Communications in Information and Systems
Volume 19, Number 3, 299–319, 2019

Analytical modeling and deep learning approaches to estimating RNA SHAPE reactivity from 3D structure

Travis Hurst*,†, Yuanzhe Zhou†, and Shi-Jie Chen‡

The selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) chemical probing method provides information about RNA structure and dynamics at single nucleotide resolution. To facilitate understanding of the relationship between nucleotide flexibility, SHAPE reactivity, and RNA 3D structure, we developed an analytical 3D Structure-SHAPE Relationship (3DSSR) method and a predictive convolutional neural network (CNN) model that predict the SHAPE reactivity from RNA 3D structures. Starting from an RNA 3D structure, the analytical model combines key factors into a composite function to predict conformational flexibility of each nucleotide and calculate the correlation between the prediction and experimental SHAPE reactivity. Here, we apply the 3DSSR and the deep learning SHAPE model to SHAPE data-assisted RNA 3D structure prediction. We show that the models provide an effective sieve to exclude 3D structures that are incompatible with experimental SHAPE data. Additionally, we compare the 3DSSR analytical model with the CNN deep learning model that recognizes structural and physical/chemical patterns to predict SHAPE data from RNA 3D structure. Depending on the training data set, the analytical model outperforms the deep learning approach for most test cases, indicating that insufficient data is available to adequately train the CNN at this juncture. For other test cases, the deep learning approach provides better predictions than the analytical model, suggesting that the deep learning approach may become increasingly promising as more SHAPE data becomes available.

*Research supported by NSF Graduate Research Fellowship Program under Grant 1443129.

†These authors contributed equally to the work.

‡Research supported by NIH Grants R01-GM063732 and R01-GM117059.

1. Introduction

Galvanized by recent progress in RNA chemical probing technology, researchers developed efficient, data-driven experimental modeling approaches that place effective constraints on RNA structure to complement established template and physics-based methods [1, 2, 3, 4, 5]. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) provides significant insights into local nucleotide structure and dynamics in RNA [6, 7]. SHAPE reagents are small ligands—such as 1-methyl-7-nitroisatoic anhydride (1M7) [8]—that covalently bind to the 2'-hydroxyl group of a nucleotide (see Fig. 1) [9]. Previous studies [10, 11, 12] suggest that unconstrained nucleotides have a greater ability to sample more conformations and to adopt SHAPE-reactive postures, which causes them to have higher SHAPE reactivity. In contrast, nucleotides that are constrained by base-pairing and stacking interactions have a lower propensity to sample a variety of poses and are much less reactive. By quantitatively measuring local nucleotide dynamics, SHAPE is an effective tool for probing whether a nucleotide is constrained by interactions with other nucleotides (in a helix or structured loop) or is located in a flexible loop/junction, without many interactions. In secondary structure modeling, use of SHAPE data substantially improves accuracy and efficiency [13, 14, 15, 16, 17], where SHAPE reactivity is used to provide additional structural constraints for free-energy based predictions [18]. Moreover, when used as the basis for advanced experimental approaches, such as differential SHAPE reactivity, mutate-and-map, and time-resolved SHAPE chemistry, SHAPE probing provides helpful information for the *in vitro* and *in vivo* determination of non-canonical tertiary interactions and RNA kinetics [19, 20, 21, 22, 23, 24, 25].

Machine learning is a general method of data analysis that automates analytical model building and is based on the idea that models can learn from data, extract patterns, and make decisions with minimal human intervention. Complex problems without clear underlying mathematical structures benefit from machine learning because manually constructed analytical models cannot easily capture all of the underlying mechanics. The appeal of machine learning methods is the ability to derive predictive models without a need for strong assumptions about underlying mechanisms, which are frequently unknown or insufficiently defined in computational biology. Machine learning has exhibited unprecedented performance in protein structure prediction [26, 27, 28, 29, 30, 31], protein-ligand binding [32, 33, 34, 35], regulatory genomics, and cellular imaging [36, 37]. Deep

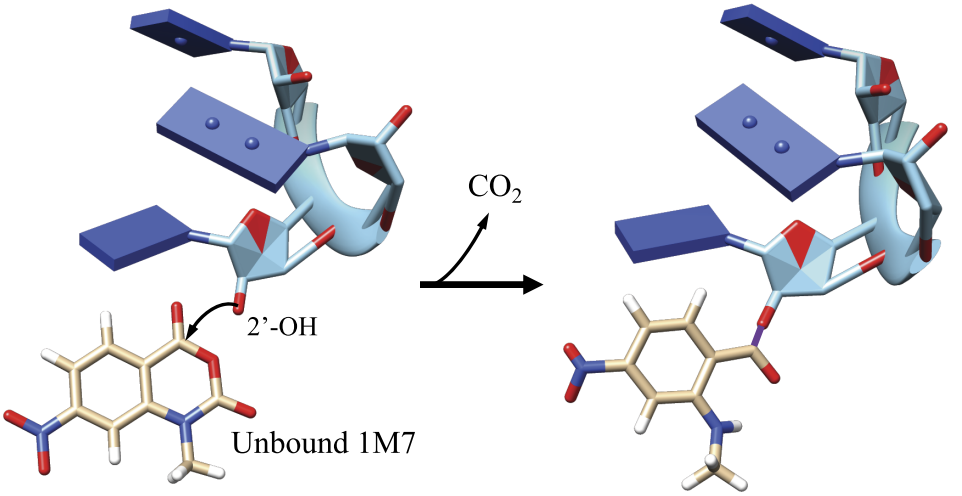

Figure 1: The SHAPE reaction. The RNA nucleotide 2'-OH group attacks the reactive carbon of 1M7, releases CO_2, and forms a covalent bond (purple) with the SHAPE reagent.

learning is a subset of machine learning based on artificial neural networks, and "deep" refers to the presence of multiple hidden layers. The convolutional neural network (CNN) is one of the deep learning network models and has gained significant attention due to its success in computer visual recognition.

