genes

MDPI

Perspective

Deep Learning for Elucidating Modifications to RNA—Status and Challenges Ahead

Sarah Rennie

Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark; sarah.rennie@bio.ku.dk

Abstract: RNA-binding proteins and chemical modifications to RNA play vital roles in the co- and post-transcriptional regulation of genes. In order to fully decipher their biological roles, it is an essential task to catalogue their precise target locations along with their preferred contexts and sequence-based determinants. Recently, deep learning approaches have significantly advanced in this field. These methods can predict the presence or absence of modification at specific genomic regions based on diverse features, particularly sequence and secondary structure, allowing us to decipher the highly non-linear sequence patterns and structures that underlie site preferences. This article provides an overview of how deep learning is being applied to this area, with a particular focus on the problem of mRNA-RBP binding, while also considering other types of chemical modification to RNA. It discusses how different types of model can handle sequence-based and/or secondary-structure-based inputs, the process of model training, including choice of negative regions and separating sets for testing and training, and offers recommendations for developing biologically relevant models. Finally, it highlights four key areas that are crucial for advancing the field.

Keywords: RNA-binding proteins (RBPs); post-transcriptional modifications; deep learning; neural networks; sequence motifs

Citation: Rennie, S. Deep Learning for Elucidating Modifications to RNA—Status and Challenges Ahead. *Genes* **2024**, *15*, 629. https://doi.org/10.3390/genes15050629

Academic Editor: Bowen Song

Received: 15 April 2024
Revised: 11 May 2024
Accepted: 11 May 2024
Published: 15 May 2024

1. Introduction

Modifications to RNA play pivotal roles in regulating gene expression, influencing a wide range of processes such as cellular differentiation, development, stress response and disease pathogenesis [1,2]. The term "epitranscriptomics" refers to the study of a broad range of modifications that influence RNA either co- or post-transcriptionally. Notable examples of modifications that play vital roles in regulating the fate of RNAs of protein-coding genes include RNA-binding proteins (RBPs) and a host of chemical modifications such as N6-methyladenosine (m^6A), m5-cytosine (m^6A), pseudouridine (ψ), and adenosine-to-inosine (A-to-I) RNA editing (Figure 1A). These modifications exhibit diverse cellular functions, including mRNA stability or degradation, splicing regulation, translation, and transport and localisation of mRNA targets (Figure 1B) [3,4]. In humans, it is estimated that there are over 1500 RBPs, which divide into a number of sub-families and possess a range of RNA-binding domains, including the RNA recognition motif (RRM), K-homology domain (KH), double-stranded RNA-binding domain (dsRBD), and zinc-finger domains. Furthermore, RBPs tend to localise to specific sub-cellular compartments, exhibit context-specific binding patterns, and often have highly specific sequence preferences [5]. The most well-studied chemical modification, m^6A, is typically found in the 3′ untranslated region (3′ UTR) and around the stop codon of modified transcripts [6], and its functions are mediated by a specific set of RBPs known as m^6A readers [7]. In contrast to m^6A, which is reversible, A-to-I RNA editing is an irreversible modification catalysed by the enzymatic activities of the RBP protein family Adenosine deaminases acting on RNA (ADAR), which act on double-stranded RNA (dsRNA) and play particularly important roles in self or non-self recognition in the regulation of immune response [8]. Whilst individual RNA modifications

have profound consequences on RNA function, they do not behave independently, and how they coordinate to exert their regulatory effects is highly complex and has significant implications for understanding gene regulatory control. Before this complexity can be fully deciphered, however, a crucial prerequisite is to determine which transcripts and precise positions are targeted by each modification and in what contexts, together with their sequence or structural preferences for binding.

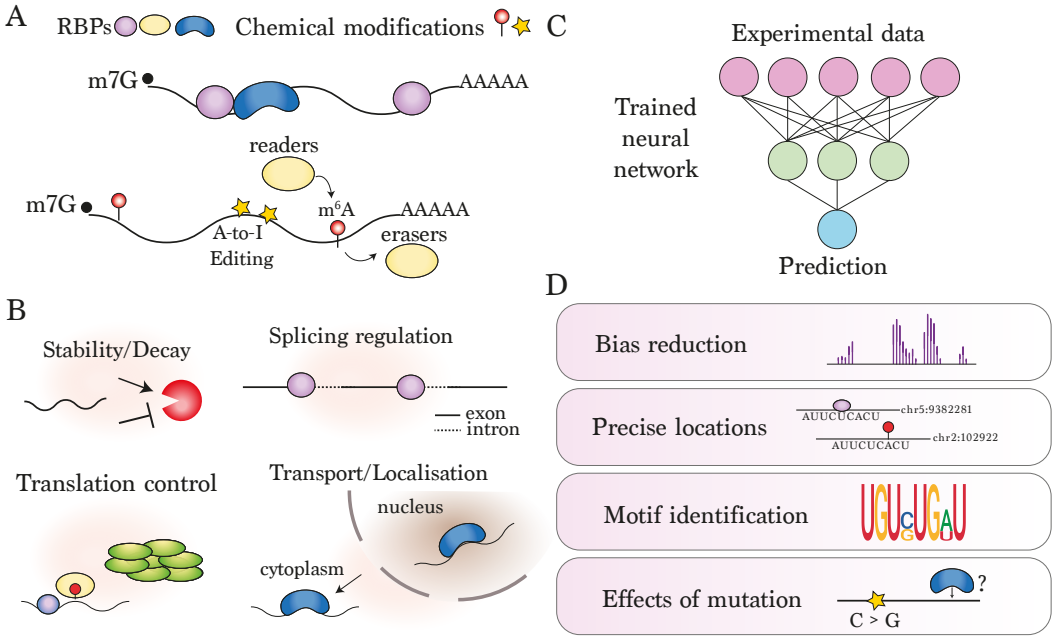

Figure 1. Introduction to deep learning for the prediction of modified sites on RNA: (**A**) RNAs are modified co- or post-transcriptionally by RNA-binding proteins (RBPs) and a range of chemical modifications. The study of these modifications transcriptome-wide is collectively termed epitranscriptomics. (**B**) Example roles of RNA modifications played by RNA-binding proteins and/or m^6A methylation. (**C**) Schematic depicting neural network, which typically starts with transcriptome-wide measurements of modified positions and makes some prediction based on the trained model. (**D**) Motivations for using deep learning approaches, including handling of bias, identification of precise binding locations, or model interpretation in terms of motif identification or assessing effects of sequence mutation.

Modified locations within the transcriptome can be detected experimentally, most commonly through the use of immunoprecipitation-based methods [9–11]. For the detection of RBP binding sites in vivo, numerous variations of the standard protocol, cross-linking followed by immunoprecipitation and sequencing (CLIP), exist. Notable examples are the high-resolution iCLIP (individual-nucleotide resolution CLIP) [12], and eCLIP (enhanced CLIP) [13], the latter of which has been extensively applied by the ENCODE consortium to profile 150 RBPs in the K562 and HepG2 cell lines, forming the largest resource of RBP binding to date [14]. However, CLIP-based methods are susceptible to several biases, including preferential binding to specific RNA sequences and the efficiency of cross-linking can vary significantly between different RBPs and RNA regions [15]. This creates challenges for the accurate determination of binding sites, potentially leading to both false positives and false negatives in the collected data. Recent advancements have led to the development of immunoprecipitation-free methods for detecting RBP protein binding, such as RNA-editor methods like HyperTRIBE and DART-seq [16,17]. These methods provide promising alternatives for modification detection, circumventing some of the biases of CLIP methods, although they do not guarantee the detection of the exact binding location. In addition to in vivo methods, in vitro approaches such as RNAcompete, RNA Bind-n-Seq, and SELEX (Systematic Evolution of Ligands by EXponential enrichment) facilitate the discovery of sequence motifs specifically recognised by RBPs [18,19]. Furthermore, various

advancements in technologies specifically for the detection of chemical modifications now allow for more precise determinants of modified bases and their associated levels [20,21]. The recent introduction of direct RNA sequencing by Oxford Nanopore Technologies also represents a significant step forward in the accurate detection of a variety of base modifications [22], with the potential for detecting multiple modifications within the same assay and at single-molecule resolution [23,24].

