RNA BIOLOGY
2020, VOL. 17, NO. 9, 1324–1330
https://doi.org/10.1080/15476286.2020.1760534

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

# Predicting RNA SHAPE scores with deep learning

Noah Bliss[a]*, Eckart Bindewald[b]*, and Bruce A. Shapiro[a]

[a]RNA Biology Laboratory, National Cancer Institute, Frederick, MD, USA; [b]Basic Science Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA

**ABSTRACT**

Secondary structure prediction approaches rely typically on models of equilibrium free energies that are themselves based on in vitro physical chemistry. Recent transcriptome-wide experiments of in vivo RNA structure based on SHAPE-MaP experiments provide important information that may make it possible to extend current in vitro-based RNA folding models in order to improve the accuracy of computational RNA folding simulations with respect to the experimentally measured in vivo RNA secondary structure. Here we present a machine learning approach that utilizes RNA secondary structure prediction results and nucleotide sequence in order to predict in vivo SHAPE scores. We show that this approach has a higher Pearson correlation coefficient with experimental SHAPE scores than thermodynamic folding. This could be an important step towards augmenting experimental results with computational predictions and help with RNA secondary structure predictions that inherently take in-vivo folding properties into account.

**ARTICLE HISTORY**
Received 30 December 2019
Revised 22 March 2020
Accepted 24 March 2020

**KEYWORDS**
SHAPE; RNA; secondary structure; SHAPE-MaP; deep learning; neural network

## Introduction

### *Importance of RNA Structure*

Local RNA structures can have important biological functions. For example, there are multiple reported cases of bacterial RNA-based riboswitches that change their conformation in reaction to the interaction with a small molecule metabolite potentially implementing a regulatory feedback mechanism important for homoeostasis [1].

It has been reported that cellular mRNAs are less structured compared to in vitro conditions as what theoretical secondary structure predictions would suggest [2–4]. The prevalence of this effect it seems can be influenced by ATP-depletion, suggesting that mRNA unfolding is subject to controlled cellular processes. Possible reasons for this are that RNA-binding proteins prevent non-specific RNA agglomeration; also translation by ribosomes and their helicase activity contribute to unfolding of mRNA coding regions. Because RNA structure can act as a regulator but is itself being regulated by a vast number of different factors such as RNA-binding proteins and metabolites, it will remain challenging to model in vivo RNA folding.

### *The promise and challenge of RNA SHAPE experiments*

Selective 2′-Hydroxyl Acetylation analysed by Primer Extension (SHAPE) is an important experimental approach for determining whether an RNA nucleotide participates in RNA base pairing [5]. The key idea is that single-stranded unpaired nucleotides exhibit a higher structural flexibility compared to base-paired nucleotides. This difference in structural flexibility and accessibility contribute to a difference in chemical reactivity of different nucleotides with respect to a chemical reagent. In the classical SHAPE approach, the adduct of the chemical reagent blocks elongation of reverse transcriptase at the modified base [5]. Each transcript can therefore harbour only one chemical adduct that contributes to the measured outcome. In contrast, the SHAPE-MaP methodology is adjusted to take full advantage of high throughput RNA-Seq experiments. Instead of blocking and terminating the reverse transcription, mutational profiling (MaP) is utilized in order to induce non-complementary nucleotide mutations corresponding to the chemically modified nucleotides [6]. RNAs can therefore be reverse-transcribed into cDNA in their entirety even if they contain multiple sites that lead to the adduction of the chemical reagent.

Recently, the SHAPE-MaP approach has been applied to *E. coli* mRNAs [7]. In that study, results corresponding to 3 different conditions (in vivo, in vitro, and in vivo with blocked co-translation) were generated and analysed. The authors found many examples of unstructured and structured regions on mRNAs. There is a tendency for less RNA structure in vivo compared to an in vitro environment; blocking translation with an antibiotic has the influence of additional measured RNA structure in coding regions. The variety of different RNA structures encountered in that study prompted the authors to coin the term ‘RNA personalities’.