Previously, we developed an analytical function to quantitatively predict the SHAPE profile from individual RNA 3D structures [38]. We showed how our function can be applied to exclude SHAPE-incompatible structures. To establish the relationship between SHAPE reactivity and nucleotide dynamics, we generated conformational ensembles with MD simulations to measure the correlation between SHAPE reactivity and the conformational propensity of each nucleotide. Then, by combining key factors that account for physical properties implicated in the SHAPE mechanism—the nucleotide interaction strength, SHAPE ligand accessibility, and base-pairing pattern—we developed the analytical 3D Structure-SHAPE Relationship (3DSSR) function, which characterizes the local nucleotide flexibility and predicts SHAPE reactivity based on information about the nucleotide posture and local energetics. To test the discriminating ability of our tool, we used the 3DSSR function to show how SHAPE-incompatible decoy structures may be excluded based on the low correlation between their predicted SHAPE profile and experimental SHAPE data.

Table 1: RNA structures used for validation. The Protein Database ID (PDBID), length of the RNA in nucleotides (nt), type of RNA, and organism of origin are displayed. The SHAPE profiles for these RNA molecules are from the published experimental data [10, 16, 17, 41, 42]

PDBID	Length (nt)	Type of RNA	Organism
2L8H	29	TAR RNA	*HIV-1*
1AUD	30	U1A protein binding site RNA	*H. sapiens*
2L1V	36	M-box riboswitch	*B. subtilis*
2K95*	48	Telomerase pseudoknot	*H. sapiens*
1Y26	71	Adenine riboswitch	*V. vulnificus*
1VTQ	75	PreQ1 riboswitch aptamer	*B. subtilis*
1EHZ	76	Aspartate tRNA	*Yeast*
1P5O*	77	IRES Domain II	*Hepatitis C*
2GDI	79	TPP riboswitch	*E. coli*
3IWN	93	Cyclic-di-GMP riboswitch	*V. cholera*
4KQY	117	SAM-I riboswitch	*B. subtilis*
1C2X*	120	5S rRNA	*E. coli*
3IVK*	128	Catalytic core of RNA polymerase ribozyme	*E. coli*
1NBS	154	Specificity domain of Ribonuclease P RNA	*B. subtilis*
3PDR	154	M-box riboswitch	*B. subtilis*
1GID*	158	Group 1 Ribozyme	*Synthetic*
3P49*	169	Glycine Riboswitch	*H. sapiens*
3DIG	174	Lysine riboswitch	*T. maritima*
4UE5*	299	SRP RNA	*C. lupus*
3G78*	421	Group II intron	*O. iheyensis*

* Denotes cases used to parameterize the CNN model, not the 3DSSR model.

Here, we revisit the 3DSSR model and develop a novel convolutional neural network (CNN) model, which uses experimental structural data to predict the SHAPE reactivity for any given nucleotide. First, we briefly describe the formulation of the 3DSSR model on a molecule that was not originally used to test or train either the 3DSSR or CNN model. Then, we describe the methods used to obtain the CNN model. Finally, we compare the ability of the two models to make useful predictions of SHAPE reactivity on RNA molecules used in training and a molecule neither algorithm has seen, emphasizing that analytical formulations often provide more insight than pattern recognition methods when limited data is available.

2. Methods

2.1. Finding structures corresponding to SHAPE data

In order to find RNA structures that correspond to our SHAPE sequences, we used the sequence searching interface equipped with NCBI's BLAST

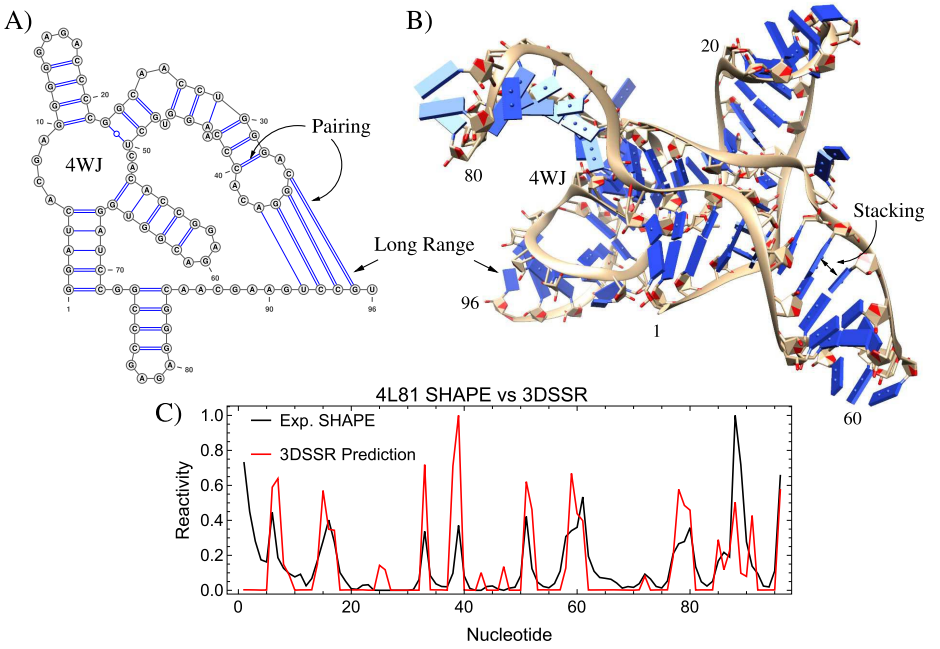

Figure 2: The 2D, 3D, and SHAPE reactivity of RNA-Puzzle 8 (PDBID: 4L81). A) The 2D structure [44] shows the four-way junction (4WJ), base-pairs, and long range interactions. B) The 3D structure shows the 4WJ and an example of a base stacking interaction. C) The experimental and predicted SHAPE profiles for the crystallized 4L81 structure show good agreement (Pearson correlation = 0.57).

(Basic Local Alignment Search Tool) program [39] provided by RCSB protein databank [40] to align the sequences. In the 3DSSR (CNN20) model 12 (20) RNA structures with an average length of ∼92 (120) nucleotides that have SHAPE reactivity data were used (see Table 1). For comparison, we also parameterized the CNN model using the same 12 structures as 3DSSR (CNN12). SHAPE reactivity data came from databases for sharing nucleic acid chemical probing data, the RNA mapping database (RMDB) [41] and the SNRNASM database [42]. To have comparable SHAPE reactivity values between different RNA structures, all of the negative values of SHAPE reactivity data are set to zero, in accordance with previous work [43]. Furthermore, the SHAPE profiles are scaled by the maximum reactivity value of each respective RNA structure, which confines SHAPE reactivity data to range from 0 to 1.