Deep learning, an advanced sub-field of machine learning, has revolutionised the analysis of genomic data in recent years [25]. Deep learning utilises deep neural networks, characterised by multiple layers incorporating potentially millions of 'weights', allowing well-defined models to learn highly complex patterns and representations from large-scale genomic datasets [26], Figure 1C). To facilitate model development, advancements in deep learning frameworks that efficiently utilise GPU capabilities have significantly enhanced the speed and performance of model training and testing, whilst also expanding the accessibility of deep learning approaches to a broader research community [27,28].

In the context of predicting RNA modification preferences, such as mRNA-RBP interactions, deep learning can serve several purposes (Figure 1D). First, it can address the high levels of noise in experimental datasets, which may be of low resolution or subject to a range of systemic biases [15]. In this regard, deep learning-based predictions can both refine the locations of observed sites and reduce the number of sites called as false positives in data processing pipelines. Second, well-trained models can extend predictions of modified locations to under-explored areas, especially where data is scarce, such as with lowly expressed RNAs, non-coding RNAs, or viral RNA [29,30], potentially leading to significant biological insights. Third, the learned feature spaces from trained models can be leveraged to uncover biologically relevant patterns, such as RNA-binding motifs [31,32], local cis-regulatory elements [33], or secondary structure features. They can also be used to address the impact of sequence variation in silico [34], such as single-nucleotide polymorphisms associated with specific diseases, thereby proving potential phenotypic insights to genetic connections. Additionally, in the context of nanopore direct RNA sequencing, several advanced deep learning approaches have recently been developed to interpret the signals captured as the RNA strand is pulled through a specialised pore, both for accurately determining the base type and modification status for each assayed molecule [23,24].

This perspective provides an overview of recent deep learning-based approaches developed to address where and how modifications select their target sites on mRNAs. Whilst the main focus is on the prediction of mRNA-RBP binding sites, we also highlight publications on other modifications, particularly m^6A and A-to-I editing. This work describes the different types of inputs typically used in training and how various types of layers can be used to connect these inputs to outputs. It discusses how models might be trained in a biologically meaningful way. Additionally, as deep learning methods and applications for RNA modifications are still in their infancy, future prospects in the field are explored, focusing on major challenges and opportunities for applying these techniques in the field.

2. Deep Learning for RNA Modifications

In order to decipher the complex regulatory roles of modifications to RNA, including their potential roles in disease, it is essential to be able to accurately pinpoint where and in which contexts these modifications occur. In recent years, numerous deep learning approaches have been published to this end, with some key examples summarised in Table 1. Whilst these approaches can vary greatly, Figure 2 outlines the most typical inputs, features and modelling strategies. First, 'positive' regions (e.g., known binding locations with some surrounding context) are usually selected for training (Figure 2A). Since deep learning models have a reputation for requiring substantial numbers of training examples, many studies focus on data from large-scale efforts like ENCODE [13,14], or from other well-documented studies in human cell-lines [35,36], often aided by databases such as POSTAR3 [37] or m6A-Atlas [38,39]. Note that many deep learning models de-

pend on proper processing of raw reads as a prerequisite. Rigorous pipelines exist for performing the necessary quality controls, conducting peak-calling and filtering regions also present in paired input samples, thereby generating a list of quality regions, or single-nucleotide positions in the case of some assays such as iCLIP [12], for use in further model training [40,41]. It is important to note, however, that even after peak-calling, certain biases, such as preferences towards certain sequence contexts, may still be present.

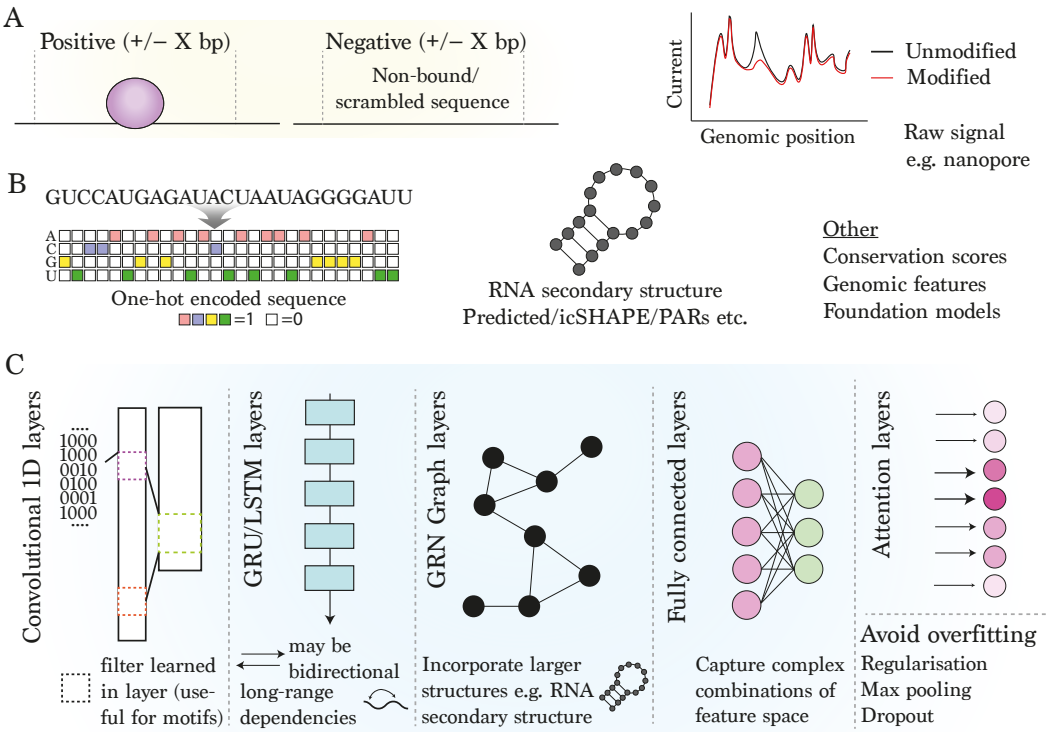

Figure 2. Strategies and outcomes for modelling RBP binding using deep learning techniques: (**A**) definition of positive and negative sites and inclusion of surrounding context. (**B**) Inputs: typical positive training examples include sequence and/or RNA secondary structure at and surrounding locations of known binding sites, together with negative regions of equal size without detected binding. (**C**) Modelling: Popular deep learning-based layers include single-dimensional (1D) convolutional layers for detecting local sequence motifs from the input data; gated recurrent unit (GRU) or long short-term memory (LSTM) layers for capturing long-range dependencies impacting RBP binding; graph (GRN) layers for capturing higher-order interactions between RNA nucleotides/structures; a fully connected layer, allowing the model to learn complex patterns; and attention, which focuses on important parts of the input data.

Table 1. Examples of some recent models for RNA modifications with emphasis on RNA-binding protein (RBP) binding. Note that this is not meant as an exhaustive list, but to cover a variety of currently available models. Abbreviations: RNASS: RNA secondary structure; CNN: convolutional neural network; RNN: recurrent neural network; SVM: support vector machine; GCN: graph convolutional network; MIL: multiple instance learning. For more detailed specifics of each model, please see cited references.

