

Devin Willmott*, David Murrugarra*, and Qiang Ye*

Improving RNA secondary structure prediction via state inference with deep recurrent neural networks

<https://doi.org/10.1515/cmb-2020-0002>

Received September 21, 2019; accepted February 23, 2020

Abstract: The problem of determining which nucleotides of an RNA sequence are paired or unpaired in the secondary structure of an RNA, which we call RNA state inference, can be studied by different machine learning techniques. Successful state inference of RNA sequences can be used to generate auxiliary information for data-directed RNA secondary structure prediction. Typical tools for state inference, such as hidden Markov models, exhibit poor performance in RNA state inference, owing in part to their inability to recognize non-local dependencies. Bidirectional long short-term memory (LSTM) neural networks have emerged as a powerful tool that can model global nonlinear sequence dependencies and have achieved state-of-the-art performances on many different classification problems.

This paper presents a practical approach to RNA secondary structure inference centered around a deep learning method for state inference. State predictions from a deep bidirectional LSTM are used to generate synthetic SHAPE data that can be incorporated into RNA secondary structure prediction via the Nearest Neighbor Thermodynamic Model (NNTM). This method produces predicted secondary structures for a diverse test set of 16S ribosomal RNA that are, on average, 25 percentage points more accurate than undirected MFE structures. Accuracy is highly dependent on the success of our state inference method, and investigating the global features of our state predictions reveals that accuracy of both our state inference and structure inference methods are highly dependent on the similarity of pairing patterns of the sequence to the training dataset. Availability of a large training dataset is critical to the success of this approach. Code available at <https://github.com/dwillmott/rna-state-inf>.

Keywords: RNA secondary structure inference, RNA state inference, NNTM, SHAPE-directed NNTM, Deep learning, Recurrent neural networks

MSC 2010: 92C40 and 92B20

1 Introduction

The secondary structure of an RNA sequence plays an important role in determining its function [10, 24, 16], but directly observing RNA secondary structure is costly and difficult [3, 9]. Therefore, researchers have used computational tools to predict the secondary structure of RNAs. One of the most popular methods is the Nearest Neighbor Thermodynamic Model (NNTM) [35]. Alternatively, comparative sequence analysis methods [13] use a set of homologous sequences to infer a secondary structure [2]. This method remains the gold standard for secondary structure prediction [30].

Corresponding Authors: All three authors assume the shared corresponding authorship.

Devin Willmott*: Bosch Center for AI, Pittsburgh, PA USA, E-mail: Devin.Willmott@us.bosch.com

David Murrugarra*: Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027 USA, E-mail: murrugarra@uky.edu

Qiang Ye*: Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027 USA, E-mail: qiang.ye@uky.edu

NNTM is based on thermodynamic optimization to find the secondary structure with the minimum free energy (MFE). There are several implementations of NTTM; some of the popular ones include RNAStructure [25], GTfold [31], UNAFold [23], and ViennaRNA package [22]. However, NTTM has been shown to be ill-conditioned [17, 18, 26]. That is, for a given sequence, significantly different secondary structures might exhibit very similar energies. Additionally, the range of accuracies of the predictions of NTTM shows significant variance [31].

More recently, high-throughput data that correlates with the state of a nucleotide being paired or unpaired has been developed. This data, called SHAPE [37] for ‘selective 2’-hydroxyl acylation analyzed by primer extension’, has been incorporated as auxiliary information into the objective function of NTTM with the goal of improving the accuracy of the predictions. This type of prediction is referred to as SHAPE-directed RNA secondary structure modeling [5, 36]. The addition of auxiliary information usually improves the accuracy of the predictions of NTTM [5] but it has been shown that the improvements are correlated with the MFE accuracy [30]. The latter result has been obtained by statistical modeling of SHAPE. The model in [30] gives distributions for the values of SHAPE if the state of the nucleotide (as paired or unpaired or helix-end) is known. Thus the model in [30] can be used to generate SHAPE data for an RNA sequence *in silico*, given the state of each of the sequence’s nucleotides.

In this paper, we present a method for improving the RNA secondary structure prediction based on state inference results. To do so, we first study the problem of determining the state of each nucleotide of an RNA sequence, which we refer to as state inference. State inference is a binary classification task on each nucleotide, which we note is in contrast to full secondary structure inference, which seeks to identify sets of base pairs. We have developed, trained, and tested a deep recurrent neural network that performs this task: given an RNA sequence, the machine outputs a probability that each nucleotide is paired. We can threshold this probability to obtain binary predictions for the state of each nucleotide.

Additionally, we use the probabilities from the state inference method to generate synthetic SHAPE. Then we use this SHAPE data for directed predictions via NTTM, leading to significant improvements in secondary structure accuracy on sequences where the state inference performed well. We note that our approach for generating SHAPE is different from the statistical models in [30], which generate synthetic SHAPE data by sampling from the distribution models.

We note that other deep learning methods for the problem of RNA secondary structure inference have been explored [38]. Although we are primarily interested in using state inference to direct secondary structure predictions, there exist other motivations for state inference. For example, such a method could be used to identify binding sites in RNA-RNA interactions [32, 6].

2 Methods

2.1 SHAPE-Directed NTTM

Under the nearest neighbor thermodynamic model (NNTM), stacks and loops are each assigned a free energy based on base pair type (Watson-Crick or wobble), with pairs contributing negative energy, and loops contributing positive energy. The energy of a secondary structure is the sum of the energies of these local features. Dynamic programming methods can be used to efficiently find the secondary structure with the smallest energy, called the minimum free energy (MFE) structure, which is usually taken as the predicted structure by the NTTM method.

The SHAPE data comes from high-throughput chemical probing experiments and associate a reactivity value to each nucleotide of an RNA sequence. It has been found experimentally that high SHAPE values are correlated with unpaired positions, and small values with paired position. SHAPE has been incorporated into NTTM by adding a pseudo-free energy term to the model [5]. This term is generated by the following relation:

$$\Delta G_{\text{SHAPE}}(i) = 2.6 \cdot \ln(\text{SHAPE}(i) + 1) - 0.8 \quad (1)$$