2.2. Reviewing 3DSSR methods and conclusions

Previously, we used simulations to show that SHAPE data corresponds with nucleotide flexibility, parameterize the 3DSSR model, and generate decoys to illustrate how our model can be used to exclude SHAPE-incompatible structures [38]. The ability of a nucleotide to react with SHAPE depends on the propensity of a nucleotide to sample SHAPE-reactive postures and the ability of the SHAPE ligand to access the reactive site. Capturing these concepts, we proposed the 3DSSR function

$$P(n) = BP(n) \cdot \frac{SAS(n) + S_0}{|II(n) - 1.0|} \tag{1}$$

to estimate the nucleotide stability and predict the SHAPE reactivity $P(n)$ for a nucleotide n. The base-pairing factor $BP(n)$ accounts for the 2D structure, which is characterized by the base-pairing pattern: a nucleotide n in a helix region is assigned $BP(n) = 0.01$ and a nucleotide in a loop or junction region is assigned $BP(n) = 1.0$. A 2D structure can always be extracted from a 3D structure (for example, using the RNApdbee 2.0 webserver [45]), and helix nucleotides are normally SHAPE-inert. The SHAPE ligand accessible 2'-OH surface area $SAS(n)$ describes the necessary requirement of a SHAPE ligand to access the nucleotide for a reaction to occur. If a nucleotide 2'-OH is buried inside the RNA structure, SHAPE reagents cannot react, which reduces the SHAPE reactivity. The unbound SHAPE ligand has an effective radius between 2.0 and 2.5 Å, and our results indicate that the 3DSSR function is not sensitive to different probe sizes within this range. The accessible surface of 2'-OH is calculated using VMD [46]. S_0 is a constant, accounting for the ability of a nucleotide to become accessible during experimental SHAPE probing. $II(n)$ is the interaction intensity for nucleotide n, which accounts for tertiary structure interactions. Through fitting, information from base-pairing and base-stacking interactions are combined to calculate the $II(n)$, a quasi-energy score for each nucleotide.

In the present study, we focus on a SHAPE data-assisted approach to RNA 3D structure prediction. For a given RNA sequence, we can generate an ensemble of possible conformations using, for example, the IsRNA coarse grained simulation model [48]. We then score each conformation by the correlation (similarity) between the (3DSSR-predicted) SHAPE profile of the conformation and the experimentally determined SHAPE data for the RNA molecule. Although due to the low-resolution energy model, the 3DSSR model might not be able to identify the native, crystal structure from SHAPE data alone, as shown below, the model can assist structure prediction by successfully excluding SHAPE-incompatible structures.

2.3. Applying the model to the SAM-I/IV riboswitch aptamer

For illustration, here we apply the 3DSSR model to the SAM-I/IV riboswitch aptamer (PDBID: 4L81) that was used in round 8 of RNA-Puzzles [49] (see Fig. 2), a community-wide, CASP-like blind test for RNA 3D structure prediction. This structure has not been previously used to train or test the 3DSSR model, and the structures submitted in the RNA-Puzzle competition by different labs give us objective decoys to show the ability of the 3DSSR model to exclude structures that are incompatible with SHAPE.

First, we access the submitted structures and assessment results from the RNA-Puzzles database (see Fig. 3A for a structure submitted to the competition). Next, we extract the 2D structures from the submitted 3D structures using the RNApdbee 2.0 webserver [45] (Fig. 3B). After that, we use RNAview software to identify the base pair types from the 3D structures [50] (Fig. 3C). Additionally, we directly calculate the stacking interaction information from the 3D structures: the angles and distances between different RNA bases (Fig. 3D). Then, we calculate the solvent accessible surface of each nucleotide 2'-OH in the 3D structures with VMD [46] (see Fig. 3E for a visual representation). Finally, we use the 3DSSR function to combine all of the structural information and predict the SHAPE reactivity for each nucleotide (Fig. 3F). To evaluate the SHAPE-compatibility, we also calculate the Pearson correlation between the experimental and 3DSSR-predicted SHAPE profiles. Comparison of the 3DSSR-predicted and experimental SHAPE profiles on the native, crystal structure can be seen in Fig. 2C. Provided with candidate 3D structures and experimental SHAPE data, we can exclude SHAPE-incompatible structures on the basis of their 3DSSR-predicted SHAPE profile.

The sensitivity of the model to structures with high RMSD and lower Interaction Network Fidelity (INF; a quantity to measure the similarity in interaction pattern) [47] can be seen in Fig. 5, where we apply the 3DSSR model on all of the submitted structures for RNA-Puzzle 8 to show the ability of the 3DSSR function to exclude SHAPE-incompatible structures. Contributing to the objectivity of the test, the submitted 3D structures and assessment results (values of RMSD and INF for each structure) for RNA-Puzzle 8 were all taken from the RNA-Puzzles database. The results suggest that many of the 43 submitted structures could be discarded because they are incompatible with SHAPE. For example, the native crystal structure is ranked in the top ten, and we could comfortably discard the bottom 20 structures, which all have a correlation < 0.45. Only one structure ranked in the bottom 20 by the 3DSSR model has RMSD (INF) $<$ ($>$) 11.2 (0.80),

