fields	
	Naïve Bayes	
	Support Vector Machine	
Clustering	Gaussian Mixture	Unsupervised
	Hierarchical Clustering	
	Hidden Markov Model	
	K-Means	
	K-Medians	
Deep Learning	Autoencoders	Supervised/Unsupervised
	Boltzmann Machine	
	Convolutional Neural Network	
	Deep Belief Networks	
	Deep Neural Network	
	Generative Adversarial Network	
	Graph Convolutional Network	
	Graph Neural Network	
	Long Short-Term Memory	
	Multilayer Hopfield Neural Network	
	Multilayer Perceptron	
	Recurrent Neural Network	
	Self-Organizing Map	
	Transformer	

provide binding pockets for SMs to bind to, which enables the discovery and design of RNA-targeted therapeutic drugs [32]. The experimental methods of determining the RNA structural information mainly include selective 2'-hydroxyl acylation analysed by primer extension (SHAPE) [33], dimethylsulfate (DMS) [34], and light-activated structural examination of RNA (LASER) [35] for secondary structures; and x-ray crystallography [36] and cryogenic electron microscopy [37] for tertiary structures [38]. However, although accurate and reliable, these methods are generally time-consuming and expensive, especially in the face of the contemporary massive amount of RNA sequences. Therefore, it is inevitable to introduce computational techniques for resolving the structure prediction problem.

The single-stranded linear RNA molecule (primary structure) folds into itself to form double-stranded complementary regions through Watson-Crick base-pairings (A-U and G-C) and possible wobble base pairing (G-U), which constitute the skeleton of RNA secondary structure. These stacking regions of at least two consecutive base pairs are known as stems. While the mismatched single-stranded regions in the rest of the RNA secondary structure could form various structural motifs (certain kinds of loops) (Figure 2) depending on where these unpaired single strands are and whether they appear on one side or both sides of the stem. A hairpin loop is a single-stranded region that forms a half-ring structure at the end of a stem. A bulge loop occurs when a stem is interrupted by one or more unpaired nucleotides on one side. An internal loop occurs when a stem is interrupted by unpaired nucleotides on both sides, which form a ring structure. A multi-branch loop is formed when more than two stems are connected by a single strand. An exterior loop is formed when two ends of both sides of the stem cannot pair. A pseudoknot contains crossing-over links of the base pairs from at least two different stems, such that it folds into knot-shaped 3D conformation. Essentially, many of the motifs are potential pocket-like sites with molecular recognition properties.

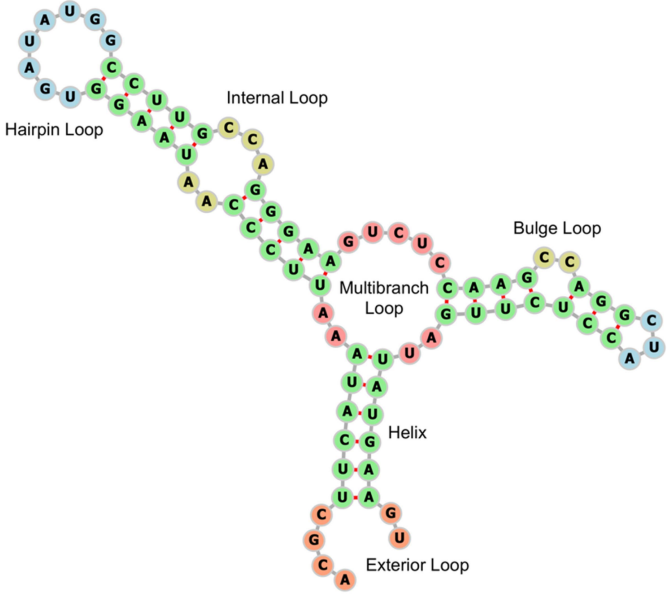


Figure 2. Common secondary structural motifs in RNA.

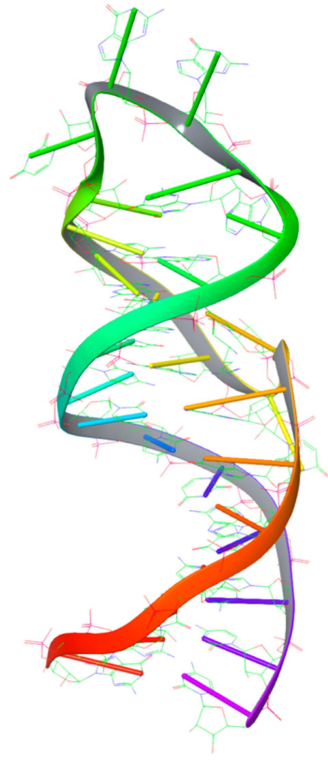


Figure 3. RNA tertiary structure example (PDB ID: 1AKX).

RNA secondary structure can further fold into its tertiary structure (three-dimensional/spatial shape) (Figure 3), in which the helical regions (base-pairings) and the unpaired regions (various loops) are organized in space via van der Waals contacts and specific hydrogen bonds [39]. The RNA tertiary structural motifs are more complex but can be generally categorized into three types: 1) interactions between two helical regions (e.g. coaxial stacking) [40]; 2) interactions between a helical region and an unpaired region (e.g. A-minor motif) [41]; 3) interactions between two unpaired regions [42]. In the following sections, we reviewed several ML-based methods for predicting RNA secondary and tertiary structures (Table 4).

ML-Based methods for predicting RNA secondary structures

The process of secondary structure prediction is essentially finding the most likely pairing scheme in a single-stranded RNA molecule that yields the lowest energy or forms the most robust structure. Compared with traditional bioinformatics techniques, e.g. minimizing the free energy of thermodynamic modes [43,44], dynamic programming algorithms [45–47], simulated annealing [48], statistical sampling [49,50], or entropy model [51,52], ML-based methods applied in RNA structure prediction have their inherent superiority in learning structural features from a large amount of data. According to specific processes ML models participate in and whether the models directly generate RNA structures, these methods can be further classified into two categories: direct secondary structure prediction and indirect secondary structure prediction.

Table 4. Summary of ML-based strategies for predicting RNA structures.

RNA Structures	Categories	ML Algorithms	Dates	References
2D	Direct Prediction	K-Nearest Neighbor	2006	[1]
		Hopfield Neural Network	2006	[2]
		Long Short-Term Memory (LSTM)	2018, 2019	[3–5]
		Ensemble Methods	2019, 2021	[6,7]
		Convolutional Neural Network (CNN)	2019	[8]
		Transformer	2020	[9]
	Indirect Prediction	Nearest Neighbor	2009	[10]
		Support Vector Machine	2013	[11]
		Ensemble Methods	2020	[12]
		LSTM	2020, 2021	[13–15]
3D	Direct Prediction	Bayesian Network	2009, 2011	[16,17]
		Conditional Random Fields	2011, 2013	[18,19]
		Markov Random Fields	2015	[20]
	Indirect Prediction	CNN	2018	[21]
		Multilayer Perceptron	2019	[22]
		Geometric Deep Learning	2021	[23]
		Graph Convolutional Network	2022	[24]