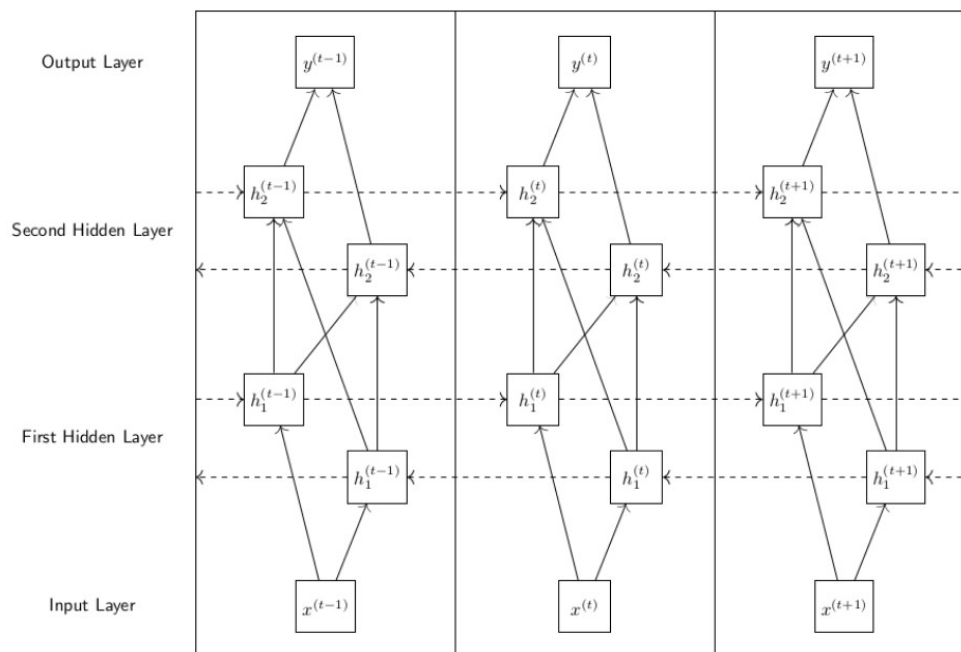


Figure 1: Bidirectional RNN. Diagram of a bidirectional RNN with two hidden layers for state inference at three different timesteps. Here x denotes the input, h_1 the first layer hidden variable, h_2 the second layer hidden variable, and y the output, with superscripts representing timesteps and arrows showing the propagation of information through the network. Each hidden layer combines information from the previous layer, earlier and later timesteps, and an internal memory state to compute its output.

This energy is added to base pair stacks involving nucleotide i . In effect, the term $\Delta G_{\text{SHAPE}}(i)$ penalizes base pairs involving nucleotides with high SHAPE values, and encourages base pairs among nucleotides with low SHAPE values. Incorporating SHAPE data consistently leads to significant increases in the accuracy of the MFE structure [5, 36, 30].

2.2 Directing NNTM With State Information

We are interested in using deep learning tools in tandem with SHAPE direction to improve secondary structure inference. However, deep learning methods require extraordinary large datasets; the relative dearth of available experimental SHAPE data prohibits us from directly learning synthetic SHAPE data with a neural network. Instead, our proposed method for secondary structure inference uses a viable method for state inference with the ability to influence the NNTM energy function on a per-nucleotide basis via SHAPE direction. Note that the same limitation applies to the secondary structure data for which a large dataset is available for 16S sequences only but not for other RNA types; see Section 4.4 for more details.

Our method is a three-step process:

1. A machine learning method for predicting the state of each nucleotide in a sequence;
2. A function converting these state predictions into artificial SHAPE data;
3. The SHAPE-directed NNTM function that takes both the original RNA sequence and the generated SHAPE data and outputs a predicted secondary structure.

Note that Step 3 is a well established method and the novelty here is in deriving the artificial SHAPE data in Steps 1 and 2. For the task of state inference, we trained a deep neural network using a set of known RNA

sequences and structures that generates a sequence of state predictions, detailed in Section 2.3. The output of this neural network is a sequence p of the same length as the original RNA sequence, where $p(i)$ is the predicted probability that the nucleotide in position i is paired.

With these predictions in hand, we convert each predicted probability $p(i)$ to a SHAPE value to be associated with nucleotide i . To construct a function for this purpose, we note that a SHAPE value of ≈ 0.3603 will not contribute any positive or negative energy to the NNTM energy function; this can be seen by setting $\Delta G_{\text{SHAPE}}(i)$ to 0 in Equation 1 and solving for SHAPE(i) [8]. We would therefore like to assign predictions of 0.5 to a SHAPE value of 0.3603, as these predictions give no information as to the state of the nucleotide. With this in mind, we use the following piecewise linear function to generate SHAPE, where a and b are constants to be specified.

$$f(i) = \begin{cases} 2(0.3603 - b)p(i) + b, & \text{if } 0 < p(i) \leq 0.5 \\ 2(a - 0.3603)(p(i) - 1) + a, & \text{if } 0.5 < p(i) \leq 1 \end{cases} \quad (2)$$

This function has range $[a, b]$, with $f(i) = a$ if $p(i) = 1$, $f(i) = b$ if $p(i) = 0$, and $f(i) = 0.3603$ if $p(i) = 0.5$. To determine values of a and b , we considered experimentally collected SHAPE data from two *E. coli* sequences, one 16S sequence and one 23S sequence [30]. Together, these two sequences contain a total of 4187 nucleotides, and represent a wide variety of structural motifs. We took the mean SHAPE value among both paired nucleotides and unpaired nucleotides; these values are 0.214 and 0.6624, respectively.

All of our experimental results in Section 3 will use $a = 0.214$ and $b = 0.6624$ in our SHAPE generation function. These choices are motivated by real SHAPE values, and thus are sensible estimations of the best values. However, they may not be the optimal values for our purposes. In Section 4.2, we explore how varying these values may affect the accuracy of our predicted secondary structures. These experiments indicate that the a and b values used in our experimental results are nearly optimal.

With set values of a and b , we can generate a sequence of artificial SHAPE data. We then use SHAPE-directed NNTM as described in the previous section to obtain our secondary structure prediction.

2.3 State Inference with Deep Neural Networks

Neural networks are tools from the realm of machine learning for solving classification and regression problems. In a neural network, model parameters are trained using a dataset of known input-output pairs: we define a loss function based on the difference between machine predictions and target outputs, retrieve gradient directions for parameters with respect to this loss using the backpropagation algorithm [27], and optimize parameters using iterative first order methods, such as gradient descent.

Recent advances in machine learning come primarily from deep neural networks [11], which are stacks of multiple neural networks: the output of one neural network in the stack acts as the input for the next. Each of these constituent neural networks is referred to as a layer of the deep neural network. These multiple layers allow the deep neural network to learn and represent complex nonlinear relationships among inputs.

For the task of state inference, we use a deep recurrent neural network (deep RNN). RNNs work specifically with sequential data by combining the learning methods of neural networks with the architecture of a discrete-time dynamical system. A single layer RNN has a state $h^{(i)}$ that is a function of the state at previous time step $h^{(i-1)}$ and the input $x^{(i)}$ at time i . The function is a composition of an elementwise nonlinear activation function with a linear map. Sequence elements are fed to the machine as inputs one at a time; at time i , the machine receives the nucleotide in position i , given as a one-hot encoding $x^{(i)}$. From this input and the previous state $h^{(i-1)}$, machine parameters generate a state $h^{(i)}$ that encodes the pairing information up to step i , from which another function produces the output $y^{(i)}$, the machine prediction of the probability that nucleotide i is paired.

