Machine learning a model for RNA structure prediction

Nicola Calonaci[1], Alisha Jones[2,3], Francesca Cuturello[1], Michael Sattler[2,3], and Giovanni Bussi[1]

[1]International School for Advanced Studies, via Bonomea 265, 34136 Trieste, Italy
[2]Institute of Structural Biology, Helmoltz Zentrum München, Neuherberg 85764, Germany
[3]Center for Integrated Protein Science München and Bavarian NMR Center at Department of Chemistry, Technical University of Munich, Garching 85757, Germany

Abstract

RNA function crucially depends on its structure. Thermodynamic models that are used for secondary structure prediction report a large number of structures in a limited energy range, often failing in identifying the correct native structure unless complemented by auxiliary experimental data. In this work we build an automatically trainable model that is based on a combination of thermodynamic parameters, chemical probing data (Selective 2' Hydroxyl Acylation analyzed via Primer Extension, SHAPE), and co-evolutionary data (Direct Coupling Analysis, DCA). Perturbations are trained on a suitable set of systems for which the native structure is known. A convolutional window is used to include neighboring reactivities in the SHAPE nodes of the network, and regularization terms limit overfitting improving transferability. The most transferable model is chosen with a cross-validation strategy that allows to automatically optimize the relative importance of heterogenous input datasets. The model architecture enlightens the structural information content of SHAPE reactivities and their dependence on local conformational ensembles. By using the selected model, we obtain enhanced populations for reference native structures and more sensitive and precise predicted structures in an independent validation set not seen during training. The flexibility of the approach allows the model to be easily retrained and adapted to incorporate arbitrary experimental information.

1 Introduction

Ribonucleic acids (RNA) transcripts, and in particular non-coding RNAs, play a fundamental role in cellular metabolism being involved in protein synthesis [1], catalysis [2], and regulation of gene expression [3]. RNAs often adopt dynamic interconverting conformations, to regulation their functional activity. Their function is however largely dependent on a specific active conformation [4], making RNA structure determination fundamental to identify the role of transcripts and the relationships between mutations and diseases [5]. The nearest-neighbor models based on thermodynamic parameters [6, 7] allow the stability of a given RNA secondary structure to be predicted with high reliability, and dynamic programming algorithms [8, 9] can be used to quickly identify the most stable structure or the entire partition function for a given RNA sequence. However, the coexistence of a large number of structures in a narrow energetic range [10] often makes the interpretation of the results difficult. Whereas there are important cases where multiple structures are indeed expected to coexist *in vivo* and might be necessary for function [11, 12], the correct identification of the dominant structure(s) is crucial to elucidate RNA function and mechanism of action. In order to compensate for the inaccuracy of thermodynamic models, it is becoming common to complement them with chemical probing data [13] providing nucleotide-resolution information that can be used to infer pairing propensities (e.g., reactive nucleotides are usually unpaired). Particularly interesting is selective $2'$ hydroxyl acylation analyzed via primer extension (SHAPE) [14, 15], as it can also probe RNA structure *in vivo* [16]. In a separate direction, novel methodologies based on direct coupling analysis (DCA) have been developed to optimally exploit co-evolutionary information in protein structure prediction [17] and found their way in the RNA world as well [18, 19]. Whereas the use of SHAPE data and of multiple sequence alignments in RNA structure prediction is becoming more and more common, these two types of information have been rarely combined [20].

In this paper, we propose a model to optimally integrate RNA thermodynamic models, SHAPE experiments, and DCA co-evolutionary information into a robust structure prediction protocol. A crucial ingredient of the model is the inclusion of chemical probing data from multiple neighboring nucleotides by means of a convolutional network. A machine learning procedure is then used to select the appropriate model and optimize the model parameters based on available experimental structures. Regularization hyperparameters are used to tune the complexity of the model thus controlling overfitting and enhancing transferability. The resulting model leads to performance in secondary structure prediction that surpasses available methods when used on a validation set not seen in the training phase. The parameters can be straightforwardly re-trained on new available data.

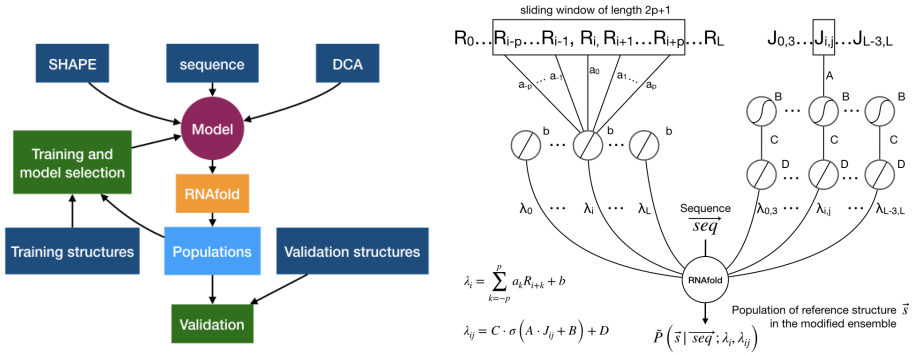


Figure 1: Graphical scheme of the machine learning procedure. (a) Models that integrate RNAfold, SHAPE experiments, and DCA scores into prediction of structure populations are trained. One among all the proposed models is selected based on a transferability criterion and validated against data that is not seen during training. Available reference structures are used as target for training and validation. (b) Sequence, SHAPE and DCA data are included through additional terms in the RNAfold model free energy. The network is split into two channels: a single-layered channel for SHAPE input (left side) and a double-layered channel for DCA couplings (right side). Along the SHAPE channel, a convolutional layer operates a linear combination on the sliding window including the SHAPE reactivity R_i of a nucleotide and the reactivities $\{R_{i+k}\}$ of its neighbors, with weights $\{a_k\}$ and bias b. The output consists in a pairing penalty λ_i for the i-th nucleotide. In the DCA channel, the first layer transforms the input DCA coupling J_{ij} via a non-linear (sigmoid) activation function, with weight A and bias B. The transformed DCA input is then mapped to a pairing penalty λ_{ij} for the specific ij pair via a second layer, implementing a linear activation function with weight C and bias D. Penalties for both individual nucleotides and for specific pairs are applied as perturbations to the RNAfold free-energy model.