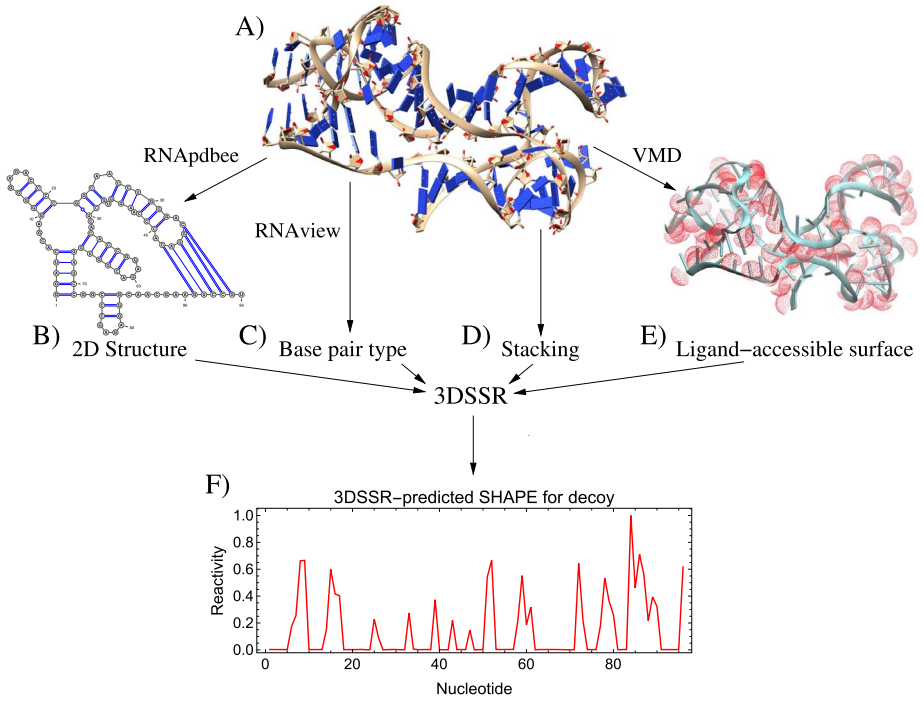

Figure 3: 3DSSR workflow on an RNA-Puzzle 8 decoy. A) A candidate 3D structure decoy is processed by RNApdbee 2.0 [45], RNAview [50], in-house software, and VMD [46] to B) produce a 2D structure, C) identify base pair types, D) extract stacking angle/distance information, and E) calculate the solvent accessible surface of the 2'-OH, respectively. The information extracted from the structure is input into the 3DSSR function to produce F) the predicted SHAPE profile for each nucleotide.

and no structure in the bottom 20 has favorable assessment values for both RMSD and INF. As can be seen in Fig. 5A, the combination of assessment results indicate that a cutoff of 0.45 is quite conservative. We could discard the bottom 65 percent of structures (the bottom 28), which would keep all of the structures with favorable assessment results for both RMSD and INF. For RNA-Puzzle 8, discarding more than the bottom 70 percent would cause us to discard the native structure. However, the quality of the SHAPE data, the size of the RNA, and the quality of candidate structures all affect the number of structures that may be comfortably excluded on the basis of SHAPE data using 3DSSR. These factors should be known so that a reasonable sieving scheme can be found.

2.4. Using a CNN to predict SHAPE reactivity from structure

2.4.1. Describing the nucleotide environment In agreement with SHAPE experiments, our CNN method probes RNA structure at single nucleotide resolution. For each nucleotide, the surrounding environment refers to neighboring atoms within a cubic volume of space around the nucleotide. Since we define the space surrounding a nucleotide as the space confined in a cube, the environment captured by this cube is not rotationally invariant. To remove the effects caused by the different choices of the cube orientation, we set a local Cartesian coordinate system for every given nucleotide. The coordinate system of a nucleotide is determined by the C1′, C4′, and O4′ atoms. Specifically, the origin of the local coordinate system is located at the atom O4′, and the local $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ axes are defined as follows. First, we denote the $\mathbf{r_{C1'}}$, $\mathbf{r_{C4'}}$, and $\mathbf{r_{O4'}}$ as the coordinates of the selected atoms, C1′, C4′, and O4′. Second, we calculate three vectors $\mathbf{v_x}$, $\mathbf{v_y}$, and $\mathbf{v_z}$ with respect to the local origin as

$$\begin{aligned} \mathbf{v_x} &= \mathbf{r_{C4'}} - \mathbf{r_{O4'}} \\ \mathbf{v_y} &= \mathbf{r_{C1'}} - \mathbf{r_{O4'}} \\ \mathbf{v_z} &= \mathbf{v_x} \times \mathbf{v_y} \end{aligned} \tag{2}$$

where $\mathbf{v_x}$ represents the vector from atom O4′ to atom C4′, $\mathbf{v_y}$ represents the vector from atom O4′ to atom C1′ and $\mathbf{v_z}$ is just the cross product of $\mathbf{v_x}$ and $\mathbf{v_y}$. Then, the $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ axes are set according to the following Eq. (3),

$$\begin{aligned} \mathbf{x} &= \frac{\mathbf{v_x}}{\|\mathbf{v_x}\|} \\ \mathbf{z} &= \frac{\mathbf{v_z}}{\|\mathbf{v_z}\|} \\ \mathbf{y} &= \mathbf{z} \times \mathbf{x} \end{aligned} \tag{3}$$

The surrounding environment of each nucleotide is captured through a cube centered and oriented according to the local coordinate system. As shown in Fig. 4, the length of the cube is 24 Å and the atoms contained in the cube will be used to generate the image for CNN model.

2.4.2. Input: defining the 3D image as input into the CNN As we described in the previous section, a 24 Å × 24 Å × 24 Å cube is used to provide the surrounding environment of each nucleotide. The corresponding

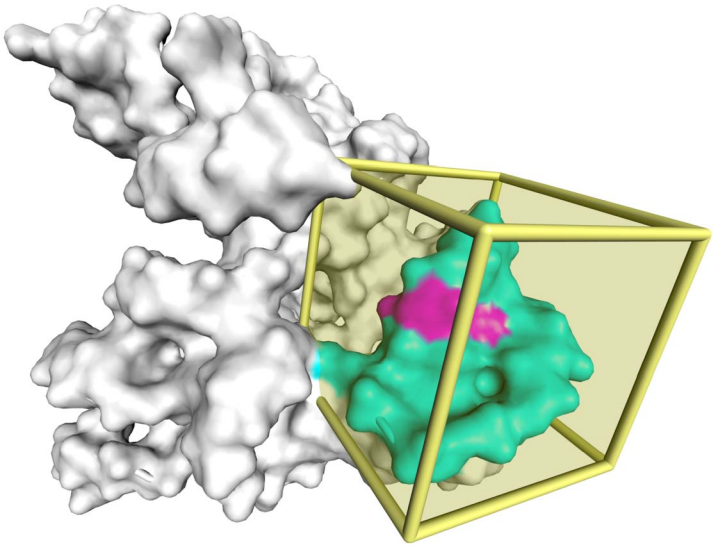

Figure 4: Extracting the 3D image of an RNA nucleotide. The magenta color depicts the nucleotide under assessment and the surrounding environment is confined within the cube with length 24 Å. The surrounding atoms are drawn in cyan, and the cube boundaries are drawn with yellow solid lines.

image associated with this nucleotide is contained within this cube. As a normal 2D digital image has three color channels (RGB) with each channel represented by a 2D pixel matrix, the 3D image that we used to capture the surrounding environment is also composed of multiple channels. However, our 3D images do not simply use RGB color channels: the channels we selected represent certain physical or chemical features. In our CNN model, we defined 5 channels, which are fully described in Table 2.