As the phrase 'direct' implies, this part of ML methods plays as core models in RNA secondary structure prediction and can directly generate corresponding structural data. They usually accept RNA primary sequences or sequence alignments as input and output RNA secondary structures. Some early works employed classic ML methods to compute the predictions. Proposed by Bindewald and Shapiro [53], an ML method that was composed of a hierarchical network of k-nearest neighbour classifiers (KNetFold) was presented to utilize RNA sequence alignment to predict whether any two nucleotides form a base pair or not. A consensus RNA secondary structure is then constructed after applying a set of rule-based filters to the output of the classifier. Around the same time, Liu et al. [54] computed a stable secondary structure of RNA by using a Hopfield neural network to find the maximum independent set in an adjacent graph mapped from all possible base pairs of an RNA sequence. Such methods have soon been overshadowed by the trend of deep learning in the recent decade, as one of the power models LSTM (Figure 4) is naturally better at dealing with sequential data and thus is extensively applied in this task. Wu et al. [55] proposed an LSTM model which consisted of Bidirectional LSTM (Bi-LSTM) (Figure 4) layers and fully connected layers. They transformed the problem of predicting RNA secondary structures into classifying base pairs in RNA primary sequences. Similarly, Lu et al. [56] constructed an adaptive LSTM with an energy-based filter that transforms RNA primary sequence to RNA secondary structure. Wang et al. [57] altered the output forms of the predictor so that they could use a sequence-to-sequence deep learning model for the problem, which uses Bi-LSTM as the encoder and fully connected layers as the decoder, to transform RNA sequences to dot-bracket sequences (The symbolic dot-bracket notation is a convenient way of representing RNA secondary structure). Singh et al. [30] proposed a pure ML approach SPOT-RNA to predict RNA secondary structure from a sequence in an end-to-end way. They ensembled five deep learning models for pre-training, each of which consists of ResNet blocks followed by a 2D Bi-LSTM layer and a fully connected block. And transfer learning, which handles the limited data problem, is then utilized for further training. The same group [58] later upgraded their model to SPOT-RNA2, which accepted evolutionary sequence profile and direct mutational coupling information in addition to RNA sequence as input and outcompeted SPOT-RNA significantly. The architecture of the deep learning model involved in SPOT-RNA2 was simplified by replacing LSTM with a dilated convolutional network, with a similar transfer learning approach applied after initial ensemble learning. Except for the successful application of LSTM, CNN (Figure 5) is also adopted for its advantages of processing spatial information. Zhang et al. [59] applied a CNN to predict the probabilities of RNA sequence base-

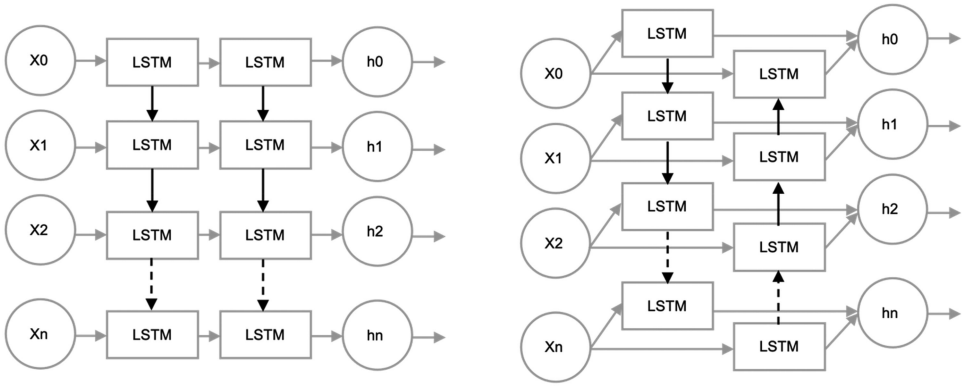


Figure 4. Network-based LSTM (left) and Bi-LSTM (right) architecture schematics.

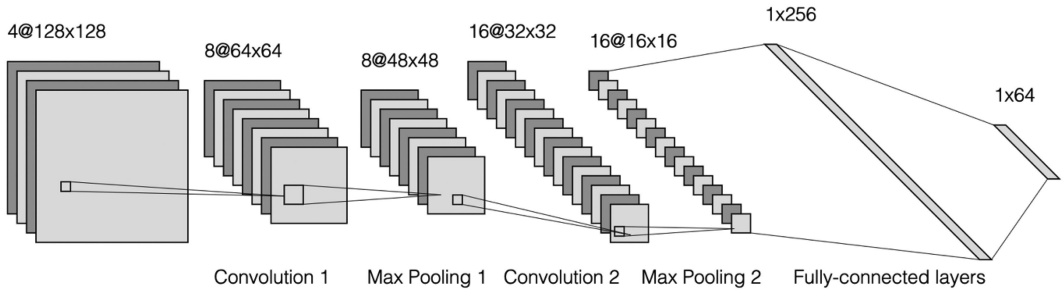


Figure 5. Data dimension-based CNN architecture schematic.

pairing, the structural predictions of which are further optimized through a dynamic programming-based correction algorithm. Chen et al. [60] proposed E2Efold, which combines a deep score network with a post-processing network to form an end-to-end deep learning model. The deep score network is built with transformer encoders and a 2D CNN, that encode RNA sequence information. The post-processing network is a multilayer network for optimizing output constraints.

For indirect RNA secondary structure prediction, ML methods are used in subprocesses like parameter estimation, scoring optimization, method selection, pre-processing, or post-processing that replace the original corresponding function schemes in the computational approaches [61]. For example, Calonaci et al. [62] trained an ensemble model of a convolutional layer and a double-layered network to predict penalties based on chemical probing data (DMS and SHAPE) and co-evolutionary data (DCA), which along with RNA sequences are further sent to RNAfold [63] to generate RNA secondary structures. Sato et al. [64] integrated convolution blocks, Bi-LSTM blocks, and multilayer perceptron blocks to calculate a folding score for each pair of nucleotides from an RNA sequence. Willmott et al. [65] improved the Nearest Neighbor Thermodynamic Model (NNTM) based RNA secondary structure prediction by incorporating synthetic SHAPE data [66] which is generated by state-predicting outputs from a Bi-LSTM. Hor et al. [28] proposed a tool based on support vector machines to preprocess RNA sequences before structure prediction. Quan et al. [67] applied Bi-LSTM to filter out the dubious base pairs derived from RNA secondary structural construction method (parallel ant colony optimization), in which the Bi-LSTM learns the base-pairing constraints for post-processing.

ML-Based methods for predicting RNA tertiary structures

Broadly speaking, predicting the 3D structure of RNA is a grand challenge, much more difficult than predicting its 2D structure. Because existing template 3D structures of RNAs are far from enough and the energy characteristics of stable 3D structures are not yet fully understood. These ML methods can also be divided into two categories: direct tertiary structure prediction and indirect tertiary structure prediction, depending on whether ML is involved in generating new 3D structures or assessing candidate 3D structures. Some successful cases from both categories are illustrated respectively below.

For direct tertiary structure prediction, most 3D structural results are obtained based on RNA secondary structures, primary sequences, or sequence alignments [68]. For example, Frellsen et al. [69] described a Bayesian network model of RNA using circular distributions and maximum likelihood estimation (BARNACLE) to construct RNA conformers, which combines a dynamic Bayesian network with directional statistics to capture the dihedral angles and their local dependencies in RNA fragments. It takes secondary structures to run and output local 3D RNA structures in continuous space. Wang et al. [70] introduced a method called TreeFolder, which firstly used a linear chain conditional random fields model to estimate the probability of an RNA conformation from both its primary sequence and secondary structure, then applied a tree-guided scheme for conformation sampling. An energy function is used to select acceptable conformation with lower energy. In addition to sampling techniques in RNA conformational space, some other approaches identified 3D RNA local modules in predefined patterns. A computational tool named RNA 3D modules detection (RMDetect) [71] proposed by Cruz et al. built a Bayesian network for modelling RNA modules in which the nodes represent individual bases and the edges represent the deficiencies between them. Four RNA modules including the G-bulge loop, kink-turn, C-loop, and tandem-GA loop can be detected from single sequences or multiple sequence alignments. The metaRNAmodules pipeline [72] further extended RMDetect with the 'RNA families' (Rfam) database alignments mapped from putative modules extracted from the 'Find RNA 3D' (FR3D) database in an automatic way. Zirbel [73] proposed hybrid Stochastic Context-Free Grammars (SCFG) and Markov Random Fields (MRF) models for the development of the software 'Java-based Alignment of RNA using 3D structure information' (JAR3D) to score RNA hairpin and internal loops against motif groups from RNA 3D Motif Atlas. Secondary sequences are aligned to the probabilistic SCFG/MRF models to check their ability to form the same pattern observed in 3D structural motifs, in which SCFG models nested pairs and insertions while MRF models crossed interactions and base triples.

For indirect tertiary structure prediction, ML techniques are generally used in structural scoring or quality evaluation that leads to the optimal 3D structure selection. Most existing works utilize the multi-layers of deep neural networks to handle the complexity of RNA 3D structural features. For example, Li et al. [74] presented a 3D CNN-based approach

Table 5. Summary of ML-based strategies for predicting RNA-SM interactions.