2 Results

The architecture of the model is summarized in Fig. 1. Chemical probing experiments provide reactivities per nucleotide (one-dimensional information, R_i) that are mapped via a single-layered convolutional network to penalties to be associated to the pairing propensity of individual nucleotides (λ_i). Similarly, direct-coupling analysis provides predicted contact scores (two-dimensional information, J_{ij}) that are mapped through a non-linear function into penalties to be associated with specific nucleotide-nucleotide pairs (λ_{ij}). The resulting penalties are integrated in the folding algorithm RNAfold from the Vienna package [9], which allows the full partition function of the system to be computed, including the population of any suboptimal structure. The parameters of the mapping functions are trained in order to maximise the population of the secondary structures as annotated in a

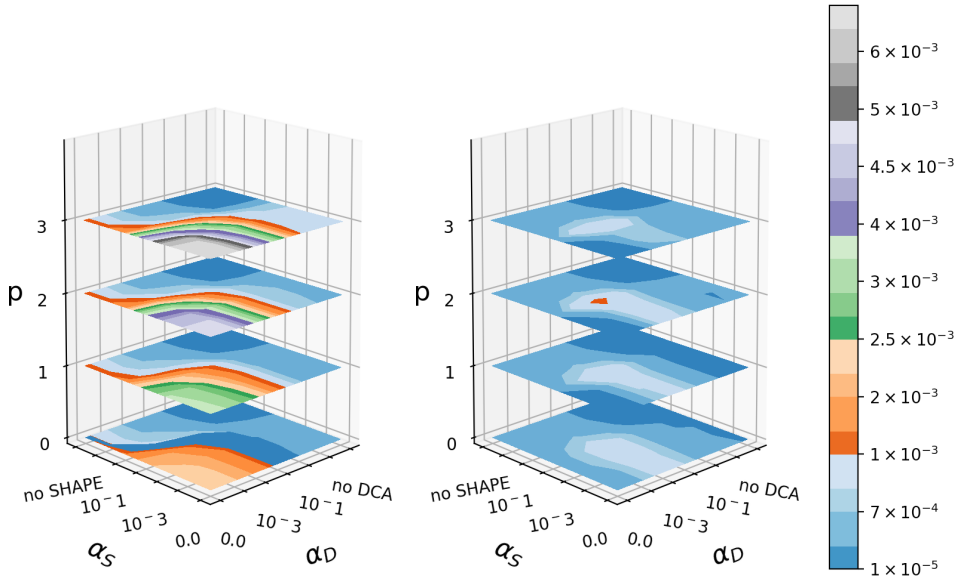
Molecule	PDB	l_{seq}	S1	S2	S3	S4	S5
yeast Phe-tRNA	1EHZ	76	T	V	T	T	T
D5,6 Yeast ai5g G-II Intron	1KXK	70	T	T	T	T	V
Adenine riboswitch	1Y26	71	T	T	T	T	T
TPP riboswtich	2GDI	78	T	T	V	V	T
SAM riboswitch	2GIS	94	T	T	V	T	T
Lysine riboswitch	3DIG	174	T	T	V	V	T
c-di-GMP riboswitch	3IRW	90	T	T	T	V	V
M-box riboswitch	3PDR	161	V	V	T	T	T
THF riboswitch	3SD3	89	V	V	T	T	V
Fluoride riboswitch	3VRS	52	V	T	T	T	T
50S ribosomal	4YBB_CB	120	T	T	T	T	T

Table 1: RNA molecules included in the dataset. For each molecule we indicate the PDB ID (second column) of the corresponding annotated structure, the number of nucleotides (l_{seq}), and, for each random dataset splitting that we used (S1 to S5), a mark to denote whether the molecule data are used for training (T) or validation (V). For PDB 4YBB, chain CB was used as a reference.

set of high-resolution X-ray diffraction experiments. The differentiability of the RNAfold model with respect to the applied penalties is crucial, since it allows the thermodynamic model to be used during the training procedure. Reference structures are obtained from the structural database [21]. Reference SHAPE data are partly taken from the RNA mapping database [22, 23] and from Ref. [24], and partly reported for the first time in this paper. Reference direct couplings are partly taken from Ref. [25] and partly obtained in this paper, using RNA families deposited on RFAM [26]. The model complexity is controlled via three hyperparameters, which are chosen using a cross-validation procedure, and the obtained model is evaluated on an independent dataset not seen during the training procedure. A more detailed explanation can be found in the Online Methods.

2.1 Model training

We randomly choose a training set of 8 systems, leaving 3 others out for later validation. Since crystal structures, SHAPE data, and DCA data for different systems might be of different quality, the specific choice of the splitting might affect the overall training and validation results. We thus generate five independent random splittings, reported in Table 1. In the following we refer to splitting S4, as it yields intermediate performance on the validation set. Results obtained with different splittings are reported in Supporting Information. The model complexity is controlled by means of three handles: a regularization parameter acting on the one-dimensional



(a) **(b)**

Figure 2: Population of native structure as function of hyperparameters. Population is indicated in the color scale. The optimized population of native structures, when averaged on the training set (a), is by construction a monotonically increasing function of the integer p controlling the window size of the convolutional network in the SHAPE channel, and a monotonically decreasing function of the regularization coefficients α_S and α_D. When averaged on the leave-one-out iterations of the CV procedure (b), the dependency of the optimized population of native structures on these hyperparameters becomes non-trivial, as it results from a combination of model complexity (controlled by p) and regularization (controlled by α_S and α_D independently). The CV procedure serves as criterion for model selection, resulting in the selection of hyperparameters $\{p = 2, \alpha_S = 0.01, \alpha_D = 0.001\}$.

penalties derived from SHAPE reactivities ($0 \leq \alpha_S \leq \infty$), a regularization parameter acting on the two-dimensional penalties derived from DCA ($0 \leq \alpha_D \leq \infty$) and an integer controlling the size of the window used for the convolutional network ($p <= 3$). When the performance of the model is evaluated on the training set, the model that better fits the data is the most complex one, with no regularization term ($\alpha_S = \alpha_D = 0$) and the largest tested window ($p = 3$) (Fig. 2a). The geometric average increase in the population of native structures, when compared with the thermodynamic model alone, is ≈ 108 times. Within our procedure it is straightforward to train the model in order to use only SHAPE data ($\alpha_D = \infty$), or only DCA data ($\alpha_S = \infty$), resulting in increases of native population of ≈ 17 times and ≈ 24 times respectively using the same randomized training set (S4 of Table 1). This result suggests that the amount of information provided by the co-evolutionary analysis is higher than that provided by chemical probing experiments.

2.2 Model selection

In order to make the parametrization transferable, we perform a leave-one-out cross-validation (CV) procedure (see Online Methods) where one of the 8 systems at a time is left out of the training procedure and the increase in the native population for the left-out system is used to estimate transferability. Overall, the average performance of the model on the left-out system shows a non-trivial dependence on the hyperparameters (Fig. 2b). The model with hyperparameters $\alpha_S = \alpha_D = 0$ and $p = 3$ results in CV populations that are *lower* than those obtained with the thermodynamic model alone, which is a signature of overfitting. The best performance in the cross-validation test is obtained when choosing $\alpha_S = 0.01$, $\alpha_D = 0.001$ and $p = 2$. We select this model as the one that yields the best balance between performance and transferability. Results obtained by using a different randomization of the training set are reported in SI. Whereas the precise set of optimal hyperparameters depends on the specific training set, sets of hyperparameters that perform well on a specific set tend to perform well for all the tested training sets.