Because of the many possible influencing factors and the apparent lack of simple reoccurring sequence-structure patterns, building a theoretical or computational model for

**CONTACT** Bruce A. Shapiro shapirbr@mail.nih.gov RNA Biology Laboratory, National Cancer Institute, Frederick, MD 271702, USA

*These authors contributed equally to this work.

in vivo mRNA structure is challenging. The accuracy of RNA secondary structure prediction approaches based on equilibrium thermodynamics can be increased if they incorporate experimental constraints [8].

### Utility of neural networks

Many of the fundamental ideas of machine learning in general and neural networks in particular were reported decades ago, where it was shown that one can solve difficult pattern recognition problems not by explicitly programming a solution but by ‘showing’ positive and negative examples to a learning algorithm. Much more recently, however, a number of important innovations dramatically expanded the scope of neural networks.

Convolutional neural networks, for example, were a breakthrough in image processing. Convolutional neural networks are based on the idea that the network self-learns a number of local features that are generated for each position of the input image. The resulting array of feature vectors is input for subsequent neural network layers. potentially leading to more complex and abstract features that are ultimately used to generate the final prediction output [9].

Furthermore, innovations in image processing dealt with the problem of scale. Object recognition in images is an inherently scale-invariant task, where translation, rotation and ‘zoom’ of an object in an image should not substantially alter the prediction result. One approach to implement this are ‘inception’ type models, whose main idea is that if the scale (i.e. the size) of an object in the input vector is not initially known, one can still obtain an approximation of scale invariance with different neural network layers that each specialize in detecting an object of a specific size range [10]. Subsequent neural network layers can use the output of the different feature vectors as input and easily detect if an object was detected for a least one particular size.

The current paper utilizes these methodologies and applies them in a novel way to the problem of predicting SHAPE scores of RNA sequences.

### Prediction of RNA SHAPE scores

RNA secondary structure depends on its RNA sequence as well as the biochemical environment. Traditionally, RNA secondary structure is predicted using thermodynamic models and algorithms that identify the minimum free energy conformation. Recently, an approach based on deep neural networks has been reported [11]. It is important to note that neural network architectures need to be of sufficient sophistication in order to capture the complexities of RNA secondary structure. Taken together it is apparent that predicting RNA SHAPE experiment values from local sequence information is an attractive target for machine learning approaches. Recently, Mautner described an approach based on a tree-based machine learning approach where predicted local RNA structure is encoded as a graph [12]. In contrast, we are presenting here a deep neural network approach were the unique network architecture is designed to capture properties of local sequence, secondary structure and SHAPE scores.

## Methods

### Data

The data used in this paper was obtained from the supplementary information of [7]. They provide data sets of experimental shape scores for 194 RNAs (each RNA possibly harbouring multiple ORFs) for 3 different conditions (in vivo, in vitro, in vivo with inhibited translation). Each of these 3 data sets was divided into testing and training datasets corresponding to 32 and 162 RNAs respectively (Table 1). Note that each nucleotide position of the respective RNA sequences is converted to a feature vector used as input for the machine learning approach, leading to a test set of 51578 feature vectors and a training set of 390843 feature vectors.

For preprocessing, the original SHAPE data combined with the corresponding RNA sequence was converted into an R data structure, augmented with a structure prediction from RNAplfold but not otherwise modified. The utilized command line parameters for the thermodynamic prediction were: RNAplfold -T 37 -p. Note that the machine learning model has for its training SHAPE data a flag that indicates missing data (‘NA’). In other words, cases of RNAs and their SHAPE values can be processed for both training and testing even if the SHAPE data has numerous missing values.