Categories	ML Algorithms	Dates	References
Machine Learning	Naïve Bayes	2012, 2020	[25–27]
	Random Forest	2012, 2019	[25,28,29]
	Support Vector Machine	2019	[30]
	Regularized Least Squares	2020	[31]
	Restricted Boltzmann Machine	2020	[32]
	K-Nearest Neighbor	2021	[33]
	Multiple Linear Regression	2022	[34]
	LASSO Regression	2023	[35]
Deep Learning	Graph Convolutional Network	2020	[36]
	Convolutional Neural Network	2021, 2022	[37,38]
	Multilayer Perceptron	2021	[33]

(RNA3DCNN), which uses the 3D image of a nucleotide and its surrounding environment to generate a nucleotide unfitness score indicating how poorly a nucleotide fits into the environment. Wang et al. [75] provided a scoring function to score RNA tertiary structure candidates using two multi-layer neural networks, which take the coarse-grained RNA structural and all-atom structural features as inputs respectively. And more recently, the Atomic Rotationally Equivariant Scorer (ARES) proposed by Townshend et al. [76] generated the best results in community-wide blind RNA structure prediction challenges. ARES is a geometric deep learning model consisting of many processing layers, with the first layer taking the 3D coordinates and chemical elements type of each atom as input and outputting the predicted RMSD of the structural model to form the true structure of the RNA molecule. Deng et al. [77] developed a computational model (RNAGCN) based on graph convolutional network (GCN) to predict RNA 3D structures, which extract features from graph representations of RNA local nucleotide environments with nodes modelling atoms and edges modelling their spatial positions to output scalar scores indicating the quality of the inputs.

Predicting RNA-SM interactions using ML

The determination of RNA structure is only halfway to the accomplishment of RNA-targeted SM drug discovery. Another crucial part is to find high-affinity SMs that could interact with the RNA. The developing process is like protein-targeted SM discovery to a certain degree – namely, both procedures require the identification of applicable pocket-like structures for SM binding. For instance, if we have obtained the RNA and SM structures, the intuitive way of finding RNA-SM binding pairs is by conducting virtual screening based on molecular docking. However, as most existing methods or tools [78–82] are originally developed and parameterized for protein targets, it would be inappropriate or invalid to directly apply them to RNA-SM systems. To overcome this shortage, some ML-based programs are proposed to score RNA-SM complex structures generated from a few available but not finely designed computational RNA docking methods. These ML-based programs don't create new RNA-SM pairs but bridge the gap between RNA structural generation and RNA-SM interaction identification. In addition to the physics-driven ways of screening RNA-SM interactions, some other ML methods take a step further. They implement a data-driven way of predicting novel RNA-SM interactions by training the ML model with existing RNA-SM associations and using the model to predict potential relationships between any pair of RNA and SM. These ML methods are usually standalone programs that directly generate binding SMs for an RNA target, which would be the mainstream development direction of ML for RNA-targeted SM discovery in the future (Table 5).

ML-Based scoring functions for evaluating RNA-SM complex structure

Some virtual screening tools like Ribodock and rDock [83,84] modified the rules of protein-SM interactions to be applied in nucleic acid-SM interactions. However, the universal problem for all related platforms is that the insufficient RNA-SM complex structures provide limited resources for judging the correct binding poses, thus the results they produce are more or less with inadequate accuracy. To address this problem, some studies used ML to apply additional criteria for evaluating the docking poses, as ML can pre-learn the complex relationships in limited structural data to distinguish between good and bad poses in future data. Chhabra et al. [85] built RNAPosers, which employed a random forest classifier to use pose fingerprint as input and output a classification score indicating the likelihood of a pose being native-like. The novel pose fingerprints were obtained by calculating distances between each SM's heavy atoms and its surrounding RNA's heavy atoms from the atomic coordinates of the poses. By taking structural-derived information into consideration, the ML-based pose classifiers outcompeted docking scores (physics-based free energy calculation) of discriminating high-confidence RNA-SM poses from decoy ones. Another successful ML method that achieved high accuracy in pose scoring is AnnapuRNA [86], a statistical scoring model designed to evaluate RNA-SM complex structures generated by any computational docking program. A coarse-grained representation of both RNA and SM was established for extracting descriptors describing their geometric relationships between the interacting partners, which were inputs for two selected supervised ML models for predicting: K Nearest Neighbors and multi-layer feedforward artificial neural network. In general, since these methods are highly dependent on the quantity and quality of determined RNA-SM complexes, the need for new structural data would be the prerequisite for algorithm improvement.

ML-Based strategies for predicting RNA-SM interactions

In most cases, when designing an ML model for predicting RNA-targeted SMs directly, the target RNA is restricted to a specific type, i.e. microRNA (miRNA). The main reason is that among various ncRNAs, miRNA is the most widely studied and now well-validated therapeutic target for SM [1,87] while coding RNAs like messenger RNA (mRNA) has always been the therapeutic target for complementary oligonucleotides rather than SM [88]. These ML models also require existing miRNA-SM association data for training, but the data do not have to be structured as that for scoring functions, because most miRNA-SM associations are established through pathway analysis without crystal structures. Several databases like SM2miR [89], Psmir [90], miRNet [91] can provide thousands of miRNA-SM association data for downloading. Based on the data, Shen et al. [92] proposed a multi-view joint learning-based computational framework (SMAJL) to predict novel miRNA-SM interactions, which exhibited superior accuracy in validation. SMAJL incorporates SM chemical and structural features, miRNA secondary structural features, and network features into the joint learning model built on a Restricted Boltzmann Machine to make predictions on association scores. Zhuo et al. [93] used a classic random forest classifier to screen potential high-affinity SMs targeting the miRNA-mRNA secondary structural motifs (loops), the results (Figure 6) of which were later experimentally validated both in vitro and in vivo. The two studies applied similar feature engineering methods to transform the chemical and biological properties of SM and miRNA by calculating the features (e.g. fingerprint, base occurrence frequency, etc.) from their structures. This is a dominant way of feature generation, as similar SMs have similar structures and incline to bind to similar target miRNAs. However, some other works adopted quite different feature generation approaches by calculating the similarity among SMs and miRNAs as their features respectively. Wang et al. [94] proposed a calculation model of random forest-based small molecule–miRNA association prediction (RFSMMA), which utilized known miRNA-SM associations from the SM2miR database to predict potential miRNA-SM pairs. In their study, they integrated four types of similarity for SM and two types of similarity for miRNA and then combined similarities of SM and miRNA as feature vectors to define training samples. These similarities do not necessarily relate to structures but involve factors such as phenotype and gene functional consistency. Similarly, Zhao et al. [95] applied identical calculations to convert SMs and miRNAs into their numerical features. They presented a model called Symmetric Nonnegative Matrix Factorization for Small Molecule-MiRNA Association prediction (SNMFSMMA), in which they firstly used symmetric nonnegative matrix factorization (SymNMF) to perform matrix decomposition and recalculation on the integrated similarity matrix of SMs and miRNAs, and secondly implemented Kronecker regularized least squares (KronRLS) to get the scores. Despite the convenience of relying on existing databases for computing, a few works tend to extract the miRNA-SM relationships directly from literature, with the aim of acquiring uncollected data and detailed information. For example, Xie et al. [96] proposed a text mining method named EmDL (Extracting miRNA-Drug interactions from Literature), which is the pioneer to establish miRNA-SM drug interactions from sentences in the literature. Firstly, features are extracted by calculating the distances between miRNA and associated SM in describing sentences retrieved from MEDLINE and PubMed Central papers. Secondly, a support vector machine was utilized to decide whether a given pair of miRNA-SM is interactive. Above all studies finally had to transform their collected data into numerical sequential representations to feed the ML models. However, there is one particular work, RNAmigos [97], that gets rid of numerical features and treated SM binding pockets in RNA 3D structures as graphs, where the nucleotides are represented by vertices and the base-pairings are represented by multi-relational edges of a graph. Oliver et al. gathered the crystal structure data of RNA-SM complexes from the Protein Data Bank (PDB) [98], and modelled the RNA binding site structure in a graphical representation that they defined as Augmented Base Pairing Network (ABPN). A Relational Graph Convolutional Network (RGCN) which operates directly on graphs, is used as the core model to accept ABPN inputs and compute node embeddings. A pooling process is followed by aggregating node embeddings into graph-level presentations, which were then fed through a multi-layer perceptron to output SM fingerprints. RNAmigos was the first one to apply a 3D RNA landscape to inferring binding SMs and showed strong performance in compound screenings. Kozlovskii and Popov [99] demonstrated a structure-based 3D CNN BiteNet$_N$, which trained with ~ 2000 nucleic acid-SM complexes to detect SM binding sites in nucleic acid structures. Similarly, Wang et al. [100] also developed a deep CNN for RNA-SM binding sites prediction: RLBind. The prediction tool used full-length RNA sequential and structural features as global information, as well as neighbouring nucleotide sequential and structural features as local information, for the capture of both long-range and short-range interactions. Meanwhile, Yazdani et al. [101] demonstrated that machine learning algorithms applied to experimentally derived sets of RNA binders are a more powerful method to inform RNA-binding chemical space.