We make a number of modifications to our deep RNN to increase state inference accuracy. Most notably, we use a popular variant of RNN called the Long Short-Term Memory (LSTM) architecture [14], which incorporates a gating mechanism and a memory cell to increase accuracy beyond that of traditional RNNs on a

variety of sequential learning tasks. Specifically, the LSTM gating architecture allows a certain component of the state to directly pass into future steps, maintaining the flow of state information or memory over the long term. This significantly increases machine capability to model long term dependence and hence capture potential long-range base pairs [21]. We also make this network bidirectional [28, 12]; this is a minor modification that allows information to flow both forward and backward through the sequence. A two-layer bidirectional RNN is shown in Figure 1.

2.4 Dataset, Implementation, and Metrics

Our experiments will focus on a test set of sixteen 16S ribosomal RNA sequences used in SHAPE direction experiments in [30]. Sequences in this set have a wide range of NNTM accuracies.

Our deep neural network requires a large dataset of RNA sequences with known states from which to learn. For this task, we used secondary structure data from the Comparative RNA Web site, run by the Gutell Lab at the University of Texas [1]. This site hosts a collection of known RNA sequences and secondary structures obtained using comparative sequence analysis. Compiling all of the available 16S rRNA results in a set of 17032 sequences and a total of over 21 million nucleotides. We refer to this as the CRW dataset.

To ensure that our model does not simply memorize large portions of sequences in the test set, we compared each CRW dataset sequence with each test set sequence and removed CRW sequences with significant similarities prior to training. In this filtering process, if the two sequences have a common block of nucleotides of more than 10% of the length of the test sequence, or if the two sequences can be aligned such that they have common nucleotides accounting for more than 80% of nucleotides of the shorter sequence, we remove it from the training set. See available code for additional details. This process leaves us with 13118 sequences and a total of approximately 16.5 million nucleotides, with a mean and median sequence length of 1264 and 1431, respectively. We then split this set into two random halves to produce a training and validation set.

For the sake of comparison, we also trained and tested a number of higher-order hidden Markov models (HMM) using the same training, validation, and test sets used by the neural network. In a k order HMM, hidden state transitions depend on the previous k states in the sequences, rather than only the previous state; this improves their representational capacity, but at the cost of a model size that increases exponentially with k . We trained our HMMs with maximum likelihood estimation over the training set, and performed inference with the Viterbi algorithm [7], a backtracking algorithm that exploits the Markovian nature of HMMs to efficiently produce the likeliest state sequence under the model's transition probabilities. HMMs are fundamentally incapable of recognizing dependencies across many timesteps, and we therefore expect the deep neural network to outperform the HMM. However, they provide a baseline against which to measure neural network output. We note that the HMM formalism has been used before for state inference of RNA but for a different purpose [29, 20, 15]. Code of the HMM implementation is available at <https://github.com/dwillmott/rna-state-inf>.

We implemented a variety of deep recurrent networks in Keras [4], a Python deep learning API, with Theano [33] as a backend; code is available at <https://github.com/dwillmott/rna-state-inf>. We found a four-layer network to be the optimal balance of representational capacity and training speed. The largest layers are the middle two, which are both bidirectional LSTMs. The first and last layers are small one-dimensional convolutional layers [19, 11]. These layers act as learnable pre- and post-processing convolutions; they take in and process local information in small regions of the sequence, and allow the two recurrent layers to focus on long-range dependencies across many timesteps. The output dimension of each of the machine's four layers at each timestep are 100, 400, 100, and 2, respectively, giving a machine with a total of 595,552 trainable parameters, which we trained using binary cross entropy loss and RMSprop [34], a gradient-descent like training algorithm. See available code material for hyperparameters and training modifications.

Our code works quite efficiently for the kind of sequences tested in this paper. On a 2080 GeForce RTX 2080 Ti GPU, training takes about 4 hours, and once trained, state inference takes about 0.005 seconds for a 5S sequence (≈ 150 nucleotides), 0.05s for a 16S sequence (≈ 1500 nucleotides), and 0.1s for a 23S sequence (≈ 3000 nucleotides), which are roughly linear with the size of the sequence.

With the sequence of state prediction probabilities p , we generated artificial SHAPE data using Equation 2. Finally, we used both the original sequence and the generated SHAPE data as input for GTfold [31], an efficient NNTM implementation, to compute a SHAPE-directed MFE structure. GTFold is used in this work for all NNTM experiments and results.

When evaluating predicted secondary structures, we compare its set of base pairs with those of the native structure. A predicted base pair is counted as true positive (TP) if it exists in both the predicted and native structure, a false positive (FP) if it appears in the predicted structure but not in the native structure, and a false negative (FN) if it appears in the native structure and not in the predicted structure. We report on three measures of performance: PPV, the proportion of true positives in the predicted structure ($\frac{TP}{TP+FP}$); sensitivity, the fraction of true positives in the native structure ($\frac{TP}{TP+FN}$); and accuracy, the arithmetic mean of PPV and sensitivity: $\frac{1}{2}(\frac{TP}{TP+FP} + \frac{TP}{TP+FN})$ [10].

In later sections, we will also consider the performance of our deep learning methods for state inference. Unlike secondary structure inference, which classifies base pairs, state inference is a binary classification of each nucleotide. When discussing the accuracy of state predictions, we will define accuracy to be the proportion of true predictions among all predictions in the sequence. This is distinct from notions of secondary structure inference accuracy, and so the metrics on state inference and structure inference cannot be directly compared.

3 Results

3.1 Native State Directed NNTM

Before analyzing the results of the entire pipeline of our method, we first examined our SHAPE generation function in detail. To do so, we used the native state of each sequence in our test set to generate SHAPE. This was done by setting $p(i)$ to 1 if the nucleotide in position i is paired, and $p(i)$ to 0 if it is unpaired. We then use Equation 2 to generate artificial SHAPE. This will result in a generated SHAPE value of 0.6624 for all paired nucleotides and 0.214 for unpaired nucleotides, which we then use to direct NNTM. We refer to the resulting predicted structures as native state directed MFE.

This experiment is similar to those run in [30], and uses the same set of data to choose appropriate SHAPE values. The difference is in the method of SHAPE generation: whereas that paper constructs SHAPE distributions from the data and stochastically samples from these distributions, we use the mean of paired and unpaired nucleotides' SHAPE values.

The results of this experiment reinforce many of the findings in [30]. A comparison of accuracy of all three methods (undirected MFE, stochastically directed MFE from [30], and native state directed MFE) is available in Table 1. Overall, native state directed MFE structures are highly accurate, with twelve of the sixteen test sequences enjoying accuracy above 80%. Both direction methods are an improvement on the accuracy of the undirected MFE structure for every test set sequence, and native state directed accuracy represents a further improvement from the stochastic model in [30]. In the case of native state direction, accuracy improvements over undirected MFE range between 15 percentage points (*H. volcanii*) and 57 percentage points (*Cryptonomas.sp.*). Consistent with observations in [30], greatest increases are concentrated in sequences with middling undirected MFE accuracy; for sequences with undirected accuracy between 25% and 45%, native state directed MFE accuracy is an improvement by more than 40 percentage points.