### Neural network model

The neural network models used in the paper are based on the Keras framework in R with a TensorFlow back-end [13,14]. We want the neural network to detect RNA secondary structures and motifs. One challenge is that one does not know ahead of time how ‘large’ a secondary structure or sequence motif is. This problem is known in image processing in the form of challenges where the task is to detect objects in an

**Table 1.** Partition of IDs of 194 RNAs reported by Mustoe into test set and training set.

| Data set | IDs |
|---|---|
| Test set | 7,11,21,22,24,27,28,38,41,44,45,51,54,59,61,62,73,75,76,94,<br>99,115,127,139,145,151,152,160,168,171,177,179 |
| Training set | 1,2,3,4,5,6,8,9,10,12,13,14,15,16,17,18,19,20,23,25,26,29,30,31,32,33,34,35,<br>36,37,39,40,42,43,46,47,48,49,50,52,53,55,56,57,58,60,63,64,65,66,67,68,69,<br>70,71,72,74,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,95,96,97,98,<br>100,101,102,103,104,105,106,107,108,109,110,111,112,113,114,116,117,<br>118,119,120,121,122,123,124,125,126,128,129,130,131,132,133,134,135,<br>136,137,138,140,141,142,143,144,146,147,148,149,150,153,154,155,156,<br>157,158,159,161,162,163,164,165,166,167,169,170,172,173,174,175,176,<br>178,180,181,182,183,184,185,186,187,188,189,190,191,192,193,194 |

image where the object’s relative size in the image is not known ahead of time. One solution is to utilize parallel units of the machine learning algorithm that are specialized for different specific sizes of the objects to be recognized. In practice, this is often implemented via convolutional neural network layers of different ‘kernel’ sizes (for example a kernel size of 5 means that the neural network only uses a subset of 5 input features from a much larger input vector, and the network is applied to the input data vector in a sliding-window manner). Additionally, stacking convolutional neural network layers with small kernel sizes reduces the number of parameters compared to a convolutional layer with a large kernel size. This is the approach utilized in this work. A schematic of the created model is shown in Fig. 1. One can see that the modules shown in red, green and purple all contain convolutional network layers of different kernel sizes. Also, these modules use parallel routes of information flow with different numbers of stacked convolutional layers. These alternative routes are in subsequent layers combined in order to predict a single number (the SHAPE score of a particular nucleotide). The primary model was based off of the inception model used in computer vision algorithms [10]. Inception models were designed to solve the issue where the salient part of an image varied dramatically in size, making it difficult to create a model which is well equipped for all scales of images. The inception networks solved this by having several different kernel sizes in parallel allowing for the network to be optimized for several different scale ranges.

To train a neural network, a large volume of data is frequently needed. The independent variable of this model is the sequence information of the RNA transcript around a specific nucleotide and the dependent variable is the SHAPE score of the nucleotide. The chosen loss function is the mean-squared-error (mse) corresponding to the average squared differences of predicted and actual SHAPE values of the training data. Using this loss function as an objective function, the training procedure minimizes the differences between predicted SHAPE scores and experimentally determined SHAPE scores utilizing each position of the sequences in the training set.

The model takes in six different inputs per nucleotide, an upstream sequence, a downstream sequence, the nucleotide of which you are trying to predict the SHAPE score of, an upstream thermodynamic matrix, a downstream thermodynamic matrix, and a thermodynamic vector corresponding to the nucleotide. The thermodynamic matrix is generated by summing for each nucleotide predicted nucleotide-base pair probabilities generated by the program RNAplfold [15]. Previously we found that local folding predictions using RNAplfold predictions (with a default window size of 70nt) lead to a higher Pearson correlation coefficient with the experimental SHAPE scores compared to global folding predictions as generated by RNAfold.

As shown in Fig. 1, the model architecture uses a variety of neural network layers that can be partitioned into four different ‘blocks’: one block for sequence input, one for thermodynamic input, and two for the combined outputs of the