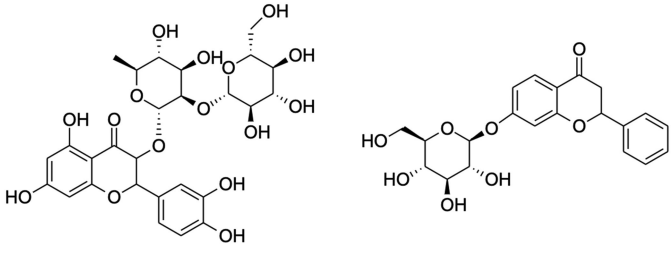


Figure 6. Chemical structures of OB – 4 (left) and OC – 3 (right).

Instead of considering as many as possible available miRNA-SM associations for training ML models to be able to handle arbitrary miRNA-SM pairs, some research works were target oriented and focused on a particular miRNA target to screen SM with practical significance. Jamal et al. [102] employed classic ML models (Naïve Bayes and Random

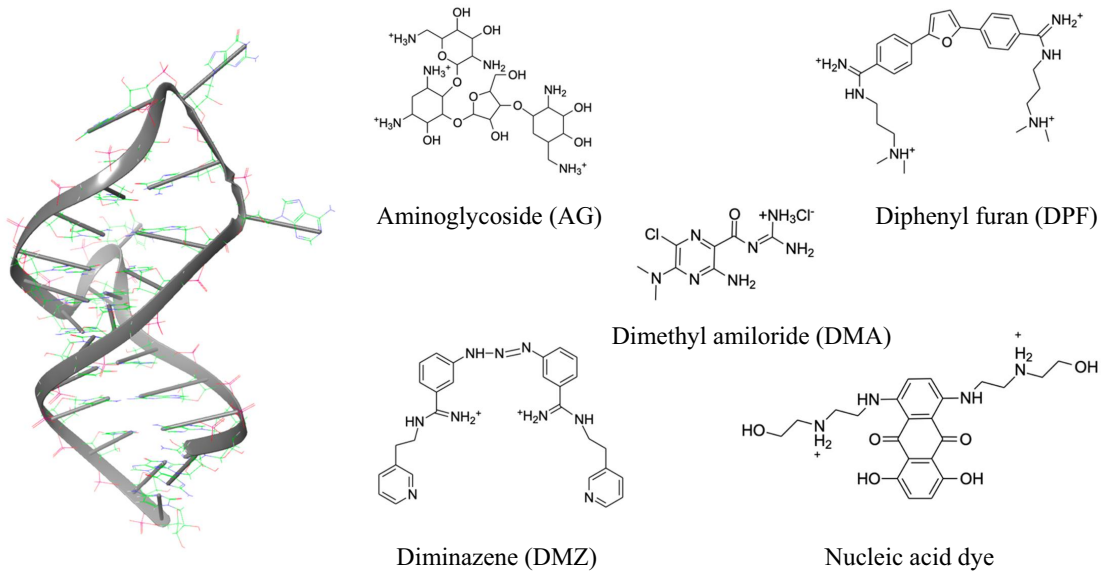


Figure 7. Crystal structure of HIV – 1 TAR RNA (PDB: 6×H0) and representative chemical structures of the scaffolds used in the work, i.e. Aminoglycoside (AG), Diphenyl furan (DPF), Dimethyl amiloride (DMA), Diminazene (DMZ), Nucleic acid dye.

Forest) to mine miR–21 inhibitors from large SM datasets with high accuracy and low false positivity. Cell-based screening for miR–21 inhibitors has already accumulated a great amount of active and inactive SMs suitable for ML, the model of which was later employed to screen approved drugs from the DrugBank database. Cai et al. [103] developed quantitative structure-activity relationship (QSAR) models to predict thermodynamic-based and kinetic-based parameters of SMs binding to a specific miRNA: HIV–1 transactivation response (TAR) element (Figure 7), in which multiple linear regression (MLR) was used to obtain linear combinations of the parameters that best fit the binding activities of SM against miRNA. The generalization ability of the model was further confirmed by comparison to ensemble tree methods.

Except for pure computational methods in identifying RNA-SM interactions, certain studies combined experimental and computational techniques, and incorporated ML methods in jobs like property determination that lead SM bias towards binding to RNA. The work of Rizvi et al. [104] is such a typical example. They used an Automated Ligand Identification System (ALIS), an affinity mass spectrometry to experimentally screen 42 RNA targets against a total of ~ 60,000 SMs. Next, naïve Bayes models were constructed for SMs selected in the preliminary screening using several descriptors such as physicochemical, topological, and biological properties. The most appropriate properties were then used to build a focused library of SM (~3700) that is enriched for binding RNA.

ML-Based strategies for De novo RNA-Targeted SM design

De novo SM design is one way of creating novel drugs with new structural and chemical properties that enhance the binding propensity to the targeted RNAs. Such design usually incorporates preferential modification into the basic molecular scaffold. Data-driven algorithms such as ML emerged as

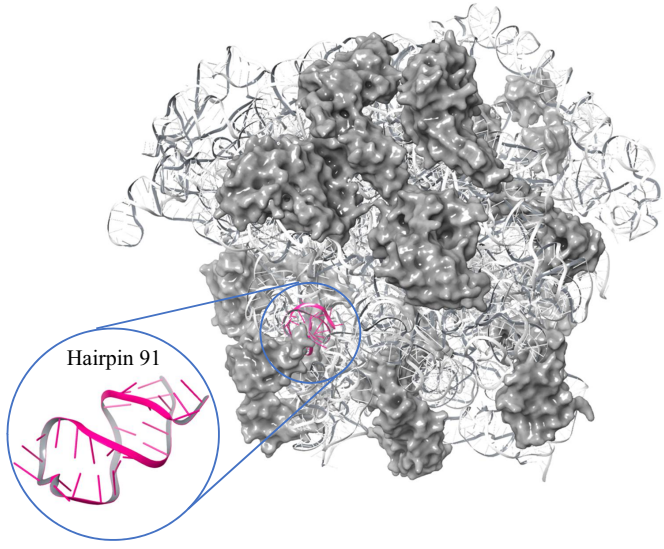


Figure 8. The complete structure of the Mycobacterium smegmatis 70S ribosome (PDB: 5O61). Hairpin 91 (pink) was used as a target for the design of small molecule inhibitors.

powerful tools that can reduce the search in chemical space for finding the chemical features that promote the RNA-SM interaction. Based on ~ 800 3D molecular structures with the 2-phenylthiazole moiety, as well as their corresponding binding values to the target hairpin 91 (Figure 8), Grimberg et al. [105] used three different ML approaches, i.e. Lasso regression, decision tree classifier, and CNN models, for discovering biosynthetic replacements facilitating the synthesis of phenylthiazole-containing molecules (Figure 9) that bind to RNA hairpin within the ribosomal peptidyl transferase centre (PTC) of *Mycobacterium tuberculosis*. Both chemical and geometrical/visual features of molecules were considered for three complementary models, and each of them revealed key

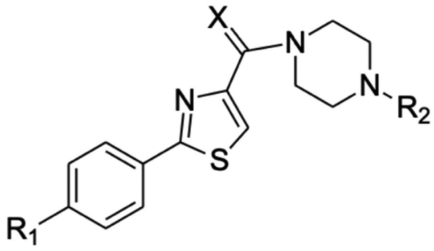


Figure 9. Chemical structure of the molecules synthesized and bio-evaluated as inhibitors for Mycobacterium smegmatis ribosomes.

features that influenced the binding affinity mostly from different perspectives, for example, higher N/NH and O/OH counts whereas lower TPSA (topological polar surface area) values, the number of N and C atoms must exceed a certain threshold, etc. After summarizing all the motifs as design principles for the synthesis of new heterocycles with improved binding abilities, they synthesized 10 SMs, of which 4 were validated to be potent inhibitors target hairpin 91.