This experiment is equivalent to assuming that our deep learning state inference method has perfect performance, and as such we can interpret the accuracy of native state directed MFE structures to be an upper bound on the performance of our method. On average, the high accuracy exhibited in this experiment gives strong evidence that there are large potential gains in MFE accuracy to be made with our method. However, several sequences with low undirected MFE accuracy sequences like *E. cuniculi* and *C. elegans* are known to be particularly resistant to SHAPE direction [30], and this is reflected in relatively poor native state directed

Table 1: Structure Inference Results on the test set

Sequence Name	Undirected MFE	Directed MFE		
		Predicted	Sükösd	Native
<i>E. cuniculi</i>	0.171	0.183	0.273	0.336
<i>V. necatrix</i>	0.181	0.314	0.503	0.705
<i>C. elegans</i>	0.203	0.248	0.308	0.519
<i>E. nidulans</i>	0.272	0.325	0.601	0.832
<i>N. tabacum</i>	0.323	0.692	0.593	0.859
<i>Cryptomonas.sp</i>	0.339	0.838	0.739	0.898
<i>Synechococcus.sp</i>	0.361	0.848	0.697	0.885
<i>M. musculus</i>	0.375	0.397	0.509	0.782
<i>M. gallisepticum</i>	0.385	0.849	0.721	0.889
<i>E. coli</i>	0.411	0.852	0.744	0.880
<i>B. subtilis</i>	0.512	0.848	0.753	0.881
<i>D. desulfuricans</i>	0.533	0.875	0.724	0.898
<i>C. reinhardtii</i>	0.537	0.845	0.702	0.868
<i>T. maritima</i>	0.562	0.881	0.733	0.896
<i>T. tenax</i>	0.619	0.766	0.754	0.861
<i>H. volcanii</i>	0.752	0.864	0.809	0.907
Mean	0.408	0.664	0.635	0.806
Median	0.380	0.841	0.712	0.874

Table of accuracy of MFE structures using NNTM with a variety of SHAPE directions. First column: undirected MFE. Second column: predicted state directed MFE (see Section 3.2). Third column: mean performance of sampled SHAPE directed NNTM in [30]. Fourth column: native state directed NNTM (see section 3.1).

MFE accuracy. We thus cannot expect our method to exhibit large improvements over undirected MFE structures in these cases.

3.2 Predicted State Directed NNTM

We now use the predictions from our deep neural network to generate SHAPE data that will in turn direct NNTM; we refer to these predictions as predicted state directed MFE structures. We emphasize that, unlike the native state direction explored in the previous section, this method does not assume prior knowledge of the state of the sequence, and thus represents a practical method of secondary structure inference.

The results of applying our method to the sequences in the test set are available in Table 1, which indicates that the extraordinary gains from native state directed NNTM are not always preserved in practice. Predicted state directed structures fall into two clear categories: five are quite inaccurate, with accuracy below 40%, while among remaining eleven structures are all near or above 70%, and nine of these are above 80%. Even with the high variance of accuracies among these structures, predicted state directed MFE structures are 25 percentage points more accurate than undirected MFE structures on average, and every sequence in the test set experiences some increase in accuracy. However, these improvements vary greatly, with several sequence staying within 5% of undirected MFE accuracy, while for four other sequences we improve by more than 40%, with the highest improvement (*Cryptomonas.sp*) at 50%.

To some extent, poor accuracy is explained by our experiment with native state directed NNTM. Indeed, the five sequences with poor accuracy from our method are the five worst-performing with native state directed using native states to direct NNTM, and only one of these exceeds 80% with native states. At worst,

native state directed NNTM gives only 34% accuracy for *E. cuniculi* and 52.6% for *C. elegans*, and this ceiling is much lower than accuracy achieved in many of our other predicted structures. Thus, the difficulties of our method with these sequences may be attributed to the RNA sequences themselves that are particularly challenging for structure inference as discussed in the directability of NNTM section in [30, p.2812]. However, in these cases and some others (*E. nidulans*, *M. musculus*), predicted state directed MFE accuracy does not come close to native state directed MFE accuracy. This is in contrast to our highest performing sequences (*D. desulfuricans*, *T. maritima*), where predicted state directed MFE is within several percentage points of native state directed MFE accuracy.

4 Discussion

4.1 State Inference Accuracy

The foundation of our method is our deep neural network for state inference: this network provides probabilities that are converted into a pseudo-free energy term in the NNTM energy function. To understand the sources of high and low performance of our structure inference method, we can directly evaluate the output of our deep neural network for state inference, using higher order HMMs as a baseline.

To calculate the accuracy of the neural network's output, we thresholded each prediction $p(i)$ above and below 0.5, taking $p(i) > 0.5$ to be a positive prediction and $p(i) < 0.5$ to be a negative prediction. The accuracy, PPV, and sensitivity of both neural network and HMM predictions are shown in Table 2. Though the table exhibits an upward trend in accuracy as the order of the HMM increases, we found that accuracy plateaued and eventually decreased beyond order 5. As expected, the LSTM clearly outperforms HMMs of all orders on the validation set. More importantly, this is the case for our test set as well, where the LSTM outperforms the best HMM in accuracy by nearly 13%.

The order 4 HMM exhibits the highest accuracy on the validation set. We further compared state inference accuracy on each test set sequence using both the order 4 HMM and the neural network. The accuracy, PPV, and sensitivity of these predictions are shown in Table 3. The accuracy of neural network state predictions were, on average, 15 percentage points higher than that of the HMM, and was higher for every sequence but one (*C. elegans*). Table 3 orders sequences in ascending order of undirected MFE accuracy; however, this ordering reveals no straightforward relationships among neural network state inference accuracy, HMM state inference accuracy, and MFE structure accuracy. Neural network accuracy varies much more among sequences: the difference between the sequences with lowest and highest accuracy (*C. elegans* and *B. subtilis*, respectively) is more than 40 percentage points. Sequences can be grouped according to accuracy: poor

Table 2: State inference accuracy of neural network vs. HMM on validation and test sets

Machine	Validation Set			Test Set		
	Acc	PPV	Sen	Acc	PPV	Sen
Order 1 HMM	0.623	0.632	0.852	0.612	0.646	0.767
Order 2 HMM	0.662	0.671	0.826	0.651	0.686	0.759
Order 3 HMM	0.674	0.693	0.794	0.672	0.713	0.750
Order 4 HMM	0.685	0.714	0.771	0.684	0.729	0.742
Order 5 HMM	0.684	0.711	0.776	0.683	0.730	0.742
Neural Network	0.954	0.950	0.972	0.839	0.858	0.873

A comparison of accuracy, PPV, and sensitivity of output from our LSTM-based neural network for state inference compared with those of HMMs of various orders. The composition of validation and test sets is described in Section 2.4.