Name	Data	Description	Model	Ref
HDRNet	RBP	Sequence + in vivo RNASS (icSHAPE) + DNABERT [42], data shared with PrismNet [43], 101 bp regions, random assignment of positions in test/training.	Attention	[44]
BERT-RBP	RBP	Sequence + RNASS + DNABERT [42], 101 bp regions, data as per RBPsuite [45], random assignment of positions in test/training.	Attention	[46]

Table 1. *Cont.*

Name	Data	Description	Model	Ref
RBPnet	RBP	Sequence to signal mixture approach for bias correction, 300 bp windows, chromosome-wise splits to test/training.	CNN	[29]
DeepPN	RBP	Sequence + RNASS, bound-genes sourced negatives, 501 bp regions, random assignment of positions in test/training.	CNN/GCN	[47]
PrismNet	RBP	Sequence + in vivo RNASS, 101 bp regions, negatives with >40% icSHAPE coverage sampled from transcriptome, random assignment of positions in test/training.	CNN/attention	[43]
RNAProt	RBP	Multiple variable features, inc. sequence, RNASS, conservation, etc., 81 bp regions, random assignment of positions in test/training.	RNN	[48]
DeepCLIP	RBP	Sequence, matched-gene negatives for training, up to 75 bp regions, random assignment of positions in test/training.	CNN/RNN	[34]
DeepRiPe	RBP	Multitask models covering 59 RBPs, Sequence (150 bp regions derived from 50 bp bins) and genomic feature information (250 bp regions), random assignment of bins in test/training.	multi-output CNN	[49]
iM6A	m^6A	Sequence-based m^6A site prediction, surrounding unmethylated sites as negatives, human+mouse	CNN	[33]
m6Anet	m^6A	Trained on nanopore signal for molecule-resolution m^6A prediction.	NN/MIL	[23]
EditPredict	A-to-I	Sequence, predict A-to-I editing sites sourced from REDIportal [50], non-edited sites as negatives, up to 200 bp regions, multi-species.	CNN	[51]

To train models to distinguish between affected (e.g., bound by an RBP or methylated) and unaffected states, positive regions are often paired with similar numbers of so-called negative regions. Many recent approaches source these negative regions either from the same gene set to minimise gene selection bias or from general transcriptomic sampling (see [52] for detailed comparison of these two strategies). Some try to address systematic biases by treating regions as negative if they are called as sites bound by other RBPs that are not the RBP of interest [49]. Other methods avoid the need to specifically select negative regions, one example being iM6A [33], which, using an architecture similar to spliceAI for the detection of splice sites [53], aims to detect m^6A sites out from surrounding nucleotides.

2.1. Features and Model Architecture

The features to the neural networks can take various forms, but the two which are by far the most prevalent in the prediction of mRNA-RBP binding sites are sequence and RNA secondary structure (Figure 2B). Generally, these types of models are supervised, that is, the ground truth is considered as known and the model is tasked with connecting the supplied input features with these ground truth measurements (e.g., presence or absence of binding) via a set of non-linear functions incorporating weights which are optimised in the training process. For this reason, we focus on supervised approaches here, but other types of models are also extremely useful in biology, such as unsupervised learning of cell clusters from single-cell RNA-sequencing [54] or semi-supervised deep learning for biological imaging analysis [55], just to name two applications.

Due to its simplicity and flexibility, the RNA sequence extracted using the coordinates of the regions surrounding selected positions is frequently processed in a one-hot encoded format, before being presented as input to a neural network (Figure 2B). In this format, each possible base (A,U,C,G) in the RNA sequence is represented as a separate "channel", which is essentially a vector of the same length as the input sequence with value of one where the sequence encodes for that base and zero otherwise. A popular alternative representation is using *k*-mers, whereby the sequence is broken up into overlapping segments of length

k, before often being collapsed into a vector of counts for each possible sequence. On the whole, there is a trend towards using wider sequence contexts around positions of interest, with some even including full transcript sequences [33,56], although note that with larger models there may be a trade-off between the maximum input sequence length and the availability of GPU capacity for training. Interestingly, recent deep learning models for RBP appear to show that sequence alone can achieve high performance scores for determining the binding status for a large number of RBPs [29,34,43,57], the locations of m^6A sites [33], and A-to-I editing sites [51].

Figure 2C outlines some popular layer types frequently used in the model architecture of supervised neural networks. Encoded RNA sequence is usually managed using one-dimensional (1D) convolutional layers (termed a convolutional neural network, CNN). Briefly, the model learns a set of weights making up fixed-sized filters, which essentially scan across the input to the layer and assign scores. These scores are subsequently passed via an activation function to the next layer. The initial convolutional layer is especially seen as informative, as it directly connects to the encoded sequence input and can thus be interpreted in terms of de novo motifs or sequence contexts relevant for RBP binding [58]. Whilst CNNs excel at identifying local 'motif-like' sequence patterns, recurrent neural networks (RNN), specifically those with LSTM (long short-term memory) or GRU (gated recurrent unit) layers, are adept at learning long-term dependencies in sequence data [59–61] and bidirectional variations of LSTM can process sequences in both directions, enriching their ability to learn contextual patterns. RNNs have been applied with success in the context of RBP-mRNA interactions, two examples being iDeepE and DeepCLIP [34,57]. In addition, models involving transformer layers implement self-attention mechanisms by assigning variable attention weights to different positions, allowing them to handle all parts of the sequence at once. Transformers have shown promising potential in recently published approaches for RBP-mRNA interaction prediction [44,46,62].

Deep learning models are particularly flexible at combining different layer types, such as LSTM layers following convolutional layers. Fully connected layers, whereby all nodes connect to all nodes in the subsequent layer, are also common, and typically feed into the output layer. Fully connected layers allow for learning highly complex feature spaces, although these feature spaces can be difficult to interpret and these layers should be used sensibly with smaller datasets since they can vastly increase the number of trainable weights in the model. Layers are also interspersed with specialised types of layers such as activation layers, pooling or drop-out layers, which respectively pass features non-linearly between layers, reduce feature dimensions, or limit the number of parameters in the subsequent layer to avoid overfitting [63,64] (Figure 2C). Overfitting occurs when a model achieves a very high performance within the same data on which it is trained, but fails to generalise to new, unseen data, such as new genomic locations. Large models with few examples are especially prone to overfitting, and for this reason, it is important to assess model performance on only unseen data (see below). Overall, due to wide possibilities for complex arrangements and parameterisations, different models with similar feature sets can potentially perform very differently. For this reason, it is important that approaches are carefully optimised for the given problem and cannot be treated as 'black-box' approaches for machine learning.

2.2. Incorporating RNA Secondary Structure

Since the RNA-binding domains of RBPs vary in their ability to recognise and bind RNA secondary structures, RNA conformation can play an important determining role in the prediction of mRNA-RBP interactions and is therefore frequently considered as an input feature in models. GraphProt, based on support vector machines (SVMs), was one of the first models to extensively incorporate predicted secondary structure information, encoded as graph kernels [65] and has since been superseded by deep learning-based methods. Many of these models encode secondary structure features into a graph, where each node represents a nucleotide in the sequence and edges symbolise their interactions [47,66,67]. This

graph representation is processed via a graph convolutional network (GCN) (Figure 2C), which applies a series of convolution operations, aggregating features from neighbouring nodes by taking into account both its individual features and the structure of its immediate surroundings in the graph.

The majority of these models focus on predicted RNA secondary structures via the application of computational tools (e.g., RNAfold or RNAshapes [68,69]. These tools work by calculating the minimum free energy (MFE) structure from a given sequence based on sophisticated dynamic programming algorithms. As RNA structure can be stochastic within cells, one approach found it advantageous to consider base-pairing probabilities instead of a single MFE configuration [70]. Alternatively, PrismNet accommodates experimentally determined in-vivo secondary structures by utilising the experimental icSHAPE method [30,44], which provides a score at genomic positions representing double-strandedness or single-strandedness of the transcribed RNA [71]. Notably, these structures appear to remain stable across different cell types [72], yet their observed variations seem able to decipher dynamic, tissue-dependent mRNA-RBP binding [43]. A more recent model, HDRNet, takes this concept a step further by testing the capabilities of in-vivo secondary structures to predict dynamic RBP binding in a given cell context using a model which was trained on another context, with promising results [44]. In addition, the same study further supported the advantage of using in vivo structures over computational predictions, finding that models using in vivo structures always outperformed their counterparts based on RNAfold-predicted structures in place of the icSHAPE scores.