Future perspectives

RNA-targeted SM drug discovery is a promising industry. The current ML technology has just entered this field and has shown its potential in solving this problem. We will discuss the other applicable ML models and some possible non-coding RNA targets.

Other ML models that can be used in RNA structure prediction

In addition to the most used models (LSTM and CNN) introduced in Section 2, which are applied to RNA secondary structure prediction, there are some other sequence models motivated by analysing sequential data that can be implemented for this task. Gated Recurrent Units (GRU) [106] is an alternative version of LSTM but has fewer parameters and can be faster to compute. Based on the similar architecture and comparable performance, GRU can be a good substitute for LSTM. Auto-encoders [107] and its variants [108] and improvements [109] can also be good models to encode sequential inputs to latent representations for further calculations. The most lately proposed Transformer [110] offers grand encoder-decoder architecture that uses attention mechanisms for information flow and is designed to process sequential input data. Employing alternative models with tailored alterations may promote prediction accuracy.

Other non-coding RNAs that can be targeted for ML drug discovery

Non-coding RNAs have a broad range, among which miRNAs, long non-coding RNAs, ASOs, siRNAs, transfer RNAs (tRNAs), and Ribosomal RNAs (rRNAs) have been experimentally identified as therapeutically target RNAs (Table 1) [87,111]. However, only miRNAs and rRNAs have been investigated for screening SMs by incorporating ML techniques. The reason why these ML methods cannot be universally applicable to different kinds of non-coding RNAs is that the properties (sequence length, structural characteristics, motif types, etc.) can vary a lot among different RNAs, thus it is difficult to encode very different RNA with a universal feature extracting method. So generally, one ML model deals with only one type of RNA. Another important factor is that the RNA-SM associations library has not been established for all kinds of non-coding RNAs. Those without enough existing published entries must first collect experimentally validated RNA-SM pairs for ML processing. However, if the local RNA motifs generated for SMs binding are similar, the existing RNA motifs database [112,113] can be easily applied to other non-coding RNAs like long non-coding RNAs. So ideally, all non-coding RNAs can be potential targets for ML drug discovery.

Conclusion

ML-based methods have already been widely applied in protein-targeted drug discovery, such as virtual screening [114], structure prediction [115], and *de novo* molecule design [116], etc. Due to the relative paucity of RNA structural data, there are fewer end-to-end SMs prediction algorithms for RNAs than for proteins. Rather, a larger part of current ML methods aims at constructing RNA secondary or tertiary structures [117]. The developmental trend is evident, after all, structural data is important in the subsequent process of drug discovery, no matter whether the method is *in silico* or not.

For the past few decades, RNA therapeutics development has mainly involved oligonucleotide drugs (e.g. antisense oligonucleotide (ASO), small interfering RNA (siRNA), mRNA, and Aptamer), which regulate target RNAs by base pairing or through RNA interference pathway [118–122]. However, oligonucleotide is not the only solution to this problem. Some medicinal chemists have already started to search for SMs drugging RNAs [123]. In 2020, FDA approved the SM drug Risdiplam (sold as Evrysdi), which was developed by Roche and PTC Therapeutics to treat spinal muscular atrophy (SMA) [124]. In the meantime, various traditional computation approaches also came to the fore [125–132]. With the advent and flourishing of artificial intelligence, the idea of developing ML-based methods for RNA-targeted small-molecule drug discovery becomes a modern topic. ML approaches benefit from their learning abilities to conclude embedded correlations from sample data and make predictions on large unknown data pools [133]. This merit makes ML methods promising as chemical space can be very large (~10^{60} molecules), and it is undesirable to derive the most possible results through manual experiments at a limited time and economic cost.

Traditional experimental way of drug discovery sticks in the dilemma of ambiguous RNA structures, and that challenge partially remains the same for ML techniques. Hence current ML is mainly involved in two parts of RNA drug discovery: RNA structure prediction and RNA-targeted SM prediction. The former produces intermediate results that can be utilized in subsequent experimental or computational research on finding binding SMs, while the latter produces final results that can be directly verified through experiments. The ML-involved studies of RNA structure

prediction are more abundant than that of RNA-targeted SM prediction at present. The reason is that ML techniques were introduced earlier in structural research, and the amount of existing suitable RNA-SM interactions limited the training of ML models. There are already credible programs for generating 2D RNA structures whose outputs can be further used for ML in predicting 3D RNA structures or RNA-targeted SMs. The end-to-end ML approaches of finding binding SMs that ignore the RNA 3D spatial coordinates structures are very common, they generally use chemical, biological, and 1-2D structural features extracted from RNA-SM associations to make predictions and can also achieve relatively good accuracy in the validation dataset. The produced results are valuable references for further experiments, as they largely narrow down the search space to find appropriate SMs. In this review, we summarized existing studies of RNA-targeted SM drug discovery where ML does its part. We believe that as more and more RNA data are being revealed, many new ML strategies will be proposed to address this issue.

Acknowledgement

We thank all of the authors who contributed to the studies reviewed in this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the National Key Research and Development Program of China [2018YFA0800802]; Hong Kong General Research Fund [CUHK 14103121, CUHK 14103420, CUHK 14108322, CUHK 14109721, HKBU 12114416, HKBU 12101117, HKBU 12100918, HKBU 12101018, HKBU 12103519, HKBU 14100218]; Theme-based Research Scheme [T12-201-20 R]; Guangdong Basic and Applied Basic Research Foundation [2019B1515120089]; Science and Technology Innovation Commission of Shenzhen Municipality Funds [JCYJ20160229210357960]; CUHK Direct Grant [2021.073]; Interdisciplinary Research Clusters Matching Scheme of Hong Kong Baptist University [RC-IRCs/17-18/02].