Table 3: State inference results on the test set from LSTM vs. HMM

Sequence Name	Accuracy		PPV		Sensitivity	
	LSTM	HMM	LSTM	HMM	LSTM	HMM
<i>E. cuniculi</i>	0.680	0.661	0.713	0.693	0.774	0.773
<i>Vairimorpha necatrix</i>	0.661	0.600	0.721	0.689	0.683	0.576
<i>C. elegans</i>	0.558	0.584	0.570	0.613	0.624	0.552
<i>Emericella nidulans</i>	0.657	0.584	0.692	0.681	0.741	0.539
<i>Nicotiana tabacum</i>	0.913	0.705	0.917	0.734	0.938	0.787
<i>Cryptomonas.sp</i>	0.926	0.676	0.935	0.730	0.941	0.728
<i>Synechococcus.sp</i>	0.938	0.700	0.943	0.740	0.953	0.769
<i>M. musculus</i>	0.608	0.603	0.626	0.655	0.637	0.520
<i>Mycoplasma gallisepticum</i>	0.919	0.639	0.933	0.713	0.932	0.668
<i>E. coli</i>	0.924	0.699	0.937	0.742	0.938	0.774
<i>Bacillus subtilis</i>	0.973	0.698	0.979	0.731	0.976	0.788
<i>Desulfovibrio desulfuricans</i>	0.926	0.712	0.940	0.741	0.938	0.803
<i>Chlamydomonas reinhardtii</i>	0.906	0.687	0.915	0.725	0.928	0.761
<i>Thermotoga maritima</i>	0.931	0.752	0.944	0.760	0.943	0.864
<i>Thermoproteus tenax</i>	0.818	0.782	0.845	0.785	0.866	0.894
<i>H. volcanii</i>	0.782	0.739	0.809	0.769	0.841	0.820
Average	0.820	0.676	0.839	0.719	0.853	0.726
Total	0.839	0.684	0.858	0.729	0.873	0.742

Table of accuracy, PPV, and sensitivity for our LSTM-based state inference model vs. an order 4 HMM. Sequences are arranged in ascending order of MFE accuracy as an indication of the difficulty of secondary structure inference for each sequence. Average indicates the average metric for each sequence, while Total gives the total metrics for all nucleotides in the test set.

(below 70%) for five sequences, medium (near 80%) for two more, and high (above 90%) for the remaining nine.

As expected, state inference accuracy exhibits a strong effect on predicted state directed MFE accuracy. State inference accuracy above 90% means that our predicted states are quite close to native states; consequently, predicted state direction and native state direction (equivalent to 100% accuracy) should produce similar predicted structures in these cases, as evidenced by their difference of only a few percentage points in Table 1. Meanwhile, the five sequences with poor state inference accuracy are exactly those where predicted state directed MFE accuracy is below 40%.

The effect of state inference accuracy is particularly evident when considering the improvement over undirected MFE accuracy: for four of the five sequences with poor accuracy (all but *V. necatrix*), predicted state directed MFE accuracy is within 6 percentage points of undirected MFE accuracy. For *V. necatrix* and both sequences with medium state inference accuracy, predicted state direction improves structure accuracy by 10-15 percentage points. The remaining nine sequences all have high state inference accuracy, and their directed structures are 30 percentage points more accurate than undirected MFE.

We note an interesting relationship between native state directed MFE accuracy and our neural network's state inference accuracy. The five sequences with state inference accuracy below 70% are the five worst performing sequence when predicting structure with native state directed NNTM. This suggests that there may be fundamental difficulties in understanding pairing structures of these sequences.

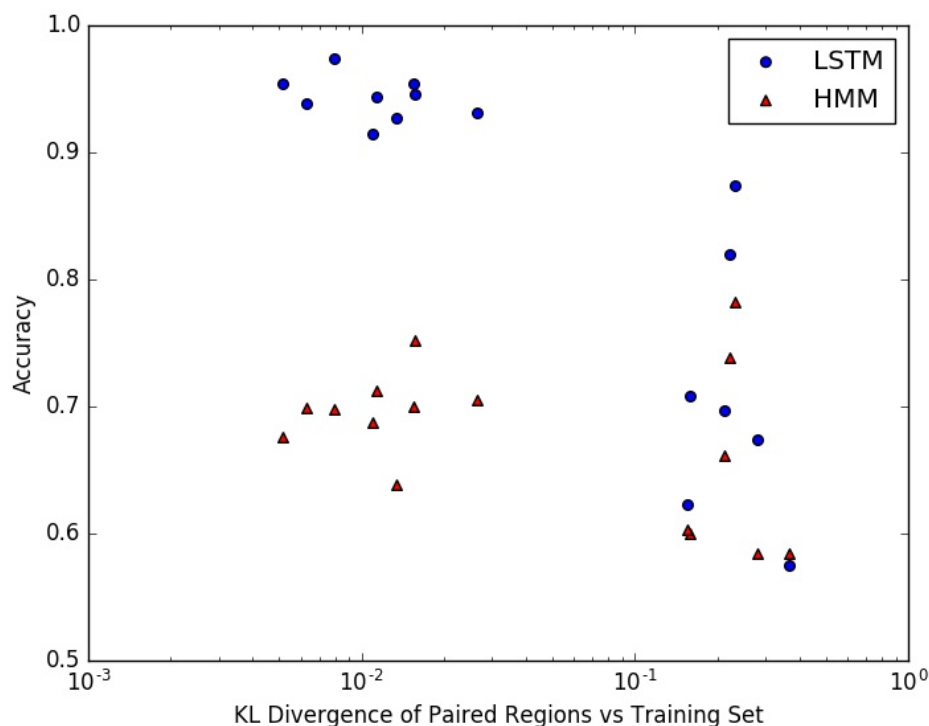


Figure 2: Test Set KL Divergence. Plot comparing each test set sequence's LSTM (neural network) and HMM state inference accuracy vs its Kullback-Leibler divergence from training set paired region distribution. KL divergence was calculated as $KL(P||Q)$, where P is the test sequence distribution and Q is the training set distribution.

4.2 Paired Regions & Global Structure

Our metrics in Table 3 give us an idea of the proportion of correct machine predictions on individual nucleotides' states, but they do not indicate whether predictions produce state sequences that preserve global properties, such as patterns of paired and unpaired states. In particular, we want the number and sizes of paired and unpaired regions of the state sequence prediction to match those in the original. A paired region in the state roughly corresponds to one half of a helix in the secondary structure, so we theorize that recognizing this information is vital for producing state predictions that successfully aid structure inference.