Table 2: Feature channels used for 3D images

Feature	Description
Hydrophobic	Aliphatic or aromatic carbon atoms
Aromatic	Aromatic carbon atoms
Positive ionizable	Gasteiger positive charge
Negative ionizable	Gasteiger negative charge
Excluded volume	All atom types

Since we extract our 3D image from a cube, each channel of the 3D image is represented by a 3D matrix, and each position in this 3D matrix has a voxel (3D pixel) value. We set the length of our 3D image equal to

the cube with an image resolution of 1 Å, so each voxel has a dimension of 1 Å × 1 Å × 1 Å. A step function fills the voxels of each channel. For example, the voxels of the excluded volume channel that are occupied by RNA atoms are filled with 1, and the rest are filled with 0, according to their Van der Waals radius. A similar procedure was used to generate other channels.

2.4.3. Describing the CNN architecture Our CNN model takes the multi-channel images as input, and outputs a predicted SHAPE reactivity for each image. The network is a basic ResNet [51] architecture with only slight modifications and has 10 convolutional layers. The detailed architecture is shown in Table 3. The first layer accepts the 3D image in a convolutional layer and has 64 $7 \times 7 \times 7$ filters with a stride of 2. The next layer has 4 residual blocks, with each block containing two convolutional layers. Downsampling is directly performed in the first convolutional layer and by the beginning convolutional layers of blocks 2-4. Finally, the network ends with a global average pooling layer and a 512-way fully-connected layer with a sigmoid activation function. Except the first layer, all of the convolutional layers use $3 \times 3 \times 3$ sized filters. Batch normalization [52] was applied right after each convolutional layer and before 'Rectified Linear Unit' [53] activation, following [52]. In our network, two hidden layers inserted residual shortcut connections for every block. The shortcut takes an identical input from the previous block and maps this identity shortcut right before the activation of the second hidden layer within the block; the block is same as the original ResNet block [51]. We initialize the weights as in [51, 54] and train all residual nets from scratch. The only preprocessing we used is the subtraction of a mean value from each image. This mean value is calculated by averaging all the voxels of all images in the training set.

For the network optimizer, we used Adam [55] with default parameters for momentum scheduling ($\beta_1 = 0.99, \beta_2 = 0.999$) provided by PyTorch [56], and a mini-batch size of 128 was used for training. The learning rate started from 0.01 and was divided by 10 when the training accuracy plateaued, and the models were trained for up to 100 epochs. For our loss function, we calculated the mean square error (MSE) loss between predicted SHAPE reactivities and experimental SHAPE reactivities as

$$\text{Loss} = \sum_{n=1}^{N} (P_n - G_n)^2 / N \tag{4}$$

Table 3: Details of CNN Architectures. Each building block is shown with two convolutional layers. Downsampling is performed in every convolutional layer with a stride of 2

	Layer name	Output size	Filter size	Filter num
first layer	conv1	$12 \times 12 \times 12$	$7 \times 7 \times 7$	64, stride 2
block1	conv2	$12 \times 12 \times 12$	$3 \times 3 \times 3$	64, stride 1
block1	conv3	$12 \times 12 \times 12$	$3 \times 3 \times 3$	64, stride 1
block2	conv4	$6 \times 6 \times 6$	$3 \times 3 \times 3$	128, stride 2
block2	conv5	$6 \times 6 \times 6$	$3 \times 3 \times 3$	128, stride 1
block3	conv6	$3 \times 3 \times 3$	$3 \times 3 \times 3$	256, stride 2
block3	conv7	$3 \times 3 \times 3$	$3 \times 3 \times 3$	256, stride 1
block4	conv8	$2 \times 2 \times 2$	$3 \times 3 \times 3$	512, stride 2
block4	conv9	$2 \times 2 \times 2$	$3 \times 3 \times 3$	512, stride 1
last layer	fc	$1 \times 1 \times 1$	average pool, 512-d fc, sigmoid	

where N is the number of images and $P_n(G_n)$ is the predicted(experimental) SHAPE reactivity for image n.

2.4.4. Output: predicting SHAPE reactivity with a CNN For any given 3D image that describes the surrounding environment of the considered nucleotide, our CNN model will output a real number characterizing the predicted SHAPE reactivity. This output value is confined within range from 0 to 1.

2.4.5. Implementation and cross-validating Based on the SHAPE data for 20 RNAs (totally 2455 nucleotides) collected by different experimental labs, we have 2455 SHAPE data along with the corresponding high-resolution atomic coordinates for all the nucleotides and their pertinent physical and chemical parameters. All the data together serve as the input for the CNN. To test and validate the deep learning approach, we used the leave-one-out cross-validation method to validate the performance of our model. Each time, our model was trained on 19 RNA cases with corresponding SHAPE reactivity data and tested on 1 RNA case. This process was carried out 20 times, leaving out each RNA in turn. The overall performance is evaluated by averaging the Pearson correlation coefficients of the 20 test cases over the leave-one-out process. We also carried out this procedure for the 12 cases used to parameterize the 3DSSR function. The results of the cross-validation process are summarized in Table 4. The Pearson correlation coefficient was used to measure the similarity between the predicted SHAPE profile and the experimentally derived SHAPE profile. For each training and validation set in the cross-validation, we chose the model that has the best performance on the validation set to avoid overfitting.