References

[1] Warner KD, Hajdin CE, Weeks KM. Principles for targeting RNA with drug-like small molecules. Nat Rev Drug Discov. 2018;17(8):547–558. doi: 10.1038/nrd.2018.93
[2] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
[3] Hangauer MJ, Vaughn IW, McManus MT, et al. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013;9(6):e1003569. doi: 10.1371/journal.pgen.1003569
[4] Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. Nature. 2012;489(7414):101–108. doi: 10.1038/nature11233
[5] Haga CL, Phinney DG. Strategies for targeting RNA with small molecule drugs. Expert Opin Drug Discov. 2022;18(2):1–13.
[6] Garner AL. Contemporary progress and opportunities in RNA-Targeted drug discovery. ACS Med Chem Lett. 2023;14(3):251–259. doi: 10.1021/acsmedchemlett.3c00020
[7] Rabie AM, Abdalla M. Forodesine and riboprine exhibit strong anti-SARS-CoV-2 repurposing potential: in Silico and in vitro studies. ACS Bio & Med Chem Au. 2022;2(6):565–585. doi: 10.1021/acsbiomedchemau.2c00039
[8] Eltayb WA, Abdalla M, Rabie AM. Novel investigational anti-SARS-CoV-2 agent ensitrelvir "S-217622": a very promising potential universal broad-spectrum antiviral at the therapeutic frontline of coronavirus species. ACS Omega. 2023;8(6):5234–5246. doi: 10.1021/acsomega.2c03881
[9] Rabie AM, Eltayb WA. Potent dual polymerase/exonuclease inhibitory activities of antioxidant aminothiadiazoles against the COVID-19 omicron virus: a promising in silico/in vitro repositioning research study. Heidelberger, Berlin: Mol Biotechnol; 2023.
[10] Crooke ST, Witztum JL, Bennett CF, et al. RNA-Targeted therapeutics. Cell Metab. 2018;27(4):714–739. doi: 10.1016/j.cmet.2018.03.004
[11] Corey DR. Nusinersen, an antisense oligonucleotide drug for spinal muscular atrophy. Nat Neurosci. 2017;20(4):497–499. doi: 10.1038/nn.4508
[12] Kristen AV, Ajroud-Driss S, Conceicao I, et al. Patisiran, an RNAi therapeutic for the treatment of hereditary transthyretin-mediated amyloidosis. Neurodegener Dis Manag. 2019;9(1):5–23. doi: 10.2217/nmt-2018-0033
[13] Manigrasso J, Marcia M, De Vivo M. Computer-aided design of RNA-targeted small molecules: a growing need in drug discovery. Chem. 2021;7(11):2965–2988. doi: 10.1016/j.chempr.2021.05.021
[14] Costales MG, Childs-Disney JL, Haniff HS, et al. How We think about targeting RNA with small molecules. J Med Chem. 2020;63(17):8880–8900. doi: 10.1021/acs.jmedchem.9b01927
[15] Childs-Disney JL, Yang X, Gibaut QMR, et al. Targeting RNA structures with small molecules. Nat Rev Drug Discov. 2022;21:736–762. doi: 10.1038/s41573-022-00521-4
[16] Seetin MG, Mathews DH. RNA structure prediction: an overview of methods. Methods Mol Biol. 2012;905:99–122.
[17] Weeks KM. Advances in RNA structure analysis by chemical probing. Curr Opin Struct Biol. 2010;20(3):295–304. doi: 10.1016/j.sbi.2010.04.001
[18] Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat Rev Genet. 2014;15(7):469–479. doi: 10.1038/nrg3681
[19] Wang X-W, Liu C-X, Chen L-L, et al. RNA structure probing uncovers RNA structure-dependent biological functions. Nat Chem Biol. 2021;17(7):755–766. doi: 10.1038/s41589-021-00805-7
[20] Ramaswamy A, Balasubramanian S, Rajagopalan M. Chapter 2 - biomolecular talks—part 1: a theoretical revisit on molecular modeling and docking approaches. In: Coumar MSeditor Molecular docking for computer-aided drug design.Academic Press; 2021. pp. 31–55. doi: 10.1016/B978-0-12-822312-3.00015-1
[21] Bibi S, Ul Islam F, Olawale OA, et al. Chapter 20 - in silico analysis for such natural compounds and COVID-19. In: Niaz KeditorApplication of natural products in SARS-CoV-2. Academic Press; 2023. pp. 463–489. doi: 10.1016/B978-0-323-95047-3.00019-8
[22] Del Rio A, Varchi G. Chapter 9 - molecular design of compounds targeting histone methyltransferases. *Epi-Informatics*. J.L. Medina-Franco Editor. Boston: Academic Press;pp. 257–272. 2016. doi: 10.1016/B978-0-12-802808-7.00009-5
[23] Greener JG, Kandathil SM, Moffat L, et al. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40–55. doi: 10.1038/s41580-021-00407-0
[24] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444. doi: 10.1038/nature14539
[25] Takefuji Y, Chen LL, Lee KC, et al. Parallel algorithms for finding a near-maximum independent set of a circle graph. IEEE Trans Neural Networks. 1990;1(3):263–267. doi: 10.1109/72.80251
[26] Sakakibara Y, Brown M, Hughey R, et al. Stochastic context-free grammers for tRNA modeling. Nucleic Acids Res. 1994;22(23):5112–5120. doi: 10.1093/nar/22.23.5112
[27] Xia T, SantaLucia J Jr., Burkard ME, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of

RNA duplexes with Watson–crick base pairs. Biochemistry. 1998;37(42):14719–14735. doi: 10.1021/bi9809425
[28] Hor CY, Yang CB, Chang CH, et al. A tool preference choice method for RNA secondary structure prediction by SVM with statistical tests. Vol. 9. Auckland, New Zealand: Evol Bioinform Online; 2013. pp. 163–184.
[29] Sample PJ, Wang B, Reid DW, et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol. 2019;37(7):803–809. doi: 10.1038/s41587-019-0164-5
[30] Singh J, Hanson J, Paliwal K, et al. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019;10(1):5407. doi: 10.1038/s41467-019-13395-9
[31] Lei X, Mudiyanselage TB, Zhang Y, et al. A comprehensive survey on computational methods of non-coding RNA and disease association prediction. Brief Bioinform. 2021;22(4):bbaa350. doi: 10.1093/bib/bbaa350
[32] Shao Y, Zhang QC. Targeting RNA structures in diseases with small molecules. Essays Biochem. 2020;64(6):955–966. doi: 10.1042/EBC20200011
[33] Wilkinson KA, Merino EJ, Weeks KM. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc. 2006;1(3):1610–1616. doi: 10.1038/nprot.2006.249
[34] Tijerina P, Mohr S, Russell R. DMS footprinting of structured RNAs and RNA-protein complexes. Nat Protoc. 2007;2(10):2608–2623. doi: 10.1038/nprot.2007.380
[35] Ackermann D, Famulok M. Pseudo-complementary PNA actuators as reversible switches in dynamic DNA nanotechnology. Nucleic Acids Res. 2013;41(8):4729–4739. doi:10.1093/nar/gkt121.
[36] Smyth MS, Martin JH. X ray crystallography. Mol Pathol. 2000;53(1):8–14. doi: 10.1136/mp.53.1.8
[37] Doerr A. Cryo-electron tomography. Nat Methods. 2017;14(1):34–34. doi: 10.1038/nmeth.4115
[38] Padroni G, Eubanks CS, Hargrove AE. Differentiation and classification of RNA motifs using small molecule-based pattern recognition. Methods Enzymol. 2019;623:101–130.
[39] Westhof E, Auffinger P. *RNA tertiary structure*, in *encyclopedia of analytical chemistry*. Hoboken, New Jersey, U.S: John Wiley & Sons; 2006.
[40] Quigley GJ, Rich A. Structural domains of transfer RNA molecules. Science. 1976;194(4267):796–806. doi: 10.1126/science.790568
[41] Nissen P, Ippolito JA, Ban N, et al. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. Proc Natl Acad Sci U S A. 2001;98(9):4899–4903. doi: 10.1073/pnas.081082398
[42] Batey RT, Rambo RP, Doudna JA. Tertiary Motifs in RNA structure and folding. Angew Chem Int Ed Engl. 1999;38(16):2326–2343. doi: 10.1002/(SICI)1521-3773(19990816)38:16<2326:AID-ANIE2326>3.0.CO;2-3
[43] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9(1):133–148. doi: 10.1093/nar/9.1.133
[44] Hofacker IL, Fontana W, Stadler PF, et al. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/Chem Monthly. 1994;125(2):167–188. doi: 10.1007/BF00818163
[45] Theis C, Janssen S, Giegerich R. 2010. Prediction of RNA secondary structure including kissing hairpin motifs. Algorithms bioinform. BerlinBerlin Heidelberg: Springer; pp. 52–64. doi: 10.1007/978-3-642-15294-8_5
[46] Reeder J, Steffen P, Giegerich R. pknotsRG: rNA pseudoknot folding including near-optimal structures and sliding windows. Nucleic Acids Res. 2007;35(suppl_2):W320–W324. doi: 10.1093/nar/gkm258
[47] Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots11Edited by I. Tinoco J Mol Biol. 1999;285(5):2053–2068. doi: 10.1006/jmbi.1998.2436
[48] Tsang HH, Wiese KC. SARNA-Predict: accuracy improvement of RNA secondary structure prediction using permutation-based simulated annealing. IEEE/ACM Trans Comput Biol Bioinform. 2010;7(4):727–740. doi: 10.1109/TCBB.2008.97
[49] Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. 2003;31(24):7280–7301. doi: 10.1093/nar/gkg938
[50] Ding Y, Chan CY, Lawrence CE. Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. 2004;32(suppl_2):W135–W141. doi: 10.1093/nar/gkh449
[51] Dawson W, Takai T, Ito N, et al. A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped? J Nucleic Acids Investig. 2014;5(1). doi: 10.4081/jnai.2014.2652
[52] Dawson WK, Fujiwara K, Kawai G, et al. Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS ONE. 2007;2(9):e905. doi: 10.1371/journal.pone.0000905
[53] Bindewald E, Shapiro BA. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA. 2006;12(3):342–352. doi: 10.1261/rna.2164906
[54] Liu Q, Ye X, Zhang Y *A hopfield neural network based algorithm for RNA secondary structure prediction*. In *First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06)*; 2006; Hangzhou, China.
[55] Wu H, Tang Y, Lu W, et al. 2018. RNA secondary structure prediction based on long short-term memory model. Intelligent computing theories and application. Cham: Springer International Publishing; pp. 595–599. doi: 10.1007/978-3-319-95930-6_59
[56] Lu W, Tang Y, Wu H, et al. Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter. BMC Bioinf. 2019;20(25):684. doi: 10.1186/s12859-019-3258-7
[57] Wang L, Liu Y, Zhong X, et al. Dmfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. Front Genet. 2019;10:10. doi: 10.3389/fgene.2019.00143
[58] Singh J, Paliwal K, Zhang T, et al. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. Bioinformatics. 2021;37(17):2589–2600. doi: 10.1093/bioinformatics/btab165
[59] Zhang H, Zhang C, Li Z., et al.Machine Learning Techniques on Gene Function Prediction. In Sangaiah, AK, editor . A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. Lausanne, Switzerland: Front Genet; 2019. p. 10.
[60] Chen X, Li Y, Umarov R, et al., *RNA secondary structure prediction by learning unrolled algorithms*. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia*, April 26-30, 2020, OpenReview.net.
[61] Zhao Q, Zhao Z, Fan X, et al. Review of machine learning methods for RNA secondary structure prediction. PLoS Comput Biol. 2021;17(8):e1009291. doi: 10.1371/journal.pcbi.1009291
[62] Calonaci N, Jones A, Cuturello F, et al. Machine learning a model for RNA structure prediction. NAR Genom Bioinform. 2020;2(4). doi: 10.1093/nargab/lqaa090
[63] RNAfold. 2022 October 18. Available from: http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
[64] Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun. 2021;12(1):941. doi: 10.1038/s41467-021-21194-4
[65] Willmott D, Murrugarra D, Ye Q. Improving RNA secondary structure prediction via state inference with deep recurrent neural networks. Comput Math Biophys. 2020;8(1):36–50. doi: 10.1515/cmb-2020-0002
[66] Deigan KE, Li TW, Mathews DH, et al. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A. 2009;106(1):97–102. doi: 10.1073/pnas.0806929106
[67] Quan L, Cai L, Chen Y, et al. Developing parallel ant colonies filtered by deep learned constrains for predicting RNA secondary