We considered the distribution of sizes of paired regions in each test set state sequence, and compared them to the distributions of neural network and HMM state predictions. Despite larger variance in state inference accuracy, we found evidence that the neural network was, on average, much more capable than the HMM of capturing this global structure. The median size of paired region in neural network predicted state differed from the median in the native state by at most one for every test set sequence, while the HMM's median paired region size was routinely several nucleotides larger. The neural network also performs better in predicting the total number of paired regions in the state, producing predictions that, on average, had 6 more paired regions than the native state, while HMM predictions had an average of 57 fewer regions.

We note that this discrepancy is to be expected in the context of nonlocal interactions. Paired region size is exactly the sort of nonlocal feature that HMMs cannot predict: at a given time, the HMM does not know how long it has been outputting positive predictions, and is thus limited in its capacity to detect large paired regions.

Considering the non-locality of paired regions can help to explain the poor performance of the neural network on certain test set sequences. High neural network accuracy is nearly always accompanied by a

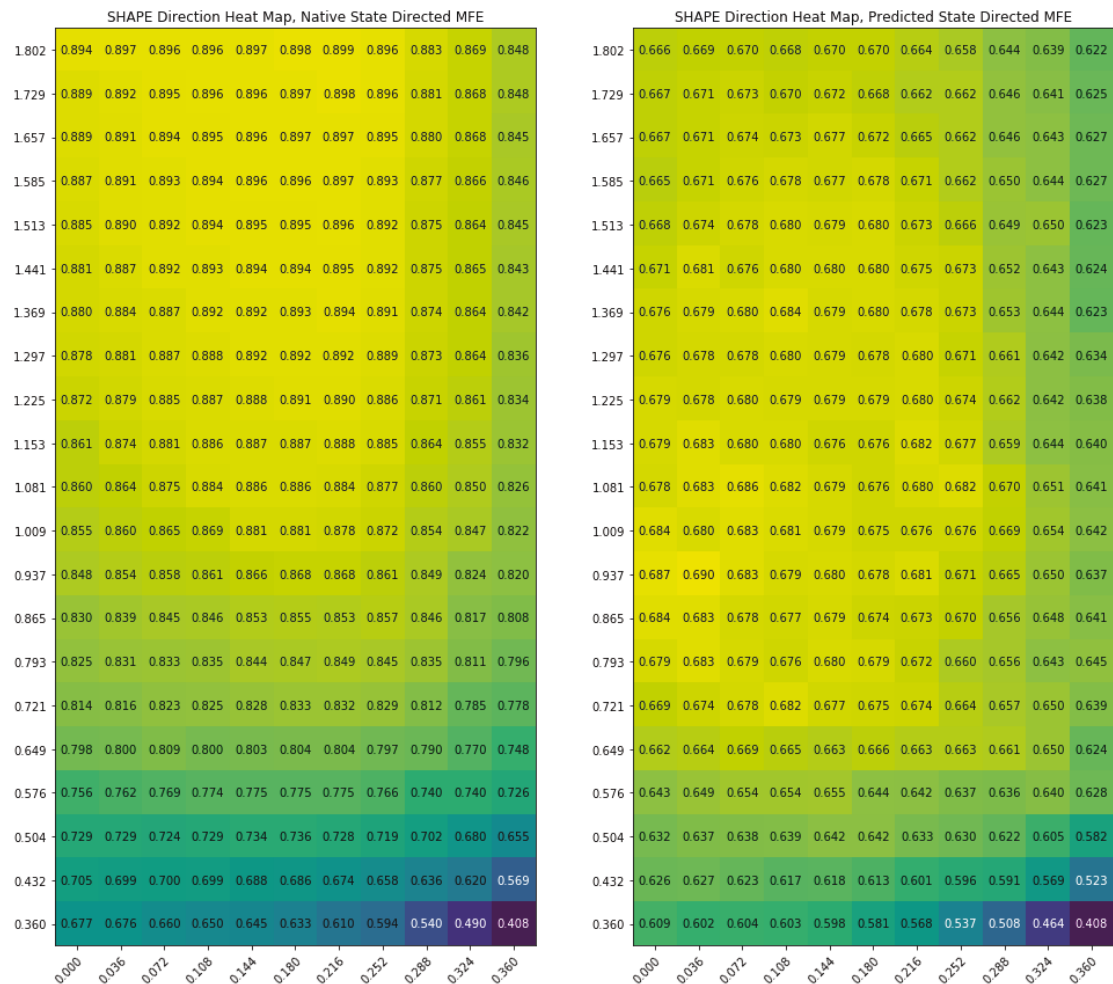


Figure 3: SHAPE Direction Heat Maps. Accuracy of native state directed MFE (left plot) and predicted state directed MFE (right plot) for various ranges $[a, b]$ of output from our SHAPE generation function. In each, the lower right corner corresponds to $a = b = 0.3603$, which is equivalent to no SHAPE direction.

particular type of distribution of large paired regions: one of length 17, one of length 13, and several more of length 12 and 11. In contrast, this pattern does not hold for those with low or medium state inference accuracy of the remaining seven, all have either paired regions of length larger than 20 or very few paired regions of length larger than 10.

We can compare the distribution of the lengths of paired regions in each of our test sequences to the distribution in the training set. We find that the training set overwhelmingly contains sequences with paired region distributions similar to the test set sequences on which the neural network performs well. In particular, we note that the training set has relatively few large paired regions: in the entire training set, there are 5 regions of length 18, 2 regions of size 19, 4 regions of size 20, and none larger than 20. Thus, during training the machine is penalized for outputting more than 20 contiguous positive predictions. Consequently, neural network predictions do not create sufficiently large regions for many test set sequences.

To quantify this difference, we considered the Kullback-Leibler (KL) divergence of the distribution of the paired region lengths between the entire training set and the distribution for each test set sequence. The KL divergence measures the similarity of each test set sequence's paired region distribution as compared with the distribution of the entire training set. Figure 2 plots state inference accuracy for each machine and test set sequence against its KL divergence.

Two clusters of sequences emerge in this plot: one with KL divergence near 0.01, and another with KL divergence near 0.5. All nine sequences with high state inference accuracy are in the former cluster, while the seven low and medium accuracy sequences are in the latter. The disparity in neural network accuracy and HMM accuracy on sequences with more similarity to the training set suggests that the increase in neural network performance comes from its ability to recognize global structure in these sequences. On the other hand, neural network accuracy is only a modest improvement from HMM accuracy in the low to medium accuracy cluster, where global structure diverges significantly from that of the training set.