2.3 Validation on an independent dataset

Finally, we evaluate the performance of the selected model on a dataset of three systems that were not seen during training. This additional test is done in the spirit of nested cross-validation [27] in order to properly evaluate the transferability of the procedure.

For the three test systems (splitting S4 of Table 1), the introduced procedure leads to a boost of the population of the native structure of ≈ 10 times, on average (Fig. 3a, right side of the vertical line), when using the

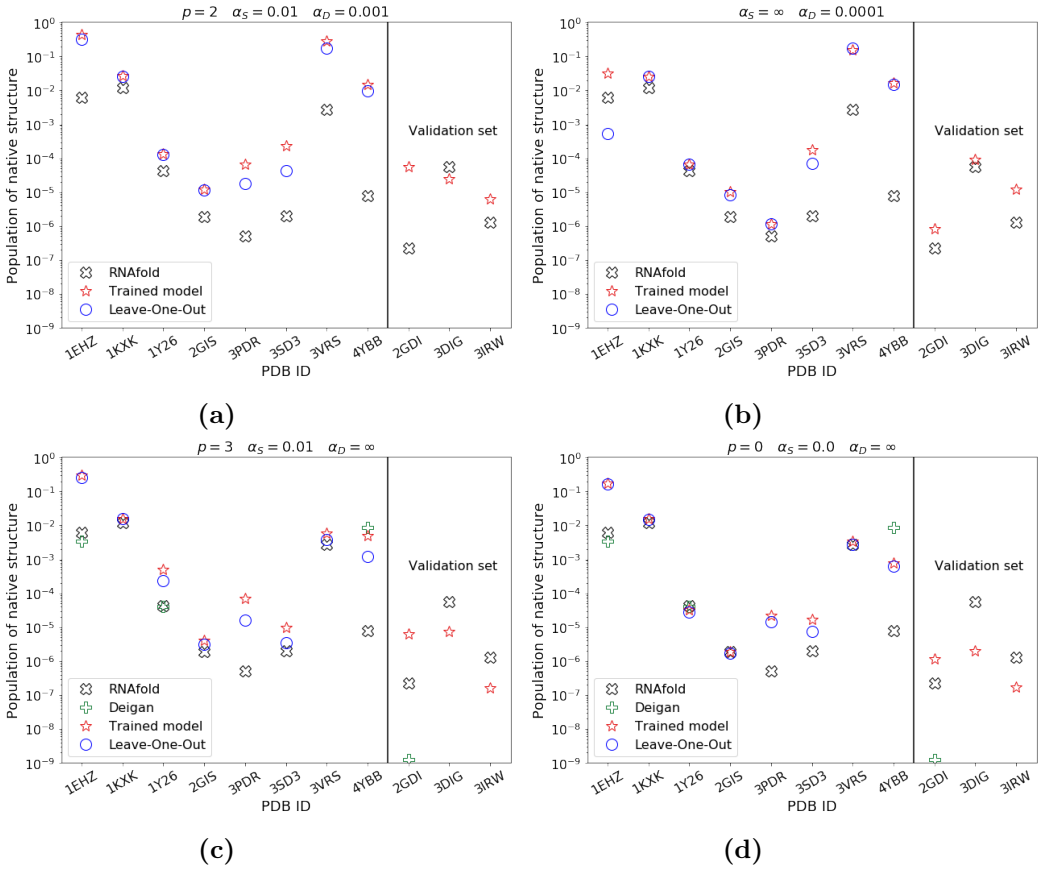
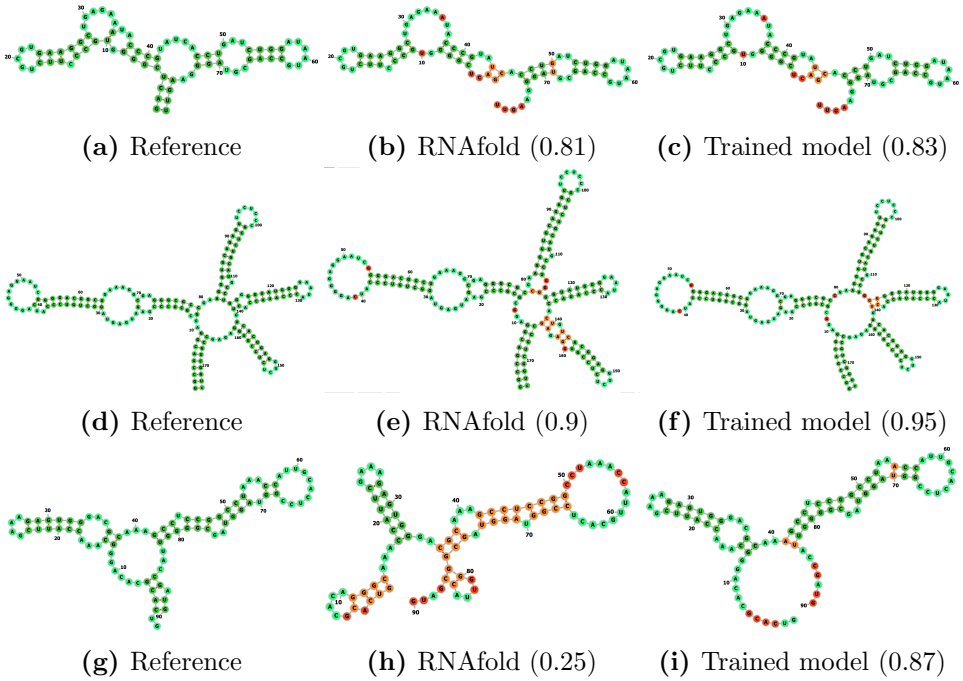


Figure 3: Comparison of populations of native structures obtained with unmodified RNAfold and with selected models, respectively: (a) the best performing model; (b) the best performing model with DCA data only; (c) the best performing model with SHAPE data only and (d) the best performing model with SHAPE data only and convolutional window width $p = 0$. Hyperparameters are noted in the figure. Native structure populations obtained with unmodified RNAfold (black cross), with our trained model (red star) and in the leave-one-out procedure (blue circle, for each molecule the model is trained on all the other molecules in the training set) are reported. Populations obtained by mapping SHAPE reactivities into penalties with the method in Ref. [15] are reported for comparison (green plus), only for molecules studied in previous work and in panels where SHAPE data only are used (c and d). The populations of native structures that we obtain with the trained model are almost always enhanced for molecules included in the training set (left side of the vertical line), whereas for molecules not included (right side of the vertical line) overfitting occur in some cases, yielding populations lower than obtained with unmodified RNAfold.

selected model $\{\alpha_S = 0.01, \alpha_D = 0.001, p = 2\}$. A side effect of targeting the population of native structures for model optimization and selection is the increase in the similarity between the predicted minimum free energy (MFE) structures and the experimental ones. Reference secondary structures, along with MFE predictions made with unmodified RNAfold and with the selected model are reported in Fig. 4 for comparison. This similarity can be quantified using the Matthews Correlation Coefficient (MCC) [28], that is routinely used to benchmark RNA structure prediction [29]. Its average is here enhanced from 0.66 to 0.88. Specific changes in the predicted secondary structures are reported in detail in the caption of Fig. 4. Considering all the tested splittings of the dataset, the average MCC of minimum free energy structure predictions is enhanced from 0.72 ± 0.22 to 0.87 ± 0.07, implying both an increased average and a decreased variance (details in SI).