However, note that the inclusion of the RNA secondary structure and sequence in parallel does not always guarantee an improvement in performance over sequence alone, and is instead likely to be dependent on the underlying biology of the protein under study [57]. Interestingly, for certain RBPs such as PABPC4, METAP2, DDX55, and DGCR8, performance was actually found to be higher for a model using only structure-based features compared to sequence-only features [30]. However, the same study did show that, on average across all tested proteins, a combination of in vivo structure and sequence resulted in the best performance (an AUROC of 0.850 compared to 0.797 and 0.758 for sequence-only and structure-only models, respectively, where an AUROC of 1 implies a perfect model; see below for description of measures of model performance). Since performance variations are likely reflecting the underlying biology of the protein under study, a closer look into these statistics may provide useful clues of properties of binding behaviours of lesser known RBP groups. In addition, note that the role of secondary structure via deep learning for the prediction of other modifications such as RNA-editing and m^6A remains less explored. Whilst the ADAR proteins catalysing RNA-editing are known to have strong preferences for double-stranded RNA, m^6A sites typically appear associated with single-stranded RNA, although the precise nature of the relationship between m^6A and RNA structure does remain somewhat unclear [73].

Finally, note that models are not limited by only sequence and/or secondary structure features, but can flexibly be extended to incorporate further features: for example, RNAprot has shown benefits in including additional features such as sequence conservation [48], and, given RBPs have preferences for specific genomic features, including information of the locality of the bound region (e.g., 5′ UTR, CDS, 3′ UTR) also seems to help predictions [49]. Moreover, progress in the use of large language models for genomic sequence, such as bidirectional encoder representations from transformers (BERT), has accelerated recently [42] and appears advantageous for the prediction of RNA-RBP interactions [44,46]. For example, HDRNet adapts BERT to encode sequences of interest broken into short *k-mer* stretches as dynamic representations that appear efficient at capturing both local contexts and long-range dependencies, reflected in the impressive performance of their models [44].

2.3. Perspective on Current Models

Due to the limited number of large data resources over which modelling can be applied systematically on a broad scale, a number of the RBP models mentioned in Table 1 focus on

the same datasets (e.g., ENCODE [14] for RBPs in the HepG2 and K562 cell lines). Despite this, the overall approaches employed by the various models are often difficult to compare due to differences and/or ambiguities in model setups, such as variations in the choice of negative sets, region sizes, and chosen strategy for dividing examples between sets used for training and testing sets used for assessing model performance (for example, holding out individual positions vs. whole genes or chromosomes, see below for further discussion on this aspect) (see Table 1 for set-ups of individual models). On the other hand, despite the different modelling approaches, the performance of a given model may be more influenced by the characteristics of the protein under study than by specific model architectures or feature choices. For instance, AGO2, involved in RNA-binding functions via miRNAs [74], and ALKBH5, an eraser of N6-methyladenosine, often receive lower performance scores relative to other RBPs [34,67]. This could be attributed to the low dependency of that protein on only sequence-based features, although factors such as poor data quality or insufficient numbers of training sites to work with likely also play a significant role.

Moreover, it should be noted that trained models are not immune to biases present in the original data. For example, one study noted an inverse correlation between AUC performance and GC-richness in the RBP's sequence preference [34]. Efforts to mitigate sequence biases in immunoprecipitation-based data are ongoing [29,67,70]. For example, a promising recent method, RBPnet, integrates matched input signal in a mixture-model approach, allowing the neural network to separate 'true' signal from systematic biases that can be inferred from the input [29]. Another method, HPNet, attempts to mitigate the influence of systematic nucleotide bias in the data by employing 'context-averaging' [67]. Whilst these attempts are important for improving the quality of binding site predictions, it should be noted that the field's heavy reliance on eCLIP data underscores the need for further attention to this issue [14]. Incorporating orthogonal experimental information, such as data from the RNA-editor methods [16,17], could lead to more confident results, although such approaches have to date not yet been tested.

In addition to the above considerations, overfitting remains a significant challenge, caused by the scarcity of reliable input data and labels for training, particularly to RBPs with few binding sites, or rarer chemical modifications. One promising option to mitigate this is via transfer learning, whereby one trains on a broad set of modifications and then fine-tunes on individual modifications [75]. The use of recent foundation models for genomic sequence can help in this regard [42,76]. These large models are pre-trained over a broad range of genomes, and sequence embedding vectors can be extracted or models fine-tuned for use in specific problems, such as mRNA-RBP site prediction, thus aiding in circumventing issues with low dataset sizes [46].

2.4. Model Performance and Choice of Background Set: The Hunt for Biologically Relevant Results

It should be noted that due to competing approaches on similar data, there is often a pressure to demonstrate the best model performance in published works. Model performance can be measured in multiple ways. The most typical scores are accuracy (the proportion of correct predictions) and the area under the curve (AUC), which balances the sensitivity of the model to find true sites and the specificity to exclude those which are not true sites, with 1 being a perfect model and 0.5 suggesting the model guesses randomly. Alternatively, the area under the precision recall curve (AUPRC) balances precision, which is the proportion those sites predicted which are actually modified, and recall, which is the proportion of modified sites that are also predicted as such. AUPRC ranges between 0 and 1 and is usually preferred over AUC in situations where numbers of positives and negatives are unbalanced, as the AUC can be highly misleading in these situations. In any case, it is normally preferable to quote a range of statistics in order to comprehensibly describe and compare model performance.

Whilst one generally would expect to trust predictions derived from a well-performing model over a lesser-performing one, it is important to realise that there are situations where a well set-up model can show poorer performance according to these above metrics, yet

yield more biologically relevant results. One example is the use of stringent background sets, which may lead to poorer performance metrics but actually ensures that the learned feature space aligns with the biological problem and avoids capturing irrelevant information. To address the impact of choice of the negative set, a recent study systematically benchmarked a range of models in the context of RBP binding [52]. They considered two different background sets: one based on random sampling within the same genes as the positive positions, and one based on sampled positions that were experimentally-defined binding sites of other RBPs. The rationale of the second set was that by sampling positions of other RBPs to the one of interest, one can avoid learning features associated with potential experimental biases between the positive and negative set. Indeed, they found that performance dropped for this second set, suggesting that a portion of the measured performance of the less strict background may be capturing over-representation of experimental bias in positive sets rather than true biological signal.

2.5. Further Considerations for Modelling Approach

Figure 3 illustrates some suggestions for designing and training deep learning models to maximise biological insight. First, as mentioned, careful selection of negative site location is essential (Figure 3A). In addition, RBPs are not uniformly distributed across transcripts, but are often localised to highly specific regions, such as near splice sites or in the 3′ UTR [14], and m^6A tends to be enriched near the stop codon. Since sequence and/or secondary structure preferences are highly variable across transcripts, general sequence patterns relevant to the transcript region or the presence of these specific features might be over-represented in intervals around the positive sites. Similar to the above situation with experimental biases, this could result in the model appearing to perform well, but in a way that is capturing region preferences independent of what is relevant to the binding of the specific modification. To circumvent this, when constructing machine learning models, it is suggested to consider carefully matching background sites according to transcript features, and preferably on genes that display similar expression distributions and/or are targets for the RBP of interest [77].

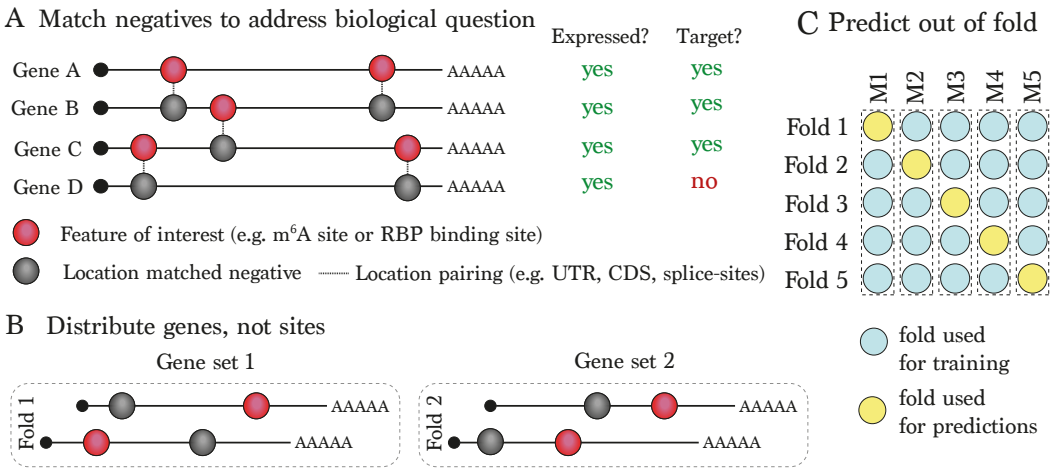

Figure 3. Recommendations for deep learning model training and prediction: (**A**) Choice of negative ('non-modified') locations can have a large influence on the biological applicability of the model. Suggested options could be to match modified positives according to feature location, expression and/or target status. (**B**) When performing cross-validation, it is recommended to avoid information leakage by distributing genes across the folds rather than individual sites, whose sequence could overlap a site in another fold. (**C**) In order to honestly assess the model, it is recommended to predict out of fold.