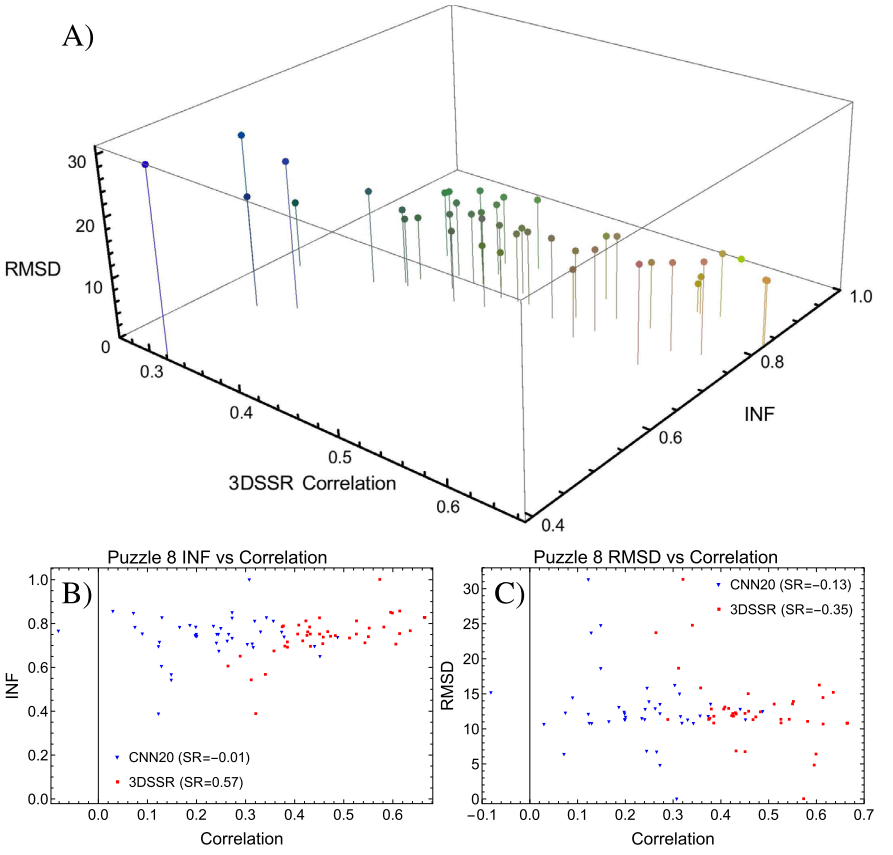

Figure 5: Sieving SHAPE-incompatible structures from RNA-Puzzle 8 submissions. A) The 3D representation shows the trend of the assessment results (RMSD and INF) with the correlation between 3DSSR-predicted SHAPE profiles and experimental SHAPE (3DSSR Correlation). Warmer colors indicate higher correlation, higher INF, and lower RMSD. The INF and RMSD values were taken from the RNA-Puzzles database. 2D plots of the B) INF and C) RMSD with respect to the 3DSSR and CNN20 correlations are also shown, along with their respective Spearman rank coefficients (SR).

2.5. Comparing 3DSSR to CNN models

As can be seen in Table 4, the 3DSSR model generally outperforms the CNN model, regardless of whether 20 or 12 structures are used to train the CNN. In contrast, the CNN model performs substantially better on 3PDR, which may indicate that information in the structure of 3PDR leading to its

Table 4: Pearson correlations between the experimental SHAPE data and the prediction algorithms: 3DSSR and the cross-validated CNN model trained on 11(19) cases and tested on the one left out, denoted as CNN12 (CNN20)

PDB	Length (nt)	3DSSR	CNN20	CNN12
2L8H	29	0.96	0.85	0.87
1AUD	30	0.92	0.90	0.71
2L1V	36	0.83	0.81	0.79
1Y26	71	0.88	0.52	0.66
1VTQ	75	0.71	0.71	0.80
1EHZ	76	0.80	0.77	0.78
2GDI	79	0.89	0.81	0.66
3IWN	93	0.74	0.33	0.38
4KQY	117	0.75	0.58	0.64
1NBS	154	0.61	0.48	0.34
3PDR	154	0.61	0.81	0.83
3DIG	174	0.70	0.64	0.68
Average	92	0.78	0.68	0.68

SHAPE reactivity profile is contained in the other cases. Because 3PDR has high performance in the CNN model in spite of its length, we may expect improvements in other cases once the amount of training data is increased. The relatively poor performance in other cases may indicate that factors that contribute to SHAPE reactivity in those RNA are not adequately represented by the structures provided in the training set. In addition, the small fluctuations captured in the 3DSSR model by using solvated, near-native representations to fit the unknown parameters may help boost its performance over the CNN.

However, the correlation alone does not show us the discerning ability of the 3DSSR and CNN models on decoy structures. For that, we turn to the results on RNA-Puzzle 8, where the Spearman rank correlation coefficient (SR) can tell us how well the models perform on ranking the structures in comparison to objective assessments (RMSD and INF). For INF(RMSD), the SR values were 0.57(−0.35) and −0.01(−0.13) for 3DSSR and CNN20, respectively, which shows that 3DSSR markedly outperforms the CNN20 model on both ranking assessments and can be used to exclude more SHAPE-incompatible structures (see Fig. 5BC).

3. Conclusion

Efficient chemical probing methods, like SHAPE, provide a wealth of information about RNA structure and dynamics. By formulating an analytical

expression that captures the key factors determining SHAPE reactivity, we can predict SHAPE reactivity from individual RNA structures. After computing predictive SHAPE profiles for a set of candidate RNA 3D structures, we can sieve the structures based on the correlation between the predicted and experimental reactivities, and SHAPE-incompatible structures can be excluded. This general method of combining efficient experimental data with computational sieving may be transferred to other efficient probing methods, enabling more confident computational determination of RNA tertiary structure at lower cost.

Machine learning techniques are incapable of creating new concepts and require the training data to be a good representative of the test data. To put it simply, a dog classification model trained with only dog images can not be used to classify cats; the model cannot be generalized to predict information it has never seen during training. Because we only have 20 RNA structures with SHAPE reactivity profiles in our data set, there is a good chance that nucleotides in a test RNA are not well represented by the other 19 structures, which results in worse performance. Additionally, using features that are important for determining SHAPE reactivity of a given nucleotide can greatly facilitate the learning process. However, finding the right combination of image channel features is not easy since the underlying mechanism that governs SHAPE reactivity is still unclear.