It is also possible to test the scenarios where only DCA data or only SHAPE data are available. In case of DCA-only information ($\alpha_S = \infty$), the best performance in CV is obtained using the model with $\alpha_D = 0.0001$ (Fig. 3b). This model is transferable to the validation set yielding an increase in the population of the native structures lower but comparable to what obtained when SHAPE and DCA data are both included. In case of SHAPE-only information ($\alpha_D = \infty$), the best performance in CV is obtained using the model with hyperparameters $\alpha_S = 0.01$ and $p = 3$. In this case our CV procedure fails to identify a parametrization that is completely transferable to the validation set, as can be seen from Fig. 3c. The population of native structure is increased for only one system out of the three, and it is slightly decreased on average, by a factor ≈ 0.8.

Our procedure to compute pairing penalties from SHAPE data can be compared with the one introduced by Deigan et. al. [15]. Since the Deigan's method requires SHAPE data normalized with a different procedure, we use normalized reactivities reported in Ref. [23]. Remarkably, our procedure leads to significantly better results both for molecules that are included in the training set (e.g. `1EHZ` and `1Y26` in Fig. 3c), and for the one included in the validation set (`2GDI` in the right side of Fig. 3c). Even when `1EHZ` is not used for training but only as a validation structure (splitting $S2$) the selected model yields an increase in population of native structure of a factor ≈ 5, outperforming both unmodified RNAfold and Deigan's method. The improvement in performance that we obtain may be due either to a more robust training and validation protocol, to the choice of a mapping function that includes neighboring reactivities, or to a combination of these two factors. In order to discriminate between these scenarios, we restrict further the comparison with Deigan's procedure to models with $p = 0$. In this case our best performing model in CV is defined by $\alpha_S = 0$. As can be seen from Fig. 3d, populations are on average lower than those obtained with $p = 3$ and $\alpha_S = 0.01$, suggesting that the inclusion of neighboring SHAPE reactivities, when combined with proper regularization, is preferable.



(a) Reference **(b)** RNAfold (0.81) **(c)** Trained model (0.83)

(d) Reference **(e)** RNAfold (0.9) **(f)** Trained model (0.95)

(g) Reference **(h)** RNAfold (0.25) **(i)** Trained model (0.87)

Figure 4: Minimum free energy (MFE) structure predictions. For each system in the validation set, reference native structure is compared with predicted MFEs. Correctly predicted base pairs (true positives) and unpaired nucleotides (true negatives) are reported in dark green and lime green, respectively. Wrongly predicted base pairs (false positives) and unpaired nucleotides (false negatives) are reported in orange and red, respectively. MCC between prediction and reference is reported in parenthesis. For `2GDI` (a-c), the accuracy of prediction is slightly increased with respect to unmodified RNAfold. In particular, our model recovers the correct structure of the internal loop (49–53:66–70). For `3DIG` (d-f), some bulges and an interior loop wrongly predicted by RNAfold are correctly discarded by our model, recovering almost completely the structure of the five-way junction. In the best-case result, `3IRW` (g-i), despite our model is not able to predict correctly the initial helix (2–5:86–90), it recovers almost completely the other stem-loop structures that are not correctly predicted by RNAfold. All secondary structure diagrams are drawn with `forna` [30].

2.4 Interpretation of parameters

In principle, different randomizations of the training set yield different hyperparameters and parameters for the functions implemented in the selected model. Here we continue focusing on splitting S4 as it yields intermediate performance in validation. The selected model is defined by hyperparameters $\{\alpha_S = 0.01, \alpha_D = 0.001, p = 2\}$. Results for different splittings are similar and are reported in SI.

DCA channel. DCA couplings are mapped into pairing penalties through a double-layered neural network, resulting in a non-linear function reported in Fig. 5a. Pairing penalties are found to be decreasing with increasing DCA coupling value, consistently with the interpretation that large couplings should correspond to co-evolutionarily related and thus likely paired nucleobases [18]. We notice however that the final bias in this non-linear function is coupled with the bias of the SHAPE channel and thus cannot be directly interpreted. We can nonetheless exclude this effect by restricting to models taking only DCA couplings as input ($\alpha_S = \infty$). The corresponding non-linear function is reported in Fig. 5b. The overall shape is consistent with that obtained fitting all the data (Fig. 5a), but the zero of this function can be straightforwardly interpreted as the threshold for penalizing or favoring base pairing. The resulting value is $J^{\text{threshold}} = 0.51$ consistent with the typical thresholds obtained in [25] with a different optimization criterion, based on the accuracy of contact predictions, and fitted on a larger dataset, thus confirming the transferability of the non-linear function reported here.
SHAPE channel. SHAPE reactivities are mapped into penalties affecting the population of individual nucleotide pairing states through a single convolutional layer with a linear activation function. For each nucleotide, the network input vector includes SHAPE reactivities from its second-nearest-neighbor upstream to its second-nearest-neighbor downstream along the sequence. The activation coefficients weight the contribution of each nucleotide in the neighbor window. Optimal parameters are reported in Fig. 5c. We first notice that the sum of the weights $\sum_{i=-2}^{2} a_i$ is negative, so that the pairing of a nucleotide in a highly reactive region is unfavored, and vice-versa for regions of low reactivity. The largest contribution arises from the a_0 term, confirming that reactivity of a nucleotide, related to its flexibility [31, 32], is maximally affected by its pairing state. The second-largest contribution is given from the reactivity of the first downstream neighbor (a_{+1}), with a weight that has the opposite sign. The combination of the a_0 and a_{+1} contributions can be interpreted as a forward finite-difference operator estimating the first derivative of the reactivity with respect to the position in the sequence. This contribution maps a local downward trend of the reactivity profile into a pairing penalty, thus providing a sort of normalization for the reactivity of the central nucleotide with respect to that of the first downstream neighbor. The remaining weights are lower and can