4.3 Modifying Synthetic SHAPE Values

Our method uses SHAPE-directed NNTM as a means of assigning pseudo-free energies to individual nucleotides. All of our results in Section 3 assign nucleotides a SHAPE value in the range $[0.214, 0.6624]$. These endpoints are based on the mean SHAPE value of paired and unpaired nucleotides from 16S and 23S *E. coli* sequences. However, our method converts to SHAPE primarily as a means of assigning pseudo-free energies to individual nucleotides with NNTM, and not as a genuine attempt to generate plausible SHAPE data. Thus, the endpoints used may not be optimal for our purposes of converting from state inference predictions.

To evaluate potential output ranges for our SHAPE generation function, we reproduced experiments with native state directed NNTM (Section 3.1) and predicted state directed NNTM (Section 3.2) while varying the endpoints a and b of our SHAPE generation function, given in Equation 2. As noted previously, a SHAPE value of 0.3603 contributes no energy to the model; thus, it is only sensible to choose paired SHAPE values below 0.3603, and unpaired SHAPE values above 0.3603. NNTM software such as GTFold ignores negative SHAPE values, so paired nucleotides' generated SHAPE must lie between 0 and 0.3603. The results of this experiment are shown in Figure 3.

In native state directed NNTM, increasing negative state SHAPE above 0.3603 and decreasing positive state SHAPE below 0.3603 consistently increased performance. This is consistent with our expectations, as in this case we are increasing the energy of all base pairs involving nucleotides that remain unpaired in the native structure. Experiments with very large unpaired SHAPE values, such as $b = 20$, were similar to the largest values shown in Figure 3, indicating that there is a ceiling of approximately 90% test set accuracy for any method centered around SHAPE-directed NNTM such as ours.

The plot for predicted state directed NNTM shows a different picture, with increasing unpaired SHAPE values eventually leading to decreasing structure inference accuracy. That this pattern appears in the predicted state experiments but not native state experiments suggests that incorrectly assigning large SHAPE values to even a small number of natively paired nucleotides can be significantly harmful to NNTM performance. There is a large region of highest accuracy, with a between 0 and 0.22 and b between 0.7 and 1.5 giving accuracies near 68%. The values of $a = 0.214$ and $b = 0.6624$ used in our results are near the boundary of this region. But we note that even optimal values of a and b give an accuracy of 69%, only 2.5 percentage points above the experimentally motivated choices of a and b used in our results.

4.4 Other RNA Types

The dependence on large amounts of data inhibited our ability to extend this work to inference on other types of RNA, such as 5S and 23S ribosomal RNA, as we were unable to amass enough secondary structure data in these other contexts to successfully train a neural network. Instead, we explored applying our trained network to other RNA sequences, but found that a neural network trained on 16S RNA sequences produces much weaker results on 5S and 23S rRNA sequences than on 16S sequences, see Tables 4-5. We suspect this is primarily due to differences in sequence length. Tables 4-5 also compare neural network state predictions with those from an HMM, and shows that the HMM is more capable of generalizing to these RNA families. However, we note that in all cases, HMM performance is poor relative to neural network performance on 16S

sequences. This is consistent with our hypothesis that our neural network is recognizing long-range dependencies specific to the family of 16S sequences that HMMs are inherently unable to capture.

Table 4: State inference results on 5S sequences

Sequence Name	Accuracy		PPV		Sensitivity	
	LSTM	HMM	LSTM	HMM	LSTM	HMM
d.5.e.S.pombe.1.no	0.647	0.605	0.735	0.685	0.676	0.676
d.5.e.P.waltl.no	0.500	0.717	0.635	0.756	0.446	0.797
d.5.e.O.sativa.1.no	0.487	0.756	0.589	0.808	0.581	0.797
d.5.e.M.glyptostroboides.no	0.467	0.700	0.586	0.757	0.460	0.757
d.5.e.M.fossilis.no	0.442	0.675	0.571	0.733	0.378	0.743
d.5.b.M.luteus	0.516	0.650	0.700	0.709	0.449	0.782
d.5.b.M.genitalium	0.466	0.636	0.622	0.754	0.319	0.597
d.5.e.L.edodes	0.583	0.658	0.682	0.709	0.608	0.757
Average	0.514	0.675	0.640	0.739	0.490	0.738

Table of accuracy, PPV, and sensitivity for our LSTM-based state inference model and an order 4 HMM for 5S sequences. Average indicates the average metric for each sequence.

Table 5: State inference results on 23S sequences

Sequence Name	Accuracy		PPV		Sensitivity	
	LSTM	HMM	LSTM	HMM	LSTM	HMM
d.233.a.H.marismortui.1	0.640	0.682	0.647	0.709	0.840	0.771
d.233.a.T.celer	0.666	0.745	0.666	0.729	0.854	0.892
d.233.m.S.sinuspaulianus	0.541	0.632	0.407	0.478	0.678	0.543
PDB_00784	0.652	0.669	0.714	0.764	0.803	0.732
PDB_00616	0.602	0.655	0.595	0.655	0.829	0.767
PDB_00953	0.595	0.663	0.567	0.622	0.868	0.856
PDB_00503	0.612	0.658	0.610	0.668	0.830	0.760
PDB_00846	0.642	0.664	0.709	0.763	0.793	0.725
PDB_00776	0.552	0.602	0.487	0.524	0.860	0.770
Average	0.611	0.663	0.600	0.657	0.817	0.757

Table of accuracy, PPV, and sensitivity for our LSTM-based state inference model and an order 4 HMM for 23S sequences. Average indicates the average metric for each sequence.

5 Conclusion

We introduced a method for secondary structure inference by connecting a deep learning method for state inference with previous work that leverages SHAPE-directed NNTM for structure inference. We tested this method on a set of 16S rRNA sequences with a wide range of MFE accuracies, and found large improvements over undirected MFE structures in most cases. These gains were not uniform throughout the test set, as several

sequences with low undirected MFE accuracy experienced very little increase from predicted state direction. However, median increase in accuracy was more than 30 percentage points, and in the best case our method improved undirected NNTM accuracy by nearly 50 percentage points.

Experiments using sequences' native states to predict structure showed that a data-directed NNTM approach has the potential to improve MFE accuracy throughout the test set. However, it also uncovered significant limitations; regardless of the accuracy of the state inference method used to supply predictions, directed NNTM will not be able to produce high accuracy MFE structures for some sequences. These findings reinforce results from [30] regarding the varying directability of sequences in the test set.

The performance of our state inference method was highly variable among test set sequences, with several clusters of high and low accuracy. We found that high performance of our state inference method (and, in turn, the accuracy of secondary structures generated from these state predictions) was strongly linked to the similarity between a sequence's paired region distribution and that of the training set.

This finding highlights the connection between performance of our method and available secondary structure data. As with any application of machine learning, the state inference method presented here is only as good as the dataset used to train the model. The task of 16S rRNA state inference explored here is feasible primarily due to the particularly large number of known 16S rRNA secondary structures. We are hopeful that future increases in available data are able to increase the efficacy of this method, both on 16S rRNA sequences as well as on 5S and 23S rRNA.