Second, when training models, it is common practice to employ cross-fold validation (Figure 3B,C). Here, one allocates positions to a number of folds, typically 5 or 10, and

trains the model on all the folds but one, using the held-out fold to assess model performance. Producing predictions for each held-out fold allows one to gain an overall picture of model performance, and since performance statistics are only calculated on sets not including within training, these statistics should not be inflated by potential overfitting of the model. Frequently, in order to allocate training examples to different folds, sites are simply randomised across these five folds, this being the case for a large number of the models highlighted above (see Table 1 for strategies employed for allocating folds by selected recent mRNA-RBP models). However, in the case of post-transcriptional modifications, there is a strong argument for exercising caution, since one single transcript region can harbour multiple modified sites, forming clusters. This means that highly overlapping input features (e.g., sequence or secondary structure) will occur across multiple different folds, potentially leading to inflated performance statistics. Thus, it is highly recommended to allocate folds in such a way that any given gene is only found within a single fold (Figure 3B). Many published deep learning models within the genomics field take this a step further, by holding out entire chromosomes in order to strictly ensure that there is no leakage between the training and testing sets, with one example from Table 1 being RBPnet [29]. In addition, SpliceAI takes an even stricter approach by holding back paralogs from the training set in order to addresses the issue of sequence similarity due to common gene ancestors [53], and it would be interesting to see how this approach affects the prediction scores of current mRNA-RBP interaction modelling approaches. Additionally, along similar lines of argument, when interpreting a model, it is also recommended to work with prediction scores or effects of in silico mutation using a model whereby the involved inputs were never seen within the training set (Figure 3C).

In conclusion, it is important for focus to shift towards biological interpretations beyond mere performance statistics when assessing a model. Lower performance scores might indicate that the problem is set up more stringently, controlling for more confounding variables. These models may have greater potential to provide interesting biological insights, and further efforts are required to fully explore these aspects.

3. Some Major Future Perspectives

Deep learning within genomics, including the field of RNA modifications and particularly the context of mRNA-RBP interactions, is gaining a lot of traction in recent years, with a lot of recently published approaches (not limited to those presented in Table 1). However, there are still a number of distinct challenges and opportunities to be taken into account, so that the field can move further forward in terms of biological discoveries. Four are briefly discussed below.

3.1. Generalisability across Cell Types and Species

Current experimental methods for studying RNA-binding proteins (RBPs) typically require substantial input material, leading to a heavy reliance on cell lines. This reliance has resulted in data skewed toward a group of specific human cell lines, most notably K562, HepG2 and HEK293T [14]. Whilst the sequence itself remains consistent across different cell types, variable expression patterns influence binding opportunities, which are likely further driven by distinct regulatory 'grammar rules' driving observed cell-type-specific patterns [78]. Furthermore, secondary structure differences have been shown to significantly affect RBP binding disparities between K562 and HepG2 cells [43], which can be leveraged for predicting dynamic binding patterns across cell types [44]. Without experimental data covering a broader range of cell types, the relevance of such differences cannot be thoroughly explored. For example, RBPs like TDP-43 and FUS are implicated in memory formation through the creation of sub-cellular RnP granules; such context-specific roles are not adequately represented in the most commonly assayed cell types [79].

Moreover, the field is heavily biased towards human data, with a notable lack of data from other species. In plants, for example, RBPs play essential roles in growth, development, and stress response, yet high-quality RBP binding data for these species

are scarce [80], although recent improvements in immunoprecipitation and RNA-editor approaches aimed specifically at plants are now boosting this area [77,81]. The POPSTAR3 database, which compiles RBP binding data across seven species [37], further highlights the limited share of non-human data available. Expanding RBP binding data in non-human species is crucial not only for the understanding modification-associated regulatory mechanisms in each specific species, but also for deciphering evolutionary relationships in RBP binding. Furthermore, note that a significant challenge with non-human species is the limited understanding of which proteins possess mRNA binding capabilities. For this reason, the integration of deep learning approaches aimed at predicting RBPs and their binding domains [82–85] and their subcellular locations [86] could be key to advancing experimental efforts across various species.

3.2. Focus On Model Interpretation

Interpreting deep learning models, given their highly non-linear feature spaces involving large numbers of weights, is challenging yet crucial for understanding biological contexts [26,87]. Reassuringly, there is a growing trend towards models considering what sequence motifs and/or secondary structure contexts might be driving RBP binding to RNA. In particular, in silico mutagenesis or related approaches can be used to interrogate how base changes in the input sequences influences the binding predictions, and are further useful for informing motif detection algorithms [88–91]. As an example, Grønning et al. used their models to show that point mutations known to cause exon skipping were predicted to result in increased binding of the RBP SRSF1, which has known roles in exon inclusion [34], and the authors of iM6A looked at the impact of single nucleotide variants on m^6A deposition probabilities [33], which showed agreement with experimental data. Moreover, two recent studies leverage their models to make predictions on viral RNAs, where there are very few training data to work with [29,30], and the results were at least in part validated by external datasets. Significant challenges persist, however, as RBPs can be highly redundant and have highly redundant binding sites, such that a single-nucleotide mutation may not be sufficient to alter the binding probability. Therefore, more work is required to build robust frameworks of how sequence variation affects molecular function and disease through its impact on RBP binding.

3.3. Extensions to Predictions on Non-Coding RNAs

Non-coding RNAs, such as long non-coding RNAs, enhancer RNAs, microRNAs, and others, exhibit highly diverse functional roles and pronounced cell-specificity [92]. Moreover, the functional roles of these RNA species in the context of their interactions with RNA modifiers such as RBPs or m^6A remain poorly understood [93]. One suggested function of lncRNAs is to act as molecular scaffolds or decoys, potentially sequestering RBPs from target genes, with implications in immune regulation [94]. Additionally, recent studies have revealed an enrichment of m^6A modification in non-coding RNAs, with m^6A-reader RBP YTHDC1 playing a role in maintaining RNA integrity [95].

On the whole, the limited availability of data on non-coding RNA modifications poses a significant challenge in terms of training deep learning models. One approach could be to employ models trained on mRNA-based data to predict modifications in non-coding RNAs. Due to the current lack of experimental methods, however, validations of these observations remain difficult; therefore, technological advancements that can lead to high quality and throughput at these regions could have both large impact in terms of validating current deep learning based observations, as well as training new models.

3.4. Cooperative Contexts and Interplay with Other Modifications

RBPs do not operate in isolation, but often exhibit redundant behaviour or act in collaboration or competition with other RBPs. For instance, the YTH-domain m^6A reader RBPs display highly redundant functions [96,97], while the RBPs HuR and AUF1 are known to compete for binding sites, affecting the stability of shared mRNA targets [98]. Moreover,

different types of modifications do not act independently. For example, the presence or absence of A-so-I RNA editing can significantly modify RBP binding patterns [99,100]. Such interactions underscore the highly complex nature of RNA regulation, where modifications and RBPs form a dynamic dependency network. Therefore, studying the binding patterns of a single RBP or modification may not sufficiently capture the nuances of RNA stability and decay. Indeed, research in this area has demonstrated that considering the full repertoire of RBPs yields a more accurate prediction of RNA half-life than analysing any individual RBP [43]. However, a significant limitation in current research is the scarcity of experimental methods capable of establishing potential cooperative binding locations on a genome-wide scale; addressing this gap would provide essential data for establishing ground truths on which to train deep learning-based modelling frameworks.