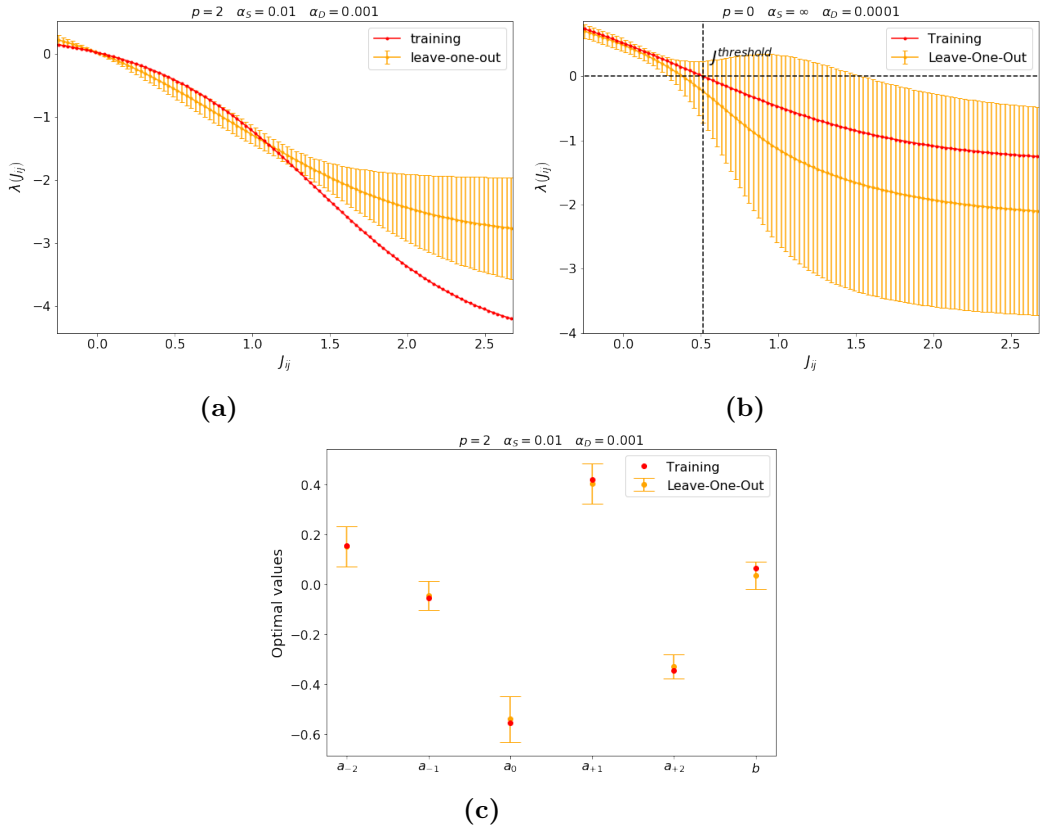


Figure 5: Properties of the optimized neural network. For the DCA channel, the optimized function mapping DCA couplings J_{ij} into pairing penalties λ_{ij}, for both (a) the selected model and (b) the best performing model with restriction to only DCA input. When trained on the whole training set (red) the activation function is consistent with the average on the leave-one-out training subsets (orange). Error bars are computed as standard deviations and are significantly lower in the region of DCA couplings around zero, as couplings lying in that region are more frequent. The trained function maps (respectively, low) high DCA coupling values to penalties favoring (respectively, disfavoring) the corresponding pairings, thus affecting the population of the structures including the specific pair. When restricting to (b) models including only DCA input, the threshold value of the coupling J^{th} between disfavored and favored pairing corresponds to the zero of the activation function, as indicated by the dashed line. For the SHAPE channel, (c) optimal values of model parameters are shown for the selected model, with hyperparameters $\{\alpha_S = 0.01, \alpha_D = 0.001, p = 2\}$. Training results (red) lie within the leave-one-out error bars (orange), indicating robustness of the minimization procedure against cross-validation. Coefficients $\{a_{-2}, \ldots, a_{+2}\}$ weighting reactivities up to the second-nearest-neighbors of a nucleotide, report the importance of the reactivity pattern for the downstream neighbor in addition to the nucleotide's own reactivity.

be interpreted as corrections to penalties taking into account higher-order variations in the local reactivity profile of the neighbor window.

3 Discussions

In this work we build a network that can be used to predict RNA structure taking as an input RNA sequence, SHAPE reactivities, and DCA scores. Whereas SHAPE reactivities and DCA scores are processed through standard linear or non-linear units, RNA sequence enters through a thermodynamic model. A crucial ingredient that we introduce here are the derivatives of the result of the thermodynamic model with respect to the pairing penalties, that allow the network to be trained using gradient-based machine learning techniques.

We built up a total of 196 models to map simultaneously SHAPE reactivities and DCA scores into free-energy terms coupling, respectively, the pairing state of individual nucleotides and that of specific pairs of nucleotides. Each model is defined by tunable hyperparameters controlling the width of the windows used to process SHAPE reactivities and the strength of the regularization terms applied to SHAPE and DCA data. The dataset is *a priori* split randomly into a training set and a validation set (8 and 3 systems respectively). Training, model selection and validation are repeated for different random splittings of the dataset, ensuring the robustness of the procedure. The whole procedure, from training to model selection, is automatic so that new parameters could be straightforwardly obtained using new SHAPE and DCA data and new crystallographic structures, allowing for a continuous refinement of the proposed structure prediction protocol. In the dataset we used, some SHAPE reactivities are taken from available experimental data. Other SHAPE reactivities are measured here for the first time so as to increase the number of systems for which both co-evolutionary data and SHAPE reactivities are available. DCA scores are based on ClustalW alignments [33] so that they are not manually curated with prior structural information. We notice however that classification of sequences in RFAM is performed including structural information, when available. In addition, co-evolutionary information might be difficult to extract for poorly conserved long non-coding RNAs. All the results obtained with different randomization of the validation set are reported in SI so that different sets of parameters can be easily tested.

The model selected via CV is defined by hyperparameters $\{p = 2, \alpha_S = 0.01, \alpha_D = 0.001\}$. The best performing method thus incorporates in the pairing state of a nucleotide reactivities from neighboring nucleotides up to the second-nearest ones. This is furthermore supported by the fact that, when restricting to SHAPE-only information, models with larger windows p perform better than models with narrower windows. In general the selected

models with $p > 0$ yield better results in population than what obtained with Deigan's method [15], which is accounted for best state-of-the-art method [34] among those based on SHAPE reactivities only. Our results thus show that the reactivity of a nucleotide is affected by the local conformational ensemble, in agreement with molecular dynamics simulations [32]. Our results also confirm that the reactivity of a nucleotide is a good indicator of its own pairing state [34], but additionally indicate that the reactivity trend with respect to its downstream nearest-neighbor plays an important role as well (see Section 2.4, *SHAPE channel* paragraph). With this approach we at least partially solve inconsistencies between SHAPE reactivities and pairing states previously pointed out [35, 36] as an issue in fitting free-energy terms computed from individual nucleotide reactivities. In perspective, the model can be extended to include additional features of the SHAPE experiments that may be related to non-canonical interactions and three-dimensional structure.

We also notice that when trained on SHAPE data alone our procedure identifies a set of hyperparameters for which the model is overfitted. This might be due to the heterogeneity of the SHAPE data used here, that come from different sources and might have been obtained in different conditions. On the other hand, the model obtained combining SHAPE and DCA data is systematically capable of predicting a higher population for the native structure.