Although the mechanism that governs SHAPE reactivity is not fully understood, our general understanding is enough to formulate a relatively simple analytical function—the 3DSSR model—to predict reactivity based on the sensitivity of SHAPE to local nucleotide dynamics and the accessibility of SHAPE-reactive nucleotides. Because there is not enough data to apply a trained, pattern recognizing CNN to new structures, our manually constructed, analytical 3DSSR function is better at ranking structures on the basis of experimental SHAPE data. Although machine learning and advanced data-processing methods are leading to rapid advances on many problems with ample data and unclear underlying mathematical structure, physics-based models can perform better in systems where limited data is available and underlying mechanisms are known well enough to mathematically express the mechanics, even if the mechanisms are incompletely understood. As more data becomes available, we expect performance of the CNN model to improve. In the meantime, we recommend using expressions of the underlying mechanics to predict SHAPE reactivity for guiding RNA structure prediction.

References

[1] C. Cheng, W. Kladwang, J. Yesselman and R. Das, *RNA structure interference through chemical mapping after accidental or intentional mutations*. Proc. Natl. Acad. Sci. USA, **114**(37), 9876–9881, 2017.

[2] S. Yang, M. Parisien, F. Major and B. Roux, *RNA structure determination using SAXS data*. J. Phys. Chem. B, **114**(31):10039–10048, 2010.

[3] M. Parisien and F. Major, *Determining RNA three-dimensional structures using low-resolution data*. J. Struct. Biol., **179**(3):252–260, 2012.

[4] F. Ding, C. A. Lavender, K. M. Weeks and N. V. Dokholyan, *Three-dimensional RNA structure refinement by hydroxyl radical probing*. Nat. Methods, **9**(6):603–608, 2012.

[5] Z. Xia, D. R. Bell, Y. Shi and P. Ren, *RNA 3D structure prediction by using a coarse-grained model and experimental data*. J. Phys. Chem. B, **117**(11):3135–3144, 2013.

[6] E. J. Merino, K. A. Wilkinson, J. L. Coughlan and K. M. Weeks, *RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension*. J. Am. Chem. Soc., **127**(12):4223–4231, 2005.

[7] K. A. Wilkinson, E. J. Merino and K. M. Weeks, *Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution*. Nat. Protoc., **1**(3):1610–1616, 2006.

[8] S. A. Mortimer and K. M. Weeks, *A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry*. J. Am. Chem. Soc., **129**(14):4144–4145, 2007.

[9] B. Lee, R. Flynn, A. Kadina, J. Guo, E. Kool and H. Chang, *Comparison of SHAPE reagents for mapping RNA structures inside living cells*. RNA, **23**(2):169–174, 2017.

[10] C. M. Gherghe, Z. Shajani, K. A. Wilkinson, G. Varani and K. M. Weeks, *Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S^2) in RNA*. J. Am. Chem. Soc., **130**(37):12244–12245, 2008.

[11] K. M. Weeks, *Advances in RNA structure analysis by chemical probing*. Curr. Opin. Struct. Biol., **20**(3):295–304, 2010.

[12] J. L. McGinnis, J. A. Dunkle, J. H. Cate and K. M. Weeks, *The mechanisms of RNA SHAPE chemistry*. J. Am. Chem. Soc., **134**(15):6617–6624, 2012.

[13] K. E. Deigan, T. W. Li, D. H. Mathews and K. M. Weeks, *Accurate SHAPE-directed RNA structure determination*. Proc. Natl. Acad. Sci. USA, **106**(1):97–102, 2009.

[14] J. T. Low and K. M. Weeks, *SHAPE-directed RNA secondary structure prediction*. Methods, **52**(2):150–158, 2010.

[15] W. Kladwang, C. C. VanLang, P. Cordero and R. Das, *Understanding the errors of SHAPE-directed RNA structure modeling*. Biochemistry, **50**(37):8049–8056, 2011.

[16] C. E. Hajdin, S. Bellaousov, W. Huggins, C. W. Leonard, D. H. Mathews and K. M. Weeks, *Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots*. Proc. Natl. Acad. Sci. USA, **110**(14):5498–5503, 2013.

[17] C. W. Leonard, C. E. Hajdin, F. Karabiber, D. H. Mathews, O. V. Favorov, N. V. Dokholyan and K. M. Weeks, *Principles for understanding the accuracy of SHAPE-directed RNA structure modeling*. Biochemistry, **52**(4):588–595, 2013.

[18] D. H. Turner and D. H. Mathews, *NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure*. Nucleic Acids Res., **38**(suppl_1):D280–D282, 2010.

[19] S. A. Mortimer and K. M. Weeks, *Time-resolved RNA SHAPE chemistry: quantitative RNA structure analysis in one-second snapshots and at single-nucleotide resolution*. Nat. Protoc., **4**(10):1413–1421, 2009.

[20] W. Kladwang, C. C. VanLang, P. Cordero and R. Das, *A two-dimensional mutate-and-map strategy for non-coding RNA structure*. Nat. Chem., **3**(12):954–962, 2011.

[21] K. A. Steen, G. M. Rice and K. M. Weeks, *Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity*. J. Am. Chem. Soc., **134**(32):13160–13163, 2012.

[22] M. Smola, T. Christy, K. Inoue, C. Nicholson, M. Friedersdorf, J. Keene, D. Lee, J. Calabrese and K. M. Weeks, *SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells*. Proc. Natl. Acad. Sci. USA, **113**(37):10322–10327, 2016.

[23] K. Watters, A. Yu, E. Strobel, A. Settle and J. Lucks, *Characterizing RNA structures in vitro and in vivo with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)*. Methods, **103**:34–48, 2016.

[24] R. Diaz-Toledano, G. Lozano and E. Martinez-Salas, *In-cell SHAPE uncovers dynamic interactions between the untranslated regions of the foot-and-mouth disease virus RNA*. Nucleic Acids Res., **45**(3):1416–1432, 2017.

[25] M. Zubradt, P. Gupta, S. Persad, A. Lambowitz, J. Weissman and S. Rouskin, *DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo*. Nat. Meth., **14**(1):75–82, 2017.

[26] R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. F.G. Green, C. Qin, A. Zidek, A. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis and A. W. Senior, *De novo structure prediction with deep-learning based scoring*. Annu. Rev. Biochem., **77**:363–382, 2018.

[27] M. Spencer, J. Eickholt and J. Cheng, *A deep learning network approach to ab initio protein secondary structure prediction*. IEEE/ACM Trans. Comput. Biol. Bioinform., **12**(1):103–112, 2015.

[28] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang and Y. Zhou, *Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning*. Sci. Rep., **5**:11476, 2015.