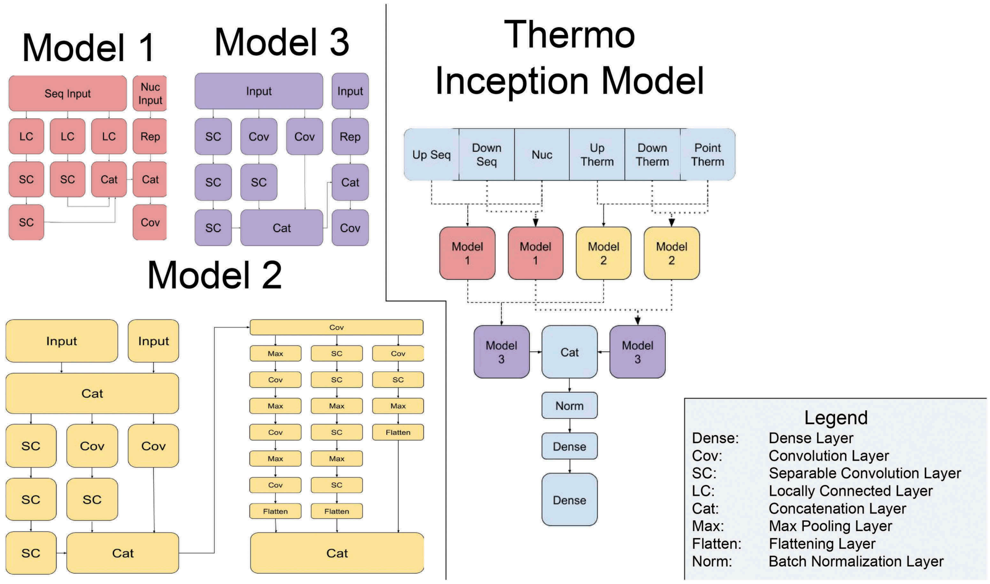

**Figure 1.** Diagram of the thermo inception model that was designed for this paper. The different sections of the model are colour coded. Each block represents a layer of the network and each arrow shows the flow of information from one layer to the next. The dashed and dotted lines running in between sections represent how the model was run on both the upstream (5′ most) data and the downstream (3′-most) data, with one represented by the dashed and the other by the dotted respectively. As described in the Methods section, the input data is processed via convolutional networks of different kernel sizes in order to be able to detect RNA secondary structures and sequence motifs of different sizes (sections highlighted in red, yellow, and purple). The upstream and downstream data are combined in a stacked set of dense neural networks. This model also utilizes several different types of neural network layers that act similar to traditional convolutional layers. The separable convolution layer is a special subset of convolutional layers where the kernel is broken up into two separate convolutions. This reduces the overall number of parameters. The locally connected layer acts as an intermediate between convolutional layers and dense layers. Corresponding nodes between layers are fully connected with their neighbors, much like the sliding-window method used by the convolutional layers. There are several other helping layers, such as the repeating layer which extrudes a vector or tensor of a lower dimension into a tensor of a dimension higher. The max pooling layer reduces the dimension of the tensor and takes the largest value in a specified axis of the input tensor. The batch normalization layer augments the tensor to have a mean of zero and a standard deviation of one.

previous blocks. Each block applies several types of different convolutional neural network layers to its input. As a final operation of each block, a point-wise convolution between the intermediate output generated by the block's convolutional layers and the unmodified point input is computed. The sequence and structure input are processed in this fashion in two separate blocks (referred to as 'Model 1' and 'Model 2' shown in red and yellow in Fig. 1, respectively). Next, another block ('Model 3' shown in purple in Fig. 1) uses as input the output from the two aforementioned blocks and computes an output vector that combines both sequence and structure information. Lastly, a chain of dense layers is used to combine the high-dimensional data vector generated by Model 3 to predict the SHAPE score of a nucleotide in the form of a single number.

### *Hyper-parameter optimization*

Not all parameters of a neural network can be optimized by gradient descent, such as the number of layers or the number of nodes in a layer. These layers are often referred to as hyper-parameters, and it is difficult to find the optimal hyper-parameters for a model. In this paper we created a grid search of the hyper parameter space to find the optimal configuration of parameters that optimizes the prediction performance of the model with respect to the training set of RNA strands. The hyper parameter optimizer was used to find the ideal configuration of nine different hyper parameters: number of training epochs, batch size of the training data, the loss function used, the number of parallel branches of the inception blocks, the length of the branches of the inception blocks, the number of units in the dense layers of the regressor block, the number of layers in the regressor block, and the amount of dropout between layers of the regressor block.