Although the model is trained to maximize the population of the individual reference structure as obtained by crystallization experiments, it can still report alternative structures. Whereas we did not investigate this issue here, alternative low-population states might be highly relevant for function. Compatibly with that, the absolute population of the native structure remains significantly low (from $\approx 10^{-6}$ to $\approx 10^{-1}$), but is still one of the highest in the ensemble. In particular, the individual structure with highest population (minimum free-energy structure) with our method is closer to the reference crystallographic structure than the one predicted by thermodynamic parameters alone on systems not seen during training.

Importantly, all the data and the used scripts are available and can be used to fit the model over larger datasets. In order to avoid overfitting, we suggest to repeat the leave-one-out procedure to select the most transferable model, whenever new independent data is added to the dataset. Scripts for training and model selection are reported in SI. In principle the model can be straighforwardly extended to include any chemical probing data that putatively correlates with base pairing state [13] or other types of experimental information that correlate with base-pairing probabilities [37]. Training on a larger set of reference structures and using more types of experimental data will make the model more robust and open the way to the reliable structure determination of non-coding RNAs.

4 Online methods

4.1 Secondary structure annotation

The secondary structures that we use as examples for training and validation are obtained annotating crystallographic structures with `x3dna-dssr` [38]. Differently from previous work, we include all the computed cis-Watson-Crick contacts as reference base pairs, with exception of pseudoknots that are forbidden in predictions made with RNAfold. All the reference structures are published in the PDB database and have a resolution better than 3 Å so that they can be assumed to be of similar quality, although crystal packing effects or other artefacts might in principle be different. The list of PDB files used in this work is reported in Table 1.

4.2 Thermodynamic model

As a starting point we use the nearest neighbor thermodynamic model [6, 7] as implemented using dynamic programming [8] in the ViennaRNA package [9]. Given a sequence $s\vec{e}q$ the model estimates the free energy associated to any possible secondary structure $\vec{s}$ by means of a sum over consecutive base pairs, with parameters based on the identity of each involved nucleobase. We denote this free energy as $F_0\left(\vec{s}|s\vec{e}q\right)$. We used here the default thermodynamic parameters of the ViennaRNA package [39], but the method could be retrained starting with alternative parameters. The probability of a structure $s\vec{e}q$ to be observed is thus

$$P_0\left(\vec{s}|s\vec{e}q\right) = \frac{e^{-\frac{1}{RT}F_0\left(\vec{s}|s\vec{e}q\right)}}{Z_0\left(s\vec{e}q\right)} \tag{1}$$

where Z_0 is the partition function, R is the gas constant and T the temperature. Importantly, the implemented algorithm is capable to find in a polynomial time not only the most stable structure associated to a sequence ($\arg\min_{\vec{s}} F_0\left(\vec{s}|s\vec{e}q\right)$) but also the full partition function Z_0 and the probability of each base pair to be formed [40].

4.3 Experimental data

4.3.1 SHAPE data

SHAPE data for systems `1EHZ`, `1Y26`, `2GDI`, `3DIG`, `3PDR` and `4YBB_CB` were taken from the literature [22, 23, 24]. SHAPE data for systems `1KXK`, `2GIS`, `3IRW`, `3SD3` and `3VRS` were collected for this work.

Single stranded DNA templates containing the T7 promoter region and the 3 and 5 SHAPE cassettes [41] were ordered from Eurofins Genomics. RNAs were transcribed using in-house prepared T7 polymerase. Briefly,

complementary T7 promoter DNA was mixed with the desired DNA template and snap cooled (95 °C for 5 minutes, followed by incubation on ice for 10 minutes) to ensure annealing of the T7 complementary promoter with template DNA. The mixture was supplemented with rNTPs, 20X transcription buffer (TRIS pH 8, 100 mM Spermidine, 200 mM DTT), PEG 8000, various concentrations of MgCl2 (final concentration ranging from 10 to 40 mM), and T7 (10 mg/mL). The RNAs were run on a 12% denaturing polyacrylamide urea gel and stained with toluidine blue to assess the most optimal MgCl2 concentration. The RNA was cut from the gel and isolated using the crush and soak method [42]. Following crush and soak, the RNAs were precipitated using ethanol, and resuspended in RNAse-free water. SHAPE modification followed by reverse transcription (using 5' FAM labeled primers) was carried out as previously described [41]. Following reverse transcription, RNAs were precipitated using ethanol, redissolved in HiDi formamide, and cDNA fragments separated using capillary electrophoresis (ABI 3130 Sequencer). Raw reads corresponding to cDNA fragments were obtained using QuSHAPE [43] and are reported in SI. Reads in each of the control and modifier channels were first normalized independently by dividing them by the sum of reads in the corresponding channel. SHAPE reactivities were then estimated by subtracting the normalized reads in the control channel from the normalized reads in the modifier channel, with negative values replaced with zeros. This normalization is a simplified version of the one proposed in Ref. [44] and does not contain position dependent corrections. These corrections are only expected to be relevant for RNA molecules significantly larger than those analyzed here.

4.3.2 DCA data

Direct couplings for all the systems were calculated using the same code and parameters reported in Ref. [25], but aligning them with ClustalW [33] so as to avoid including indirectly known structural information. For systems where the sequences used in Ref. [25] were different from those reported in the PDB or used in SHAPE experiments, DCA calculations were performed again. Coupling J_{ij}'s were computed as the Frobenius norm of the couplings between positions i and j, as detailed in Ref. [25]. All the used alignments and couplings are reported in Supporting Information.

4.4 Penalties

We integrate SHAPE reactivities R_i and direct couplings J_{ij} into the model by mapping them into penalties to pairing propensity of, respectively, individual nucleotides and specific nucleotide pairs. This is obtained through a

modification of the original model free energy by two additional terms:

$$F\left(\vec{s}|s\vec{e}q;\vec{R},\vec{J}\right) = F_0\left(\vec{s}|s\vec{e}q\right) + RT\sum_{i=1}^{l_{seq}}\lambda_i(\vec{R})\cdot(1-s_i) + RT\sum_{j>i+2}^{l_{seq}}\lambda_{ij}(\vec{J})\cdot s_{ij} \tag{2}$$

where s_i is the pairing status of the i-th nucleotide in the structure $\vec{s}$

$$s_i = \begin{cases} 1 & \text{if nucleotide i is paired} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and s_{ij} is the pairing status of the specific couple of nucleotides i and j

$$s_{ij} = \begin{cases} 1 & \text{if nucleotide i is paired with nucleotide j} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

We implement both kinds of penalties in the folding algorithm using the soft constraints functions from RNAlib `vrna_sc_add_up` and `vrna_sc_add_bp`, respectively. We notice that penalties on individual nucleotides are used in several methods developed to account for chemical probing experiments [45, 46] though the way these penalties are computed can differ. Also notice that the most used model to include SHAPE data in secondary structure prediction [15] uses slightly different penalties that are associated to consecutive base pairs rather than to individual ones.