structure with pseudo-knots. Neurocomputing. 2020;384:104–114. doi: 10.1016/j.neucom.2019.12.041
[68] Huang B, Du Y, Zhang S, et al. Computational prediction of RNA tertiary structures using machine learning methods*. Chin Phys B. 2020;29(10):108704. doi: 10.1088/1674-1056/abb303
[69] Frellsen J, Moltke I, Thiim M, et al. A probabilistic model of RNA conformational space. PLoS Comput Biol. 2009;5(6):e1000406. doi: 10.1371/journal.pcbi.1000406
[70] Wang Z, Xu J. A conditional random fields method for RNA sequence–structure relationship modeling and conformation sampling. Bioinformatics. 2011;27(13):i102–i110. doi: 10.1093/bioinformatics/btr232
[71] Cruz JA, Westhof E. Sequence-based identification of 3D structural modules in RNA with RMDetect. Nat Methods. 2011;8(6):513–519. doi: 10.1038/nmeth.1603
[72] Theis C, Höner Zu Siederdissen C, Hofacker IL, et al. Automated identification of RNA 3D modules with discriminative power in RNA structural alignments. Nucleic Acids Res. 2013;41(22):9999–10009. doi: 10.1093/nar/gkt795
[73] Zirbel CL, Roll J, Sweeney BA, et al. Identifying novel sequence variants of RNA 3D motifs. Nucleic Acids Res. 2015;43(15):7504–7520. doi: 10.1093/nar/gkv651
[74] Li J, Zhu W, Wang J, et al. RNA3DCNN: local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. PLoS Comput Biol. 2018;14(11):e1006514. doi: 10.1371/journal.pcbi.1006514
[75] Wang YZ, Li J, Zhang S, et al. An RNA scoring function for tertiary structure prediction based on multi-layer neural networks. Mol Biol. 2019;53(1):118–126. doi: 10.1134/S0026893319010175
[76] Townshend RJL, Eismann S, Watkins AM, et al. Geometric deep learning of RNA structure. Science. 2021;373(6558):1047–1051. doi: 10.1126/science.abe5650
[77] Deng C, Tang Y, Zhang J, et al. RNAGCN: rNA tertiary structure assessment with a graph convolutional network. Chin Phys B. 2022;31(11):118702. doi: 10.1088/1674-1056/ac8ce3
[78] Comeau SR, Gatchell DW, Vajda S, et al. ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res. 2004;32(Web Server issue):W96–9. doi: 10.1093/nar/gkh354
[79] Schneidman-Duhovny D, Inbar Y, Nussinov R, et al. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 2005;33(Web Server issue):W363–7. doi: 10.1093/nar/gki481
[80] Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. Nucleic Acids Res. 2006;34(Web Server issue):W310–4. doi:10.1093/nar/gkl206
[81] Kozakov D, Brenke R, Comeau SR, et al. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins. 2006;65(2):392–406. doi: 10.1002/prot.21117
[82] Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res. 2008;36(Web Server issue):W233–8. doi: 10.1093/nar/gkn216
[83] Morley SD, Afshar M. Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. J Comput Aided Mol Des. 2004;18(3):189–208. doi: 10.1023/B:JCAM.0000035199.48747.1e
[84] Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, et al. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol. 2014;10(4):e1003571. doi: 10.1371/journal.pcbi.1003571
[85] Chhabra S, Xie J, Frank AT. Rnaposers: machine learning classifiers for ribonucleic acid-ligand poses. J Phys Chem B. 2020;124(22):4436–4445. doi: 10.1021/acs.jpcb.0c02322
[86] Stefaniak F, Bujnicki JM, Schlick T. AnnapuRNA: a scoring function for predicting RNA-small molecule binding poses. PLoS Comput Biol. 2021;17(2):e1008309. doi: 10.1371/journal.pcbi.1008309
[87] Winkle M, El-Daly SM, Fabbri M, et al. Noncoding RNA therapeutics - challenges and potential solutions. Nat Rev Drug Discov. 2021;20(8):629–651. doi: 10.1038/s41573-021-00219-z
[88] Matsui M, Corey DR. Non-coding RNAs as drug targets. Nat Rev Drug Discov. 2017;16(3):167–179. doi: 10.1038/nrd.2016.117
[89] Liu X, Wang S, Meng F, et al. Sm2mir: a database of the experimentally validated small molecules' effects on microRNA expression. Bioinformatics. 2013;29(3):409–411. doi: 10.1093/bioinformatics/bts698
[90] Meng F, Wang J, Dai E, et al. Psmir: a database of potential associations between small molecules and miRnas. Sci Rep. 2016;6(1):19264. doi: 10.1038/srep19264
[91] Chang L, Zhou G, Soufan O, et al. miRnet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res. 2020;48(W1):W244–W251. doi: 10.1093/nar/gkaa467
[92] Shen C, Luo J, Lai Z, et al. *Multiview joint learning-based method for identifying small-molecule-associated MiRNAs by Integrating pharmacological*, genomics, and network knowledge. J Chem Inf Model. 2020;60(8):4085–4097. doi: 10.1021/acs.jcim.0c00244
[93] Zhuo Z, Wan Y, Guan D, et al. A loop-based and AGO-Incorporated virtual screening model targeting AGO-Mediated miRNA-Mrna interactions for drug discovery to rescue bone phenotype in genetically modified mice. Adv Sci. 2020;7(13):1903451. doi: 10.1002/advs.201903451
[94] Wang CC, Chen X, Qu J, et al. RFSMMA: a new computational model to identify and prioritize potential small molecule-MiRNA associations. J Chem Inf Model. 2019;59(4):1668–1679. doi: 10.1021/acs.jcim.9b00129
[95] Zhao Y, Chen X, Yin J, et al. SNMFSMMA: using symmetric nonnegative matrix factorization and Kronecker regularized least squares to predict potential small molecule-microRNA association. RNA Biol. 2020;17(2):281–291. doi: 10.1080/15476286.2019.1694732
[96] Xie WB, Yan H, Zhao XM. EmDL: extracting miRNA-Drug Interactions from Literature. IEEE/ACM Trans Comput Biol Bioinform. 2019;16(5):1722–1728. doi: 10.1109/TCBB.2017.2723394
[97] Oliver C, Mallet V, Gendron RS, et al. Augmented base pairing networks encode RNA-small molecule binding preferences. Nucleic Acids Res. 2020;48(14):7690–7699. doi:10.1093/nar/gkaa583
[98] Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res. 2000;28(1):235–242. doi: 10.1093/nar/28.1.235
[99] Kozlovskii I, Popov P. Structure-based deep learning for binding site detection in nucleic acid macromolecules. NAR Genom Bioinform. 2021;3(4):lqab111. doi: 10.1093/nargab/lqab111
[100] Wang K, Zhou R, Wu Y, et al. Rlbind: a deep learning method to predict RNA–ligand binding sites. Brief Bioinform. 2022;24(1). doi: 10.1093/bib/bbac486
[101] Yazdani K, Jordan D, Yang M, et al. Machine learning informs RNA-Binding chemical space. Angew Chem Int Ed Engl. 2023;62(11):e202211358. doi: 10.1002/anie.202211358
[102] Jamal S, Periwal V, Open Source Drug Discovery C, et al. Computational analysis and predictive modeling of small molecule modulators of microRNA. J Cheminform. 2012;4(1):16. doi: 10.1186/1758-2946-4-16
[103] Cai Z, Zafferani M, Akande OM, et al. Quantitative structure-activity relationship (QSAR) study predicts small-molecule binding to RNA structure. J Med Chem. 2022;65(10):7262–7277. doi: 10.1021/acs.jmedchem.2c00254
[104] Rizvi NF, Santa Maria JP Jr., Nahvi A, et al. Targeting RNA with small molecules: identification of selective, RNA-Binding small molecules occupying drug-like chemical space. SLAS Discov. 2020;25(4):384–396. doi: 10.1177/2472555219885373
[105] Grimberg H, Tiwari VS, Tam B, et al. Machine learning approaches to optimize small-molecule inhibitors for RNA targeting. J Cheminform. 2022;14(1):4. doi: 10.1186/s13321-022-00583-x
[106] Cho K, Merrienboer Bv., Bahdanau D, and Bengio Y *On the properties of neural machine translation: encoder–decoder approaches in* SSST@EMNLP 2014.