Several thousand models were generated, and each were trained on their own GPU node on the NIH Biowulf cluster, then evaluated on the same set of data and the model with the highest Pearson correlation coefficient was chosen.

## Results

We applied the machine learning model on the transcript sequences provided by Mustoe [7]. Note that while training the complex model is very memory intensive (we found that up to 250GB RAM were required), the application of the trained model is reasonably fast and in need of only modest memory (< 7GB RAM required).

Fig. 2 depicts the Pearson correlation coefficients of predicted SHAPE scores as well as scores derived from the secondary structure prediction program called RNAplfold [15]. Also shown are the Pearson correlation coefficients between base pairing probabilities derived from the predicted secondary structures provided by Mustoe and their underlying corresponding SHAPE scores. One can see that the deep learning model leads to a higher Pearson correlation coefficient compared to the thermodynamic prediction provided by RNAplfold. In Fig. 3 absolute differences between predicted

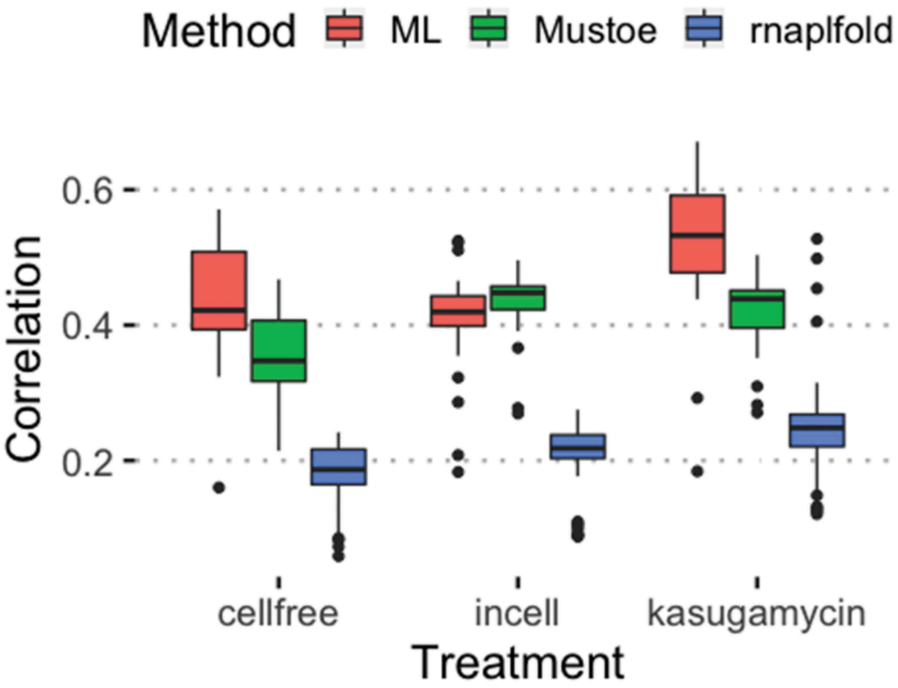

**Figure 2.** Correlations with experimental SHAPE scores for the machine learning method (red, 'ML'), thermodynamic folding (blue, 'rnaplfold') and predicted structures informed by the experimental SHAPE scores (green, 'Mustoe'). The middle portion ['Mustoe') depicts the correlation between the secondary structures probabilities accompanying the [7],publication (corresponding to thermodynamic folding informed by experimental SHAPE scores]. One can see that the machine learning method shown in red (which did not have access to the experimental SHAPE scores for the used test cases) is performing similar or better compared to the predicted structures informed by SHAPE scores [shown in green). The 3 experimental conditions indicated as 'cellfree', 'incell' and 'kasugamycin' correspond to the three dataset provided in [7],of i] cell-free lysates, ii) in-cell conditions and iii) conditions of deactivated protein translation due to the presence of the kasugamycin antibiotics.