4.5 Neural network

An important ingredient in our procedure is the way experimental data (SHAPE and direct couplings) are mapped into single and pairwise penalties respectively.

The penalties associated with individual nucleotides are mapped from SHAPE reactivities via a single-layered convolutional network:

$$\lambda_i\left(\vec{R}\right) = \sum_{k=-p}^{p} a_k \cdot R_{i+k} + b \tag{5}$$

We include the reactivities of the first p neighbor nucleotides on both the $3'$ and $5'$ sides along the sequence. Hence, the hyperparameter p determines the size of the convolutional window, namely $2p+1$. The parameters a_k of the linear activation function control the relative weights of neighbors, and b is the bias.

The penalties on specific nucleotide pairs are mapped from direct couplings via a double-layered network:

$$\lambda_{ij}\left(J_{ij}\right) = C\cdot\sigma\left(A\cdot J_{ij}+B\right)+D \tag{6}$$

The activation function of the output layer is linear with parameters C and D, whereas we apply a sigmoid activation $\sigma(x) = \frac{1}{1+e^{-x}}$ at the innermost layer, with weight A and bias B.

The model has thus $2p + 6$ free parameters: $\{a_k, b\}$ for the penalties associated to the SHAPE data and $\{A, B, C, D\}$ for those associated to the DCA data.

4.6 Training

The modifications to the model free energy affect the whole ensemble of structures for a given sequence, resulting in modified populations:

$$P\left(\vec{s}|s\vec{e}q, \vec{R}, \vec{J}\right) = \frac{e^{-\frac{1}{RT}F\left(\vec{s}|s\vec{e}q;\vec{R},\vec{J}\right)}}{Z\left(s\vec{e}q, \vec{R}, \vec{J}\right)} \tag{7}$$

Our aim is to enhance the population of the native structure, under the assumption that the native structure is the one obtained by X-ray crystallography. We thus consider a set of given sequence-structure pairs $\{s\vec{e}q, \hat{s}\}$ (one for each system in the training set), where $\hat{s}$ denotes an available crystallographic structure, and for each system we train the model to minimize the cost function

$$\mathcal{C}\left(\{a_k, b\}, \{A, B, C, D\}\right) = -RT \ln P\left(\hat{s}|s\vec{e}q, \vec{R}, \vec{J}\right) \tag{8}$$

Its minimization, in the training procedure, is equivalent to maximizing the population of the target structures.

For each system we decompose the cost function into two terms, namely $F\left(\vec{s}|s\vec{e}q; \vec{R}, \vec{J}\right)$ and $-RT \ln Z\left(s\vec{e}q, \vec{R}, \vec{J}\right)$ that we can compute using, respectively, the functions `vrna_eval_structure` and `vrna_pf` from RNAlib. The derivatives of Eq. (8) with respect to model paramaters, that are required for cost minimization, are proportional to pairing probabilities of

individual nucleotides p_i and of specific nucleotide pairs p_{ij}:

$$
\begin{aligned}
\frac{\partial \mathcal{C}}{\partial a_n} &= RT \sum_{i=1}^{l_{seq}} (p_i - \hat{s}_i) \frac{\partial \lambda_i}{\partial a_n} = RT \sum_{i=1}^{l_{seq}} (p_i - \hat{s}_i) R_{i+n} \\
\frac{\partial \mathcal{C}}{\partial b} &= RT \sum_{i=1}^{l_{seq}} (p_i - \hat{s}_i) \frac{\partial \lambda_i}{\partial b} = RT \sum_{i=1}^{l_{seq}} (p_i - \hat{s}_i) \\
\frac{\partial \mathcal{C}}{\partial A} &= RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij}) \frac{\partial \lambda_{ij}}{\partial A} = RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij}) J_{ij} \cdot C\sigma' (AJ_{ij} + B) \\
\frac{\partial \mathcal{C}}{\partial B} &= RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij}) \frac{\partial \lambda_{ij}}{\partial B} = RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij}) \cdot C\sigma' (AJ_{ij} + B) \\
\frac{\partial \mathcal{C}}{\partial C} &= RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij}) \frac{\partial \lambda_{ij}}{\partial C} = RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij}) \sigma (AJ_{ij} + B) \\
\frac{\partial \mathcal{C}}{\partial D} &= RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij}) \frac{\partial \lambda_{ij}}{\partial D} = RT \sum_{j>i+2}^{l_{seq}} (\hat{s}_{ij} - p_{ij})
\end{aligned}
\tag{9}
$$

These derivatives are then used to back propagate derivatives from the output layer to the input nodes. Here $\hat{s}_i$ and $\hat{s}_{ij}$ represent the pairing state in the reference structure for nucleotide i and for pair ij respectively. Base-pair probabilities in the penalty-driven ensembles

$$
\begin{aligned}
p_{ij} &= \sum_{\{\vec{s}\}} P\left(\vec{s} | s\vec{e}q, \vec{R}, \vec{J}\right) s_{ij} \\
p_i &= \sum_{j=1}^{l_{seq}} p_{ij}
\end{aligned}
\tag{10}
$$

can be straightforwardly computed using the function `vrna_bpp` from RNAlib.

4.7 Regularization

In order to reduce the risk of overfitting we include $l-2$ regularization in the training procedure. Direct couplings (two-dimensional data) and SHAPE profiles (one-dimensional data) differ in the amount of structural information they contain. For this reason, instead of adding to the cost function a standard single regularization term on all parameters, we add two regularization terms, each with an independent coefficient, directly on the penalties

mapped from each type of data:

$$\mathcal{C}\left(\{a_k, b\}, \{A, B, C, D\}\right) = -RT \ln P\left(\hat{s}|s\vec{e}q, \vec{R}, \vec{J}\right) + \\ + \alpha_S \sum_i \lambda_i^2 + \alpha_D \sum_{ij} \lambda_{ij}^2 \tag{11}$$

This procedure keeps the penalties that we add to the model free energy from becoming too large, and thus helps preventing the occurence of overfitting during the minimization of the cost function. The introduction of regularization terms must be taken into account in the cost function derivatives of Eq.(9) by addition of corresponding derivative terms that are easily computed.