Availability of data and materials: The sequence and secondary structure data used to train the neural network and HMMs is available from the Comparative RNA Web (CRW) site repository, <https://doi.org/10.1186/1471-2105-3-2>. This data, along with the native secondary structure data for the test set, is available at <http://ms.uky.edu/~dwi239/rnastateinf-data.zip>.

Acknowledgements: We would like to thank John Hirdt, who had initially participated in this project, for many valuable discussions and suggestions. DM thanks Christine Heitsch for introducing him into research on SHAPE-directed RNA structure prediction. The research of Qiang Ye is supported in part by NSF under DMS-1821144 and DMS-1620082.

References

- [1] Jamie J Cannone, Sankar Subramanian, Murray N Schnare, James R Collett, Lisa M D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi V Madabusi, Kirsten M Müller, et al. The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC bioinformatics*, 3(1):2, 2002.
- [2] Jamie J. Cannone, Sankar Subramanian, Murray N. Schnare, James R. Collett, Lisa M. D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi V. Madabusi, Kirsten M. Müller, Nupur Pande, Zhidi Shang, Nan Yu, and Robin R. Gutell. The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC Bioinformatics*, 3(1):2, 2002.
- [3] Jonathan L Chen, Stanislav Bellaousov, and Douglas H Turner. Rna secondary structure determination by nmr. *Methods Mol Biol*, 1490:177–86, 2016.
- [4] François Chollet et al. Keras, 2015.
- [5] Katherine E Deigan, Tian W Li, David H Mathews, and Kevin M Weeks. Accurate shape-directed rna structure determination. *Proc Natl Acad Sci U S A*, 106(1):97–102, Jan 2009.
- [6] Laura DiChiacchio, Michael F Sloma, and David H Mathews. Accessfold: predicting rna–rna interactions with consideration for competing self-structure. *Bioinformatics*, 32(7):1033–1039, 2015.
- [7] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
- [8] Sean R Eddy. Computational analysis of conserved rna secondary structure in transcriptomes and genomes. *Annu Rev Biophys*, 43:433–56, 2014.
- [9] Boris Fürtig, Christian Richter, Jens Wöhnert, and Harald Schwalbe. Nmr spectroscopy of rna. *ChemBioChem*, 4(10):936–962, 2003.

- [10] Paul P Gardner and Robert Giegerich. A comprehensive comparison of comparative rna structure prediction approaches. *BMC Bioinformatics*, 5:140, Sep 2004.
- [11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [12] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.
- [13] Robin R Gutell, Jung C Lee, and Jamie J Cannone. The accuracy of ribosomal rna comparative structure models. *Curr Opin Struct Biol*, 12(3):301–10, Jun 2002.
- [14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
- [15] Risa Kawaguchi, Hisanori Kiryu, Junichi Iwakiri, and Jun Sese. reactidr: evaluation of the statistical reproducibility of high-throughput structural analyses towards a robust rna structure prediction. *BMC Bioinformatics*, 20(Suppl 3):130, Mar 2019.
- [16] Wan-Jung C Lai, Mohammad Kayedkhordeh, Erica V Cornell, Elie Farah, Stanislav Bellaousov, Robert Rietmeijer, Enea Salsi, David H Mathews, and Dmitri N Ermolenko. mrnas and lncrnas intrinsically form secondary structures with short end-to-end distances. *Nat Commun*, 9(1):4328, 10 2018.
- [17] D M Layton and R Bundschuh. A statistical analysis of rna folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res*, 33(2):519–24, 2005.
- [18] S Y Le, J H Chen, and J V Maizel, Jr. Prediction of alternative rna secondary structures based on fluctuating thermodynamic parameters. *Nucleic Acids Res*, 21(9):2173–8, May 1993.
- [19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [20] Mirko Ledda and Sharon Aviran. Patterna: transcriptome-wide search for functional rna elements via structural data signatures. *Genome Biology*, 19(1), Mar 2018.
- [21] Thomas J X Li and Christian M Reidys. The rainbow spectrum of rna secondary structures. *Bull Math Biol*, 80(6):1514–1538, 06 2018.
- [22] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms Mol Biol*, 6:26, Nov 2011.
- [23] Nicholas R Markham and Michael Zuker. Unafold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, 453:3–31, 2008.
- [24] David H Mathews and Douglas H Turner. Prediction of rna secondary structure by free energy minimization. *Curr Opin Struct Biol*, 16(3):270–8, Jun 2006.
- [25] Jessica S Reuter and David H Mathews. Rnastructure: software for rna secondary structure prediction and analysis. *BMC Bioinformatics*, 11:129, 2010.
- [26] Emily Rogers, David Murrugarra, and Christine Heitsch. Conditioning and robustness of rna boltzmann sampling under thermodynamic parameter perturbations. *Biophysical Journal*, 113(2):321–329, 2017.
- [27] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.
- [28] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [29] Alina Selega, Christel Sirocchi, Ira Iosub, Sander Granneman, and Guido Sanguinetti. Robust statistical modeling improves sensitivity of high-throughput rna structure probing experiments. *Nat Methods*, 14(1):83–89, 01 2017.
- [30] Zsuzsanna Sükösd, M Shel Swenson, Jørgen Kjems, and Christine E Heitsch. Evaluating the accuracy of shape-directed rna secondary structure predictions. *Nucleic Acids Res*, 41(5):2807–16, Mar 2013.
- [31] M Shel Swenson, Joshua Anderson, Andrew Ash, Prashant Gaurav, Zsuzsanna Sukosd, David A Bader, Stephen C Harvey, and Christine E Heitsch. Gtfold: Enabling parallel rna secondary structure prediction on multi-core desktops. *BMC Res Notes*, 5(1):341, Jul 2012.
- [32] Hakim Tafer, Fabian Amman, Florian Eggenhofer, Peter F Stadler, and Ivo L Hofacker. Fast accessibility-based prediction of rna-rna interactions. *Bioinformatics*, 27(14):1934–40, Jul 2011.
- [33] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.
- [34] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.
- [35] Douglas H Turner and David H Mathews. Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, 38(Database issue):D280–2, Jan 2010.
- [36] Stefan Washietl, Ivo L Hofacker, Peter F Stadler, and Manolis Kellis. Rna folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res*, 40(10):4261–72, May 2012.
- [37] Kevin A Wilkinson, Robert J Gorelick, Suzy M Vasa, Nicolas Guex, Alan Rein, David H Mathews, Morgan C Giddings, and Kevin M Weeks. High-throughput shape analysis reveals structures in hiv-1 genomic rna strongly conserved across distinct biological states. *PLoS Biol*, 6(4):e96, Apr 2008.
- [38] Devin Willmott. *Recurrent Neural Networks and Their Application to RNA Secondary Structure Inference*. PhD thesis, University of Kentucky, 8 2018.