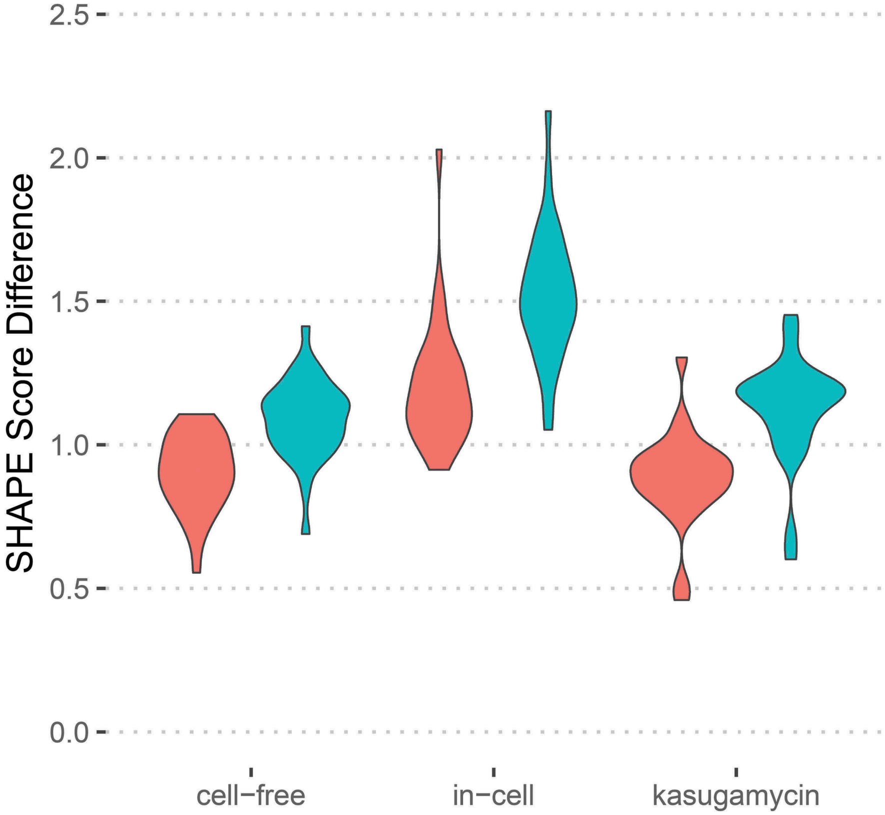

Figure 3. Violin plots of absolute difference between predicted and experimental SHAPE scores (red) compared to control where the correspondence between input feature vector and output has been randomly shuffled (blue).

and experimentally determined SHAPE scores are shown in the form of violin plots.

### *Example: RNA Structure around translation start sites*

As an example of the utility of the approach we plotted the experimental and predicted SHAPE scores around translation start sites of mRNAs (Fig. 4). Also plotted are scores derived from a thermodynamic folding prediction (see Methods). One can see that the SHAPE score predicted with the method described in this paper captures the tendency of nucleotides at and immediately upstream of the translation start sites to be relatively unstructured. Indeed, a statistical test results in statistically significant differences in the position corresponding to the start codon compared to its surrounding positions for the cases of the experimental SHAPE score and SHAPE scores predicted by the machine learning approach (Fig. 4A,B left and middle). In contrast, the differences in the amount of predicted secondary structures for start codons versus their surrounding nucleotides failed to reach statistical significance using the thermodynamic folding approach (Fig. 4A,B right). This suggests that the deep learning method provides added value compared to thermodynamic folding predictions that do not take particular sequence motifs into account.

## Discussion

There is a rich history of attempting to predict RNA secondary structure from its sequence. Typical approaches rely on an equilibrium thermodynamic model that represent RNA base pairing, entropic and stacking effects. A search algorithm would then determine RNA secondary structures corresponding to the minimum free energy of the utilized model. It should be emphasized that such models are inherently representing RNA folding energetics akin to ‘in-vitro’ settings, where potential interactions with other factors, such as proteins are not taken into account.