4. Conclusions

This article has provided an overview of recent advances in deep learning in the context of RNA modifications, highlighting areas in which there are distinct challenges and opportunities. It is important to emphasise that this is a cyclical process, with experimental data forming a basis for deep learning models, and these improved models, in turn, can guide the development of either improved or more targeted experimental methodologies. Consequently, future collaboration between experimental and computational biologists will be key for driving progress in the RNA modification field, allowing for the construction of powerful and highly interpretable models able to answer biological questions in a range of species and contexts.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Delaunay, S.; Helm, M.; Frye, M. RNA modifications in physiology and disease: Towards clinical applications. *Nat. Rev. Genet.* **2024**, *25*, 104–122. [CrossRef] [PubMed]
2. Barbieri, I.; Kouzarides, T. Role of RNA modifications in cancer. *Nat. Rev. Cancer* **2020**, *20*, 303–322. [CrossRef] [PubMed]
3. Gerstberger, S.; Hafner, M.; Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **2014**, *15*, 829–845. [CrossRef] [PubMed]
4. Hentze, M.W.; Castello, A.; Schwarzl, T.; Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 327–341. [CrossRef] [PubMed]
5. Dominguez, D.; Freese, P.; Alexis, M.S.; Su, A.; Hochman, M.; Palden, T.; Bazile, C.; Lambert, N.J.; Van Nostrand, E.L.; Pratt, G.A.; et al. Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell* **2018**, *70*, 854–867. [CrossRef] [PubMed]
6. Ke, S.; Alemu, E.A.; Mertens, C.; Gantman, E.C.; Fak, J.J.; Mele, A.; Haripal, B.; Zucker-Scharff, I.; Moore, M.J.; Park, C.Y.; et al. A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev.* **2015**, *29*, 2037–2053. [CrossRef] [PubMed]
7. Patil, D.P.; Pickering, B.F.; Jaffrey, S.R. Reading m6A in the transcriptome: M6A-binding proteins. *Trends Cell Biol.* **2018**, *28*, 113–127. [CrossRef] [PubMed]
8. Eisenberg, E.; Levanon, E.Y. A-to-I RNA editing—immune protector and transcriptome diversifier. *Nat. Rev. Genet.* **2018**, *19*, 473–490. [CrossRef]
9. Ule, J.; Jensen, K.B.; Ruggiu, M.; Mele, A.; Ule, A.; Darnell, R.B. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **2003**, *302*, 1212–1215. [CrossRef]
10. Dominissini, D.; Moshitch-Moshkovitz, S.; Schwartz, S.; Salmon-Divon, M.; Ungar, L.; Osenberg, S.; Cesarkas, K.; Jacob-Hirsch, J.; Amariglio, N.; Kupiec, M.; et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **2012**, *485*, 201–206. [CrossRef]
11. Linder, B.; Grozhik, A.V.; Olarerin-George, A.O.; Meydan, C.; Mason, C.E.; Jaffrey, S.R. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* **2015**, *12*, 767–772. [CrossRef]

12. König, J.; Zarnack, K.; Rot, G.; Curk, T.; Kayikci, M.; Zupan, B.; Turner, D.J.; Luscombe, N.M.; Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **2010**, *17*, 909–915. [CrossRef] [PubMed]
13. Van Nostrand, E.L.; Pratt, G.A.; Shishkin, A.A.; Gelboin-Burkhart, C.; Fang, M.Y.; Sundararaman, B.; Blue, S.M.; Nguyen, T.B.; Surka, C.; Elkins, K.; et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **2016**, *13*, 508–514. [CrossRef] [PubMed]
14. Van Nostrand, E.L.; Freese, P.; Pratt, G.A.; Wang, X.; Wei, X.; Xiao, R.; Blue, S.M.; Chen, J.Y.; Cody, N.A.; Dominguez, D.; et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **2020**, *583*, 711–719. [CrossRef] [PubMed]
15. Wheeler, E.C.; Van Nostrand, E.L.; Yeo, G.W. Advances and challenges in the detection of transcriptome-wide protein–RNA interactions. *Wiley Interdiscip. Rev. Rna* **2018**, *9*, e1436. [CrossRef] [PubMed]
16. Rahman, R.; Xu, W.; Jin, H.; Rosbash, M. Identification of RNA-binding protein targets with HyperTRIBE. *Nat. Protoc.* **2018**, *13*, 1829–1849. [CrossRef] [PubMed]
17. Meyer, K.D. DART-seq: An antibody-free method for global m6A detection. *Nat. Methods* **2019**, *16*, 1275–1280. [CrossRef] [PubMed]
18. Ray, D.; Kazan, H.; Cook, K.B.; Weirauch, M.T.; Najafabadi, H.S.; Li, X.; Gueroussov, S.; Albu, M.; Zheng, H.; Yang, A.; et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **2013**, *499*, 172–177. [CrossRef] [PubMed]
19. Lambert, N.; Robertson, A.; Jangi, M.; McGeary, S.; Sharp, P.A.; Burge, C.B. RNA Bind-n-Seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell* **2014**, *54*, 887–900. [CrossRef]
20. Dai, Q.; Zhang, L.S.; Sun, H.L.; Pajdzik, K.; Yang, L.; Ye, C.; Ju, C.W.; Liu, S.; Wang, Y.; Zheng, Z.; et al. Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nat. Biotechnol.* **2023**, *41*, 344–354. [CrossRef]
21. Liu, C.; Sun, H.; Yi, Y.; Shen, W.; Li, K.; Xiao, Y.; Li, F.; Li, Y.; Hou, Y.; Lu, B.; et al. Absolute quantification of single-base m6A methylation in the mammalian transcriptome using GLORI. *Nat. Biotechnol.* **2023**, *41*, 355–366. [CrossRef] [PubMed]
22. Garalde, D.R.; Snell, E.A.; Jachimowicz, D.; Sipos, B.; Lloyd, J.H.; Bruce, M.; Pantic, N.; Admassu, T.; James, P.; Warland, A.; et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **2018**, *15*, 201–206. [CrossRef] [PubMed]
23. Hendra, C.; Pratanwanich, P.N.; Wan, Y.K.; Goh, W.S.; Thiery, A.; Göke, J. Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat. Methods* **2022**, *19*, 1590–1598. [CrossRef]
24. Mateos, P.A.; Sethi, A.; Ravindran, A.; Guarnacci, M.; Srivastava, A.; Xu, J.; Woodward, K.; Yuen, Z.; Mahmud, S.; Kanchi, M.; et al. Simultaneous identification of m6A and m5C reveals coordinated RNA modification at single-molecule resolution. *bioRxiv* **2022.** . [CrossRef]
25. Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **2016**, *12*, 878. [CrossRef] [PubMed]
26. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nat. Genet.* **2019**, *51*, 12–18. [CrossRef] [PubMed]
27. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 2019. Available online: https://dl.acm.org/doi/10.5555/3454287.3455008 (accessed on 10 May 2024).
28. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for Large-Scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
29. Horlacher, M.; Wagner, N.; Moyon, L.; Kuret, K.; Goedert, N.; Salvatore, M.; Ule, J.; Gagneur, J.; Winther, O.; Marsico, A. Towards In-Silico CLIP-seq: Predicting Protein-RNA Interaction via Sequence-to-Signal Learning. *Genome Biol.* **2022**, *24*, 180. [CrossRef]
30. Xu, Y.; Zhu, J.; Huang, W.; Xu, K.; Yang, R.; Zhang, Q.C.; Sun, L. PrismNet: Predicting protein–RNA interaction using in vivo RNA structural information. *Nucleic Acids Res.* **2023**, *51*, W468–W477. [CrossRef]
31. Zhang, S.; Zhou, J.; Hu, H.; Gong, H.; Chen, L.; Cheng, C.; Zeng, J. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.* **2016**, *44*, e32. [CrossRef]
32. Laverty, K.U.; Jolma, A.; Pour, S.E.; Zheng, H.; Ray, D.; Morris, Q.; Hughes, T.R. PRIESSTESS: Interpretable, high-performing models of the sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Res.* **2022**, *50*, e111. [CrossRef]
33. Luo, Z.; Zhang, J.; Fei, J.; Ke, S. Deep learning modeling m6A deposition reveals the importance of downstream cis-element sequences. *Nat. Commun.* **2022**, *13*, 2720. [CrossRef]
34. Grønning, A.G.B.; Doktor, T.K.; Larsen, S.J.; Petersen, U.S.S.; Holm, L.L.; Bruun, G.H.; Hansen, M.B.; Hartung, A.M.; Baumbach, J.; Andresen, B.S. DeepCLIP: Predicting the effect of mutations on protein–RNA binding with deep learning. *Nucleic Acids Res.* **2020**, *48*, 7099–7118. [CrossRef]
35. Mukherjee, N.; Wessels, H.H.; Lebedeva, S.; Sajek, M.; Ghanbari, M.; Garzia, A.; Munteanu, A.; Yusuf, D.; Farazi, T.; Hoell, J.I.; et al. Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res.* **2019**, *47*, 570–581. [CrossRef] [PubMed]
36. Stražar, M.; Žitnik, M.; Zupan, B.; Ule, J.; Curk, T. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics* **2016**, *32*, 1527–1535. [CrossRef]
37. Zhao, W.; Zhang, S.; Zhu, Y.; Xi, X.; Bao, P.; Ma, Z.; Kapral, T.H.; Chen, S.; Zagrovic, B.; Yang, Y.T.; et al. POSTAR3: An updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.* **2022**, *50*, D287–D294. [CrossRef] [PubMed]