[107] Cho K, Merrienboer BV, Gülçehre Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Doha, Qatar: Association for Computational Linguistics; 2014. p. 1724–1734.
[108] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst. 2014;2:27.
[109] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, CA, USA; 2015.
[110] Vaswani A, Shazeer N, Parmar N, et al., *Attention is all you need*. in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, Curran Associates Inc.: Long Beach, California, USA. p. 6000–6010.
[111] Falese JP, Donlic A, Hargrove AE. Targeting RNA with small molecules: from fundamental principles towards the clinic. Chem Soc Rev. 2021;50(4):2224–2243. doi: 10.1039/D0CS01261K
[112] Disney MD, Winkelsas AM, Velagapudi SP, et al. Inforna 2.0: a platform for the sequence-based design of small molecules targeting structured RNAs. ACS Chem Biol. 2016;11 (6):1720–1728. doi: 10.1021/acschembio.6b00001
[113] Sun S, Yang J, Zhang Z. Rnaligands: a database and web server for RNA-ligand interactions. RNA. 2022;28(2):115–122. doi: 10.1261/rna.078889.121
[114] Ghislat G, Rahman T, Ballester PJ. Recent progress on the prospective application of machine learning to structure-based virtual screening. Curr Opin Chem Biol. 2021;65:28–34. doi: 10.1016/j.cbpa.2021.04.009
[115] AlQuraishi M. Machine learning in protein structure prediction. Curr Opin Chem Biol. 2021;65:1–8. doi: 10.1016/j.cbpa.2021.04.005
[116] Mouchlis VD, Afantitis A, Serra A, et al. Advances in de novo drug design: from conventional to machine learning methods. Int J Mol Sci. 2021;22(4):1676. doi: 10.3390/ijms22041676
[117] Bagnolini G, Luu TB, Hargrove AE. Recognizing the power of machine learning and other computational methods to accelerate progress in small molecule targeting of RNA. RNA. 2023;29 (4):473–488. doi: 10.1261/rna.079497.122
[118] Zogg H, Singh R, Ro S. Current advances in RNA therapeutics for human diseases. Int J Mol Sci. 2022;23(5):2736. doi: 10.3390/ijms23052736
[119] Rabie AM. Two antioxidant 2,5-disubstituted-1,3,4-oxadiazoles (CoVitris2020 and ChloViD2020): successful repurposing against COVID-19 as the first potent multitarget anti-SARS-CoV-2 drugs. New J Chem. 2021;45(2):761–771. doi: 10.1039/D0NJ03708G
[120] Rabie AM. Teriflunomide: a possible effective drug for the comprehensive treatment of COVID-19. Current Res Pharmacol Drug Discov. 2021;2:100055. doi: 10.1016/j.crphar.2021.100055
[121] Rabie AM. Potent inhibitory activities of the adenosine analogue cordycepin on SARS-CoV-2 replication. ACS Omega. 2022;7 (3):2960–2969. doi: 10.1021/acsomega.1c05998
[122] Rabie AM, Abdalla M. Evaluation of a series of nucleoside analogs as effective anticoronaviral-2 drugs against the Omicron-B.1.1.529/BA.2 subvariant: a repurposing research study. Med Chem Res. 2023;32(2):326–341. doi: 10.1007/s00044-022-02970-3
[123] Koehn JT, Felder S, Weeks KM. Innovations in targeting RNA by fragment-based ligand discovery. Curr Opin Struct Biol. 2023;79:102550. doi: 10.1016/j.sbi.2023.102550
[124] Sheridan C. First small-molecule drug targeting RNA gains momentum. Nat Biotechnol. 2021;39(1):6–8. doi: 10.1038/s41587-020-00788-1
[125] Jiang W, Chen X, Liao M, et al. Identification of links between small molecules and miRnas in human cancers based on transcriptional responses. Sci Rep. 2012;2(1):282. doi: 10.1038/srep00282
[126] Wang J, Meng F, Dai E, et al. Identification of associations between small molecule drugs and miRnas based on functional similarity. Oncotarget. 2016;7(25):38658–38669. doi: 10.18632/oncotarget.9577.
[127] Qu J, Chen X, Sun YZ, et al. Inferring potential small molecule-miRNA association based on triple layer heterogeneous network. J Cheminform. 2018;10(1):30. doi: 10.1186/s13321-018-0284-9
[128] Lv Y, Wang S, Meng F, et al. Identifying novel associations between small molecules and miRnas based on integrated molecular networks. Bioinformatics. 2015;31(22):3638–3644. doi: 10.1093/bioinformatics/btv417
[129] Li J, Lei K, Wu Z, et al. Network-based identification of microRnas as potential pharmacogenomic biomarkers for anticancer drugs. Oncotarget. 2016;7(29):45584–45596. doi: 10.18632/oncotarget.10052
[130] Qu J, Chen X, Sun YZ, et al. In Silico prediction of small molecule-miRNA associations based on the HeteSim algorithm. Mol Ther Nucleic Acids. 2019;14:274–286. doi: 10.1016/j.omtn.2018.12.002
[131] Meng F, Dai E, Yu X, et al. Constructing and characterizing a bioactive small molecule and microRNA association network for Alzheimer's disease. J R Soc Interface. 2014;11(92):20131057. doi: 10.1098/rsif.2013.1057
[132] Guan NN, Sun YZ, Ming Z, et al. Prediction of potential small molecule-associated MicroRNAs using graphlet interaction. Front Pharmacol. 2018;9:1152. doi: 10.3389/fphar.2018.01152
[133] Lu M, Yin J, Zhu Q, et al. Artificial intelligence in pharmaceutical sciences. Eng. 2023. doi:10.1016/j.eng.2023.01.014