Building an explicit in-vivo model for RNA folding may be a daunting task, because it might involve explicitly modelling a large number of possible RNA-protein and RNA-RNA interaction events. Building instead a quasi-probabilistic description based on machine learning, trained with known experimental results may instead be tractable. Modelling RNA SHAPE scores have the advantage of providing a rather direct observation of a one-dimensional relationship between sequence and score. Another advantage is that there is due to recent transcriptome-wide experiments sufficient data available to build non-trivial machine learning models [7]. This approach is also attractive because it is, to our knowledge, the first machine learning model that is geared towards SHAPE scores. The model could be used in future machine learning experiments as a constituent of a larger model. Because the model seems to have detected at least some of the novel aspects of in vivo RNA folding, it could be used to inform a larger model about the insights of RNA folding it has learned.

We noticed that the prediction accuracy of the machine learning algorithm is similar for the testing and training data sets. A possible explanation for this observation is that RNA structure (and therefore SHAPE scores) has non-local aspects that go beyond the utilized 121 nucleotide window, for example due to long-range secondary structure base pairing or due to

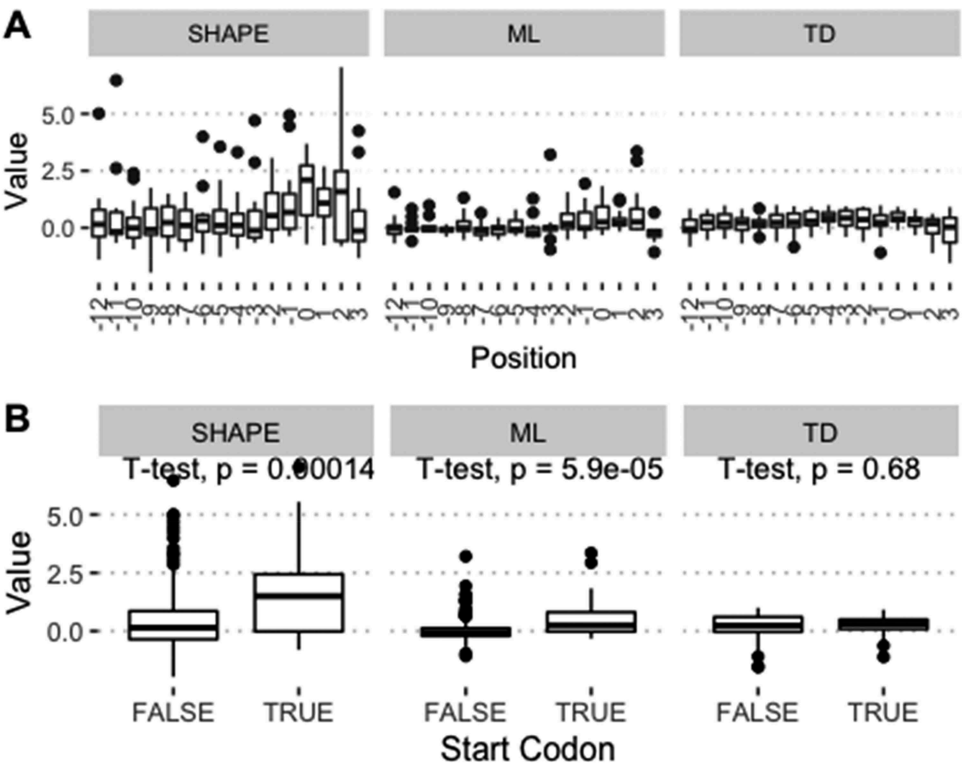

**Figure 4.** Predicted SHAPE scores capture the essence of experimental SHAPE scores around translation start sites. A) The box-whisker plots show experimental SHAPE scores (left), SHAPE score via deep learning (middle) and scores derived from thermodynamic predictions (right, see Methods). For each of the 3 types of scores, the values are plotted with respect to translation start sites of genes (the 'A' of an AUG start codon corresponds to position 0). One can see that the SHAPE scores predicted via our deep learning model capture the essence of scores better compared to scores derived from thermodynamic folding predictions (see Methods). B) Comparing the scores corresponding to the AUG start codon (positions 0,1,2 in panels shown in A) to the remaining positions, one finds that a t-test shows statistically significant differences (indicating that start codon sites are less structured) in the cases of SHAPE scores (SHAPE) and our machine learning approach (ML) but not in the case of thermodynamic folding.