38. Tang, Y.; Chen, K.; Song, B.; Ma, J.; Wu, X.; Xu, Q.; Wei, Z.; Su, J.; Liu, G.; Rong, R.; et al. m6A-Atlas: A comprehensive knowledgebase for unraveling the N 6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Res.* **2021**, *49*, D134–D143. [CrossRef] [PubMed]
39. Liang, Z.; Ye, H.; Ma, J.; Wei, Z.; Wang, Y.; Zhang, Y.; Huang, D.; Song, B.; Meng, J.; Rigden, D.J.; et al. m6A-Atlas v2. 0: Updated resources for unraveling the N 6-methyladenosine (m6A) epitranscriptome among multiple species. *Nucleic Acids Res.* **2024**, *52*, D194–D202. [CrossRef]
40. Krakau, S.; Richard, H.; Marsico, A. PureCLIP: Capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biol.* **2017**, *18*, 240 . [CrossRef]
41. Uren, P.J.; Bahrami-Samani, E.; Burns, S.C.; Qiao, M.; Karginov, F.V.; Hodges, E.; Hannon, G.J.; Sanford, J.R.; Penalva, L.O.; Smith, A.D. Site identification in high-throughput RNA–protein interaction data. *Bioinformatics* **2012**, *28*, 3013–3020. [CrossRef]
42. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R.V. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112–2120. [CrossRef]
43. Sun, L.; Xu, K.; Huang, W.; Yang, Y.T.; Li, P.; Tang, L.; Xiong, T.; Zhang, Q.C. Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. *Cell Res.* **2021**, *31*, 495–516. [CrossRef]
44. Zhu, H.; Yang, Y.; Wang, Y.; Wang, F.; Huang, Y.; Chang, Y.; Wong, K.c.; Li, X. Dynamic characterization and interpretation for protein-RNA interactions across diverse cellular conditions using HDRNet. *Nat. Commun.* **2023**, *14*, 6824. [CrossRef] [PubMed]
45. Pan, X.; Fang, Y.; Li, X.; Yang, Y.; Shen, H.B. RBPsuite: RNA-protein binding sites prediction suite based on deep learning. *BMC Genom.* **2020**, *21*, 884 . [CrossRef] [PubMed]
46. Yamada, K.; Hamada, M. Prediction of RNA–protein interactions using a nucleotide language model. *Bioinform. Adv.* **2022**, *2*, vbac023. [CrossRef] [PubMed]
47. Zhang, J.; Liu, B.; Wang, Z.; Lehnert, K.; Gahegan, M. DeepPN: A deep parallel neural network based on convolutional neural network and graph convolutional network for predicting RNA-protein binding sites. *BMC Bioinform.* **2022**, *23*, 257. [CrossRef] [PubMed]
48. Uhl, M.; Tran, V.D.; Heyl, F.; Backofen, R. RNAProt: An efficient and feature-rich RNA binding protein binding site predictor. *GigaScience* **2021**, *10*, giab054. [CrossRef] [PubMed]
49. Ghanbari, M.; Ohler, U. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res.* **2020**, *30*, 214–226. [CrossRef]
50. Picardi, E.; D'Erchia, A.M.; Lo Giudice, C.; Pesole, G. REDIportal: A comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **2017**, *45*, D750–D757. [CrossRef]
51. Wang, J.; Ness, S.; Brown, R.; Yu, H.; Oyebamiji, O.; Jiang, L.; Sheng, Q.; Samuels, D.C.; Zhao, Y.Y.; Tang, J.; et al. EditPredict: Prediction of RNA editable sites with convolutional neural network. *Genomics* **2021**, *113*, 3864–3871. [CrossRef]
52. Horlacher, M.; Cantini, G.; Hesse, J.; Schinke, P.; Goedert, N.; Londhe, S.; Moyon, L.; Marsico, A. A Systematic Benchmark of Machine Learning Methods for Protein-RNA Interaction Prediction. *Briefings Bioinform.* **2023**, *24*, bbad307. [CrossRef]
53. Jaganathan, K.; Panagiotopoulou, S.K.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.I.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting splicing from primary sequence with deep learning. *Cell* **2019**, *176*, 535–548. [CrossRef] [PubMed]
54. Li, X.; Wang, K.; Lyu, Y.; Pan, H.; Zhang, J.; Stambolian, D.; Susztak, K.; Reilly, M.P.; Hu, G.; Li, M. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* **2020**, *11*, 2338. [CrossRef] [PubMed]
55. Han, K.; Sheng, V.S.; Song, Y.; Liu, Y.; Qiu, C.; Ma, S.; Liu, Z. Deep semi-supervised learning for medical image segmentation: A review. *Expert Syst. Appl.* **2024**, *245*, 123052. [CrossRef]
56. Han, H.; Talpur, B.A.; Liu, W.; Wang, L.; Ahmed, B.; Sarhan, N.; Awwad, E.M. RNA-RBP interactions recognition using multi-label learning and feature attention allocation. *J. Cloud Comput.* **2024**, *13*, 54. [CrossRef]
57. Pan, X.; Rijnbeek, P.; Yan, J.; Shen, H.B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genom.* **2018**, *19*, 511. [CrossRef] [PubMed]
58. Trabelsi, A.; Chaabane, M.; Ben-Hur, A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. *Bioinformatics* **2019**, *35*, i269–i277. [CrossRef] [PubMed]
59. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
60. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
61. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
62. Wang, X.; Zhang, M.; Long, C.; Yao, L.; Zhu, M. Self-attention based neural network for predicting RNA-protein binding sites. *IEEE/Acm Trans. Comput. Biol. Bioinform.* **2022**, *20*, 1469–1479. [CrossRef]
63. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
64. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
65. Maticzka, D.; Lange, S.J.; Costa, F.; Backofen, R. GraphProt: Modeling binding preferences of RNA-binding proteins. *Genome Biol.* **2014**, *15*, R17. [CrossRef]
66. Uhl, M.; Tran, V.; Heyl, F.; Backofen, R. GraphProt2: A novel deep learning-based method for predicting binding sites of RNA-binding proteins. *BioRxiv* **2019**. [CrossRef]