4.8 Minimization

The inclusion of regularization terms in the cost function brings in two hyperparameters, α_S and α_D, in addition to p, the hyperparameter that determines the width of the convolutional window. The collection of models that we train is thus defined by the triplet of hyperparameters $\{p, \alpha_S, \alpha_D\}$. We then explore all hyperparameters combination within the ranges $p \in [0, 1, 2, 3]$ and $\alpha_S, \alpha_D \in [\infty, 1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 0.0]$ for a total of $4 \times 7 \times 7 = 196$ models. For each model we minimize the corresponding cost function using the sequential quadratic programming algorithm as implemented in the `scipy.optimize` optimization package [47]. The minimization problem is non-convex whenever α_D is finite, so we expect the cost function landscape to be rough, with multiple local minima. The result of the minimization will thus depend on the initial set of model parameters. For each minimization we try multiple initial values for the model parameters, extracting them from a random uniform distribution, and we select those that yield the minimum cost function. For each minimization we include in the set of starting parameters also three specific sets of starting points:

- parameter values from the optimized $\{p-1, \alpha_S, \alpha_D\}$ model, with the new a_{-p} and a_p set to 0.0; if $p = 0$, we ignore this starting point.
- parameter values from the optimized $\{p, 10 \cdot \alpha_S, \alpha_D\}$ model; if $\alpha_S = 0.0$, we use values from the optimized $\{p, 10^{-4}, \alpha_D\}$ model; if $\alpha_S = 1$, we use values from the optimized $\{p, \infty, \alpha_D\}$ model; if $\alpha_S = \infty$, we ignore this starting point.
- parameter values from the optimized $\{p, \alpha_S, 10 \cdot \alpha_D\}$ model; if $\alpha_D = 0.0$, we use values from the optimized $\{p, \alpha_S, 10^{-4}\}$ model; if $\alpha_D = 1$, we use values from the optimized $\{p, \alpha_S, \infty\}$ model; if $\alpha_D = \infty$, we ignore this starting point.

This ensures that models with higher complexity (i.e., higher p or lower α_S or α_D) will, by construction, fit the data better than models with lower complexity. In this way the performance of the models, as evaluated on the training set, is by construction a monotonically decreasing function of α_D and α_S, and a monotonically increasing function of p.

4.9 Leave-one-out

Among the models optimized in the training procedure, we select the one that yields the best performance without overfitting the training data, in order to ensure the transferability of its structure and optimal parameters. As a test for transferability, we use a leave-one-out cross-validation. This procedure consists in iteratively leaving each of the 8 systems at a time out of the training set, and using the optimal parameters resulting from optimization on the reduced training set to compute the population of the native structure for the left-out system. The population of native structures, averaged on the left-out systems, is used to rank all the tested models. We consider the model with the highest score as the most capable of yielding an increase in population of native structures for systems on which it was not trained.

4.10 Validation

The resulting model is then validated on a set of systems that were not used in the parameter or hyperparameter optimization. For these systems we compute the population of the native structure. In addition, we compute the similarity between the most stable structure in the predicted ensemble (minimum free energy structure) and the native structure using the Mathews correlation coefficient, that optimally balances sensitivity and precision.

References

[1] Cech, T. R. The ribosome is a ribozyme. *Science* **289**, 878–879 (2000).

[2] Doudna, J. & Cech, T. The chemical repertoire of natural ribozymes. *Nature* **418**, 222–228 (2002).

[3] Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**, 423 (2014).

[4] Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* **12**, 641 (2011).

[5] Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).

[6] Tinoco, I. *et al.* Improved estimation of secondary structure in ribonucleic acids. *Nature* **246**, 40–41 (1973).

[7] Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with watson- crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).

[8] Nussinov, R., Pieczenik, G., Griggs, J. R. & Kleitman, D. J. Algorithms for loop matchings. *SIAM J. Appl. Math.* **35**, 68–82 (1978).

[9] Lorenz, R. *et al.* ViennaRNA package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).

[10] Wuchty, S., Fontana, W., Hofacker, I. L. & Schuster, P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**, 145–165 (1999).

[11] Dethoff, E. A., Petzold, K., Chugh, J., Casiano-Negroni, A. & Al-Hashimi, H. M. Visualizing transient low-populated structures of RNA. *Nature* **491**, 724–728 (2012).

[12] Serganov, A. & Nudler, E. A decade of riboswitches. *Cell* **152**, 17–24 (2013).

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

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

[15] Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 97–102 (2009).

[16] Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486 (2015).

[17] Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).

[18] De Leonardis, E. *et al.* Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.* **43**, 10444–10455 (2015).

[19] Weinreb, C. *et al.* 3d RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963–975 (2016).

[20] Lavender, C. A. *et al.* Model-free RNA sequence and structure alignment informed by SHAPE probing reveals a conserved alternate secondary structure for 16s rRNA. *PLoS Comput. Biol.* **11**, 1–19 (2015).

[21] Burley, S. K. *et al.* Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2018).

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

[23] Loughrey, D., Watters, K. E., Settle, A. H. & Lucks, J. B. SHAPE-seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res.* **42**, e165 (2014).

[24] Hajdin, C. E. *et al.* SHAPE-directed RNA structure modeling. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5498–5503 (2013).

[25] Cuturello, F., Tiana, G. & Bussi, G. Assessing the accuracy of direct-coupling analysis for RNA contact prediction. *RNA* 10.1261/rna.074179.119 (2020).

[26] Nawrocki, E. P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2014).

[27] Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).

[28] Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442 – 451 (1975).

[29] Miao, Z. *et al.* RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* **23**, 655–672 (2017).

[30] Kerpedjiev, P., Hammer, S. & Hofacker, I. L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 33773379 (2015).

[31] Weeks, K. M. & Mauger, D. M. Exploring RNA structural codes with SHAPE chemistry. *Acc. Chem. Res.* **44**, 1280–1291 (2011).

[32] Mlýnský, V. & Bussi, G. Molecular dynamics simulations reveal an interplay between SHAPE reagent binding and RNA flexibility. *J. Phys. Chem. Lett.* **9**, 313–318 (2018).

[33] Larkin, M. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

[34] Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F. & Wolfinger, M. T. SHAPE directed RNA folding. *Bioinformatics* **32**, 145–147 (2016).

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

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

[37] Ziv, O. *et al.* Comrades determines in vivo RNA structures and interactions. *Nat. Methods* **15**, 785–788 (2018).

[38] Lu, X.-J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**, e142–e142 (2015).

[39] Turner, D. H. & Mathews, D. H. Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* **38**, D280–D282 (2009).

[40] McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105–1119 (1990).

[41] Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.* **1**, 1610–6 (2006).

[42] Mörl, M. & Schmelzer, C. A simple method for isolation of intact RNA dried from polyacrylamide gels. *Nucleic Acids Res.* **21**, 2016–2016 (1993).

[43] Karabiber, F., McGinnis, J. L., Favorov, O. V. & Weeks, K. M. Qushape: Rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19**, 63–73 (2013).

[44] Aviran, S., Lucks, B. & Patcher, L. RNA structure characterization from chemical mapping experiments. *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 1743–1750 (2011).

[45] Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H. & Clote, P. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS One* **7**, e45160 (2012).

[46] Washietl, S., Hofacker, I. L., Stadler, P. F. & Kellis, M. Rna folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.* **40**, 4261–4272 (2012).

[47] Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* 1–12 (2020).