RNA-protein binding events. A second possibility might be that SHAPE scores may contain inherent noise due to peculiarities of the utilized chemical reagent, 3D structure effects or effects due to the noise introduced in the complex SHAPE score normalization procedure. For example, we noticed previously in a study that correlated SHAPE experiment data with known RNA 3D structures, that SHAPE scores corresponding to stacking of adjacent single-stranded nucleotides can have similarities with SHAPE scores corresponding to base paired regions [16]. Since stacking effects are not captured by regular secondary structure data used for training and testing in the current study, they may appear as 'noise' to the machine learning algorithm.

Interestingly, the approach is more successful for the cell-free and kasugamycin data sets as opposed to the in-cell dataset. One interpretation is that in the case of the in-cell system, active translation leads to continuous disruption and melting of local RNA secondary structures in mRNA coding regions. The algorithm does not have an internal representation of whether a coding region or a non-coding region is analysed. Augmenting the machine learning model with additional biological knowledge of this kind should improve the prediction accuracy further and will be part of future work.

## Acknowledgments

We are grateful for the computational resources provided by the NIH Biowulf facility. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Funding

This work was supported in whole or in part with Federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health, under Contract No. HHSN261200800001E. This research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

## Data availability

The R package 'deepshape.ecoli' is available on a BitBucket repository https://bitbucket.org/nossahbli/deepshape.ecoli

## References

[1] Winkler WC, Breaker RR. Regulation of bacterial gene expression by riboswitches. Annu Rev Microbiol. 2005;59:487–517. Annual Reviews.

[2] Leamy KA, Assmann SM, Mathews DH, et al. Bridging the gap between in vitro and in vivo RNA folding. Q Rev Biophys. 2016;49. DOI:10.1017/S003358351600007X

[3] Rouskin S, Zubradt M, Washietl S, et al. Genome-wide probing of Rna structure reveals active unfolding of mRNA

structures in vivo. Nature. 2014;505(7485):701.Nature Publishing Group.

[4] Yiliang D, Tang Y, Kwok CK, et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014;505(7485):696–700.

[5] Wilkinson KA, Merino EJ, Weeks KM. Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc. 2006;1(3):1610.Nature Publishing Group. .

[6] Siegfried NA, Busan S, Rice GM, et al. RNA motif discovery by shape and mutational profiling (SHAPE-MaP). Nat Methods. 2014;11(9):959.Nature Publishing Group.

[7] Mustoe AM, Busan S, Rice GM, et al. Pervasive regulatory functions of mRNA structure revealed by high-resolution shape probing. Cell. 2018;173(1):181–195.Elsevier.

[8] Low JT, Weeks KM. SHAPE-directed RNA secondary structure prediction. Methods. 2010;52(2):150–158.Elsevier.

[9] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86 (11):2278–2324.IEEE.

[10] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition; Boston (MA); 2015. pp. 1–9.

[11] Li G, Zrimec J, Ji B, et al. Performance of regression models as a function of experiment noise. arXiv Preprint arXiv. 2019:1–13. 1912.08141.

[12] Mautner S, Montaseri S, Miladi M, et al. ShaKer: RNA SHAPE prediction using graph kernel. Bioinformatics. 2019;35(14):i354–i359.

[13] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. 2016 Mar 14; 1 – 19.

[14] Chollet F, Allaire JJ. Deep learning with R. Shelter Island (New York): Manning Publications Company; 2018.

[15] Lorenz R, Bernhart SH, Zu Siederdissen CH, et al. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6(1):26.BioMed Central.

[16] Bindewald E, Wendeler M, Legiewicz M, et al. Correlating shape signatures with three-dimensional Rna structures. RNA. 2011;17 (9):1688–1696. Cold Spring Harbor Lab.