67. Zhao, X.; Chang, F.; Lv, H.; Zou, G.; Zhang, B. A Novel Deep Learning Method for Predicting RNA-Protein Binding Sites. *Appl. Sci.* **2023**, *13*, 3247. [CrossRef]
68. Gruber, A.R.; Lorenz, R.; Bernhart, S.H.; Neuböck, R.; Hofacker, I.L. The vienna RNA websuite. *Nucleic Acids Res.* **2008**, *36*, W70–W74. [CrossRef] [PubMed]
69. Steffen, P.; Voß, B.; Rehmsmeier, M.; Reeder, J.; Giegerich, R. RNAshapes: An integrated RNA analysis package based on abstract shapes. *Bioinformatics* **2006**, *22*, 500–503. [CrossRef]
70. Yan, Z.; Hamilton, W.L.; Blanchette, M. Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. *Bioinformatics* **2020**, *36*, i276–i284. [CrossRef] [PubMed]
71. Spitale, R.C.; Flynn, R.A.; Zhang, Q.C.; Crisalli, P.; Lee, B.; Jung, J.W.; Kuchelmeister, H.Y.; Batista, P.J.; Torre, E.A.; Kool, E.T.; et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **2015**, *519*, 486–490. [CrossRef]
72. Sun, L.; Fazal, F.M.; Li, P.; Broughton, J.P.; Lee, B.; Tang, L.; Huang, W.; Kool, E.T.; Chang, H.Y.; Zhang, Q.C. RNA structure maps across mammalian cellular compartments. *Nat. Struct. Mol. Biol.* **2019**, *26*, 322–330. [CrossRef]
73. Chan, D.; Feng, C.; Spitale, R.C. Measuring RNA structure transcriptome-wide with icSHAPE. *Methods* **2017**, *120*, 85–90. [CrossRef] [PubMed]
74. Hutvagner, G.; Zamore, P.D. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **2002**, *297*, 2056–2060. [CrossRef] [PubMed]
75. Vaculík, O.; Chalupová, E.; Grešová, K.; Majtner, T.; Alexiou, P. Transfer Learning Allows Accurate RBP Target Site Prediction with Limited Sample Sizes. *Biology* **2023**, *12*, 1276. [CrossRef] [PubMed]
76. Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N.L.; Grzywaczewski, A.H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B.P.; Sirelkhatim, H.; et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv* **2023**. . [CrossRef]
77. Arribas-Hernández, L.; Rennie, S.; Köster, T.; Porcelli, C.; Lewinski, M.; Staiger, D.; Andersson, R.; Brodersen, P. Principles of mRNA targeting via the Arabidopsis m6A-binding protein ECT2. *eLife* **2021**, *10*, e72375. [CrossRef]
78. Uhl, M.; Houwaart, T.; Corrado, G.; Wright, P.R.; Backofen, R. Computational analysis of CLIP-seq data. *Methods* **2017**, *118*, 60–72. [CrossRef] [PubMed]
79. Hanan, M.; Soreq, H.; Kadener, S. CircRNAs in the brain. *RNA Biol.* **2017**, *14*, 1028–1034. [CrossRef]
80. Mateos, J.L.; Staiger, D. Toward a systems view on RNA-binding proteins and associated RNAs in plants: Guilt by association. *Plant Cell* **2023**, *35*, 1708–1726. [CrossRef] [PubMed]
81. Lewinski, M.; Brüggemann, M.; Köster, T.; Reichel, M.; Bergelt, T.; Meyer, K.; König, J.; Zarnack, K.; Staiger, D. Mapping protein–RNA binding in plants with individual-nucleotide-resolution UV cross-linking and immunoprecipitation (plant iCLIP2). *Nat. Protoc.* **2024**, *19*, 1183–1234. [CrossRef]
82. Peng, X.; Wang, X.; Guo, Y.; Ge, Z.; Li, F.; Gao, X.; Song, J. RBP-TSTL is a two-stage transfer learning framework for genome-scale prediction of RNA-binding proteins. *Brief. Bioinform.* **2022**, *23*, bbac215. [CrossRef]
83. Zhang, J.; Yan, K.; Chen, Q.; Liu, B. PreRBP-TL: Prediction of species-specific RNA-binding proteins based on transfer learning. *Bioinformatics* **2022**, *38*, 2135–2143. [CrossRef]
84. Arican, O.C.; Gumus, O. PredDRBP-MLP: Prediction of DNA-binding proteins and RNA-binding proteins by multilayer perceptron. *Comput. Biol. Med.* **2023**, *164*, 107317. [CrossRef] [PubMed]
85. Jin, W.; Brannan, K.W.; Kapeli, K.; Park, S.S.; Tan, H.Q.; Gosztyla, M.L.; Mujumdar, M.; Ahdout, J.; Henroid, B.; Rothamel, K.; et al. HydRA: Deep-learning models for predicting RNA-binding capacity from protein interaction association context and protein sequence. *Mol. Cell* **2023**, *83*, 2595–2611. [CrossRef]
86. Wang, J.; Horlacher, M.; Cheng, L.; Winther, O. DeepLocRNA: An interpretable deep learning model for predicting RNA subcellular localisation with domain-specific transfer-learning. *Bioinformatics* **2024**, *40*, btae065. [CrossRef] [PubMed]
87. Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.M.; Zietz, M.; Hoffman, M.M.; et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387. [CrossRef]
88. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the International Conference on Machine Learning. PMLR, Sydney, Australia, 6–11 August 2017; pp. 3145–3153.
89. Shrikumar, A.; Tian, K.; Avsec, Ž.; Shcherbina, A.; Banerjee, A.; Sharmin, M.; Nair, S.; Kundaje, A. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. *arXiv* **2018**, arXiv:1811.00416.
90. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; 2017; pp. 4768–4777. Available online: https://dl.acm.org/doi/10.5555/3295222.3295230 (accessed on 10 May 2024).
91. Nair, S.; Shrikumar, A.; Schreiber, J.; Kundaje, A. fastISM: Performant in silico saturation mutagenesis for convolutional neural networks. *Bioinformatics* **2022**, *38*, 2397–2403. [CrossRef] [PubMed]
92. Marchese, F.P.; Raimondi, I.; Huarte, M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* **2017**, *18*, 206. [CrossRef]
93. Ferre, F.; Colantoni, A.; Helmer-Citterich, M. Revealing protein–lncRNA interaction. *Brief. Bioinform.* **2016**, *17*, 106–116. [CrossRef]
94. Fatica, A.; Bozzoni, I. Long non-coding RNAs: New players in cell differentiation and development. *Nat. Rev. Genet.* **2014**, *15*, 7–21. [CrossRef]

95. Akhtar, J.; Lugoboni, M.; Junion, G. m6A RNA modification in transcription regulation. *Transcription* **2021**, *12*, 266–276. [CrossRef] [PubMed]
96. Zaccara, S.; Jaffrey, S.R. A unified model for the function of YTHDF proteins in regulating m6A-modified mRNA. *Cell* **2020**, *181*, 1582–1595. [CrossRef] [PubMed]
97. Arribas-Hernández, L.; Rennie, S.; Schon, M.; Porcelli, C.; Enugutti, B.; Andersson, R.; Nodine, M.D.; Brodersen, P. The YTHDF proteins ECT2 and ECT3 bind largely overlapping target sets and influence target mRNA abundance, not alternative polyadenylation. *eLife* **2021**, *10*, e72377. [CrossRef] [PubMed]
98. Lal, A.; Mazan-Mamczarz, K.; Kawai, T.; Yang, X.; Martindale, J.L.; Gorospe, M. Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs. *EMBO J.* **2004**, *23*, 3092–3102. [CrossRef] [PubMed]
99. Hu, X.; Zou, Q.; Yao, L.; Yang, X. Survey of the binding preferences of RNA-binding proteins to RNA editing events. *Genome Biol.* **2022**, *23*, 169. [CrossRef] [PubMed]
100. Weirick, T.; Militello, G.; Hosen, M.R.; John, D.; Moore IV, J.B.; Uchida, S. Investigation of RNA Editing Sites within Bound Regions of RNA-Binding Proteins. *High-Throughput* **2019**, *8*, 19. [CrossRef]