[29] S. Wang, J. Peng, J. Ma and J. Xu, *Protein secondary structure prediction using deep convolutional neural fields*. Sci. Rep., **6**:18962, 2016.

[30] J. Zhou and O. G. Troyanskaya, *Deep supervised and convolutional generative stochastic network for protein secondary structure prediction*. arXiv preprint arXiv:1403.1347, 2014.

[31] S. Wang, S. Sun, Z. Li, R. Zhang and J. Xu, *Accurate de novo prediction of protein contact map by ultra-deep learning model*. PLoS Comput. Biol., **13**(1):e1005324, 2017.

[32] H. Li, K. S. Leung, M. H. Wong and P. J. Ballester, *Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets*. Mol. Inform., **34**(2-3):115–126, 2015.

[33] Z. Cang and G. W. Wei, *TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions.* PLoS Comput. Biol., **13**(7):e1005690, 2017.

[34] J. Jiménez, M. Skalic, G. Martínez-Rosell and G. De Fabritiis, *KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks.* J. Chem. Inf. Model., **58**(2):287–296, 2018.

[35] I. Wallach, M. Dzamba and A. Heifets, *AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery.* arXiv preprint arXiv:1510.02855, 2015.

[36] C. Angermueller, T. Pärnamaa, L. Parts and O. Stegle, *Deep learning for computational biology.* Mol. Syst. Biol., **12**(7), 2016.

[37] W. Jones, K. Alasoo, D. Fishman and L. Parts, *Computational biology: deep learning.* Emerging Top. Life Sci., **1**(3):257–274, 2017.

[38] T. Hurst, X. Xu, P. Zhao and S.-J. Chen, *Quantitative understanding of SHAPE mechanism from RNA structure and dynamics analysis.* J. Phys. Chem. B., **122**(18):4771–4783, 2018.

[39] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *Basic local alignment search tool.* J. Mol. Biol, **215**(3):403–410, 1990.

[40] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *The protein data bank.* Nucleic Acids Res., **28**(1):235–242, 2000.

[41] P. Cordero, J. B. Lucks and R. Das, *An RNA Mapping DataBase for curating RNA structure mapping experiments.* Bioinformatics, **28**(22):3006–3008, 2012.

[42] P. Rocca-Serra, S. Bellaousov, A. Birmingham, C. Chen, P. Cordero, R. Das, L. Davis-Neulander, C. D. Duncan, M. Halvorsen, R. Knight, N. B. Leontis, D. H. Mathews, J. Ritz, J. Stombaugh, K. M. Weeks, C. L. Zirbel and A. Laederach, *Sharing and archiving nucleic acid structure mapping data.* RNA, **17**(7):1204–1212, 2011.

[43] J. T. Low and K. M. Weeks, *SHAPE-directed RNA secondary structure prediction.* Methods, **52**(2):150–158, 2010.

[44] K. Darty, A. Denise and Y. Ponty, *VARNA: Interactive drawing and editing of the RNA secondary structure.* Bioinformatics, **25**(15):1974–1975, 2009.

[45] T. Zok, M. Antczak, M. Zurkowski, M. Popenda, J. Blazewicz, R. W. Adamiak and M. Szachniuk, *RNApdbee 2.0: multifunctional tool for RNA structure annotation*. Nucleic Acids Res., **46**(W1), W30–W35, 2018.

[46] W. Humphrey, A. Dalke and K. Schulten, *VMD: visual molecule dynamics*. J. Mol. Graph., **14**(1):33–38, 1996.

[47] Parisien, M.; Cruz, J. A.; Westhof, E.; Major, F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **2009**, *15*(10), 1875–1885.

[48] D. Zhang and S.-J. Chen, *IsRNA: an iterative simulated reference state approach to modeling correlated interactions in RNA folding*. J. Chem. Theory. Comput., **14**(4):2230–2239, 2018.

[49] Z. Miao, R. W. Adamiak, M. Antczak, R. T. Batey, A. J Becka, M. Biesiada, M. J. Boniecki, J. M. Bujnicki, S. J. Chen, C. Y. Cheng, F.C. Chou, A. R. Ferré-D'Amaré, R. Das, W. K. Dawson, F. Ding, N. V. Dokholyan, S. Dunin-Horkawicz, C. Geniesse, K. Kappel, W. Kladwang, A. Krokhotin, G. E. Łach, F. Major, T. H. Mann, M. Magnus, K. Pachulska-Wieczorek, D. J. Patel, J. A. Piccirilli, M. Popenda, K. J. Purzycka, A. Ren, G. M. Rice, J. Santalucia Jr., J. Sarzynska, M. Szachniuk, A. Tandon, J. J. Trausch, S. Tian, J. Wang, K. M. Weeks, B. Williams 2nd, Y. Xiao, X. Xu, D. Zhang, T. Zok and E. Westhof. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. RNA, **23**(5):655–672, 2017.

[50] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. Berman and E. Westhof, *Tools for the automatic identification and classification of RNA base pairs*. Nucleic Acids Res., **31**(13):3450–3460, 2003.

[51] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*. Proc. IEEE Conf. Comput. Vis. Pattern Recogni., 770–778, 2016.

[52] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. Proc. Mach. Learn. Res., **37**:448–456, 2015.

[53] V. Nair and G. E. Hinton, *Rectified linear units improve restricted boltzmann machines*. Proc. 27th Int. Conf. Mach. Learn. (ICML-10), 807–814, 2010.

[54] K. He, X. Zhang, S. Ren and J. Sun, *Spatial pyramid pooling in deep convolutional networks for visual recognition*. IEEE Trans. Pattern Anal. Mach. Intell., **37**(9):1904–1916, 2015.

[55] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014.

[56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer, *Automatic differentiation in Pytorch*. In "NIPS-W", 2017.

Travis Hurst
Department of Physics
University of Missouri–Columbia
Columbia, MO 65211
USA
E-mail address: tchvw5@mail.missouri.edu

Yuanzhe Zhou
Department of Physics
University of Missouri–Columbia
Columbia, MO 65211
USA
E-mail address: yzbn4@mail.missouri.edu

Shi-Jie Chen
Department of Physics, Department of Biochemistry, and
Institute of Data Sciences & Informatics
University of Missouri–Columbia
Columbia, MO 65211
USA
E-mail address: ChenShi@missouri.edu

Received August 1, 2019