Decisive Roles of Sequence Distributions in the Generalizability of *de novo* Deep Learning Models for RNA Secondary Structure Prediction

Xiangyun Qiu

Department of Physics, George Washington University, Washington DC 20052

ABSTRACT

The availability of sizeable RNA structure databases and powerful deep learning (DL) frameworks has prompted recent developments of DL models for RNA secondary structure prediction. Taking RNA sequences as the only inputs, the class of *de novo* DL models has demonstrated far superior performances than traditional algorithms. However, key questions remain over the statistical underpinning of such DL models which make no use of co-evolutionary information or physical laws of RNA folding. Here we present a quantitative study of the capacity and generalizability of a series of *de novo* DL models, with a minimal two-module architecture and no post-processing, under varied sequence distributions of seen and unseen datasets. Excellent performances are observed from our models with as few as 16K parameters, affirming their remarkable learning capacity. Our DL models show to generalize well over non-identical unseen sequences, but the model generalizability degrades rapidly as the sequence distributions between the seen and unseen become dissimilar. Examinations of RNA family-specific behaviors manifest not only disparate family-dependent model performances but substantial generalization gaps within the same RNA family. We further determine how model generalization decreases with the decrease of sequence similarity via pairwise sequence alignment, providing quantitative insights into the limitations of statistical learning. Model generalizability thus poses major hurdles for practical uses of current single-sequence-based DL models and we discuss avenues for future advances of such *de novo* DL models for RNA secondary structure prediction.

INTRODUCTION

As a linear chain of nucleotides capable of base pairing, an RNA molecule readily forms various secondary structure motifs such as stems and loops, regardless of the existence of stable tertiary structures (1-3). Particularly for the diverse families of non-coding RNAs (ncRNA) (4), their secondary structures are more conserved than sequences and provide important cues for their biological functions (5,6). And even messenger RNAs (mRNA) possess key secondary structure motifs for translation initiation and regulation (7-9). As such, there have been major interests in determining and understanding RNA secondary structures, via both experiment and computation (10-12). In recent years, with the emergence of sizeable RNA structure databases and the accessibility of powerful machine learning algorithms, data-centric deep-learning-based models, the subject of this study, have been successfully developed for RNA secondary structure prediction (13).

Predicting RNA secondary structures comes down to the identification of all native base pairs. As any two nucleotides (i.e., AUCG) can theoretically pair up (14), the native set of base pairs is regarded as the optimal among all possible sets (10). Traditional computational approaches generally represent a given set of base pairs as a graph with nucleotides as nodes and base pairs as edges. A score is then computed according to pre-defined structural elements of the graph, using parameters derived from physical measurements of thermodynamic energies (15), data mining of known RNA secondary structures, or a combination of both (16). The onus for the predictive task is to optimize the score by searching the entire structure space, which however grows exponentially with the RNA sequence length. In order to reduce the computational complexity, various rules or constraints of RNA secondary structures have been introduced, such as non-nested base pairing (i.e., no pseudoknots), canonical base pairs only (i.e., AU, GC, and GU), and no sharp turns (i.e., loop length > 3). Varieties of efficient algorithms, especially based on dynamic programming and related techniques (17), have also been introduced along with improved scoring parameters (15,16). However, traditional algorithms have struggled to make significant gains in performance in the recent decade (13).

A major recent advance in prediction performance comes from the application of deep learning (DL) to the RNA secondary structure problem. Instead of the graph search in traditional methods, DL models represent the secondary structure as a 2D pairing probably matrix (PPM) and directly predict every PPM_{ij} (*i* and *j* are indices to the nucleotides) in a parallel and quasi-independent manner. In order to directly map an input sequence to its predicted PPM, DL models often employ many abstraction layers enlisting a large number of parameters that must be learned via training on known RNA secondary structures. As existing RNA secondary structure datasets are curated largely via comparative sequence analysis (18,19), this study focuses on the class of single-sequence-based DL models (i.e., without the use of co-evolutionary sequences), also referred to as *de novo* DL models herein. Several recent publications demonstrated that such *de novo* DL models can be successfully developed with up to several millions of parameters, such as ATTfold (20), DMfold (21), E2Efold (22), MXfold2 (23), SPOT-RNA (24), and Ufold (25). These DL models substantially outperform traditional algorithms, with even close-to-perfect predictions for certain benchmarking datasets (e.g., Ufold on the ArchiveII dataset (26)). As a result, single-sequence-based DL models have emerged as a promising and powerful solution to the RNA secondary structure problem.

Despite the recent successes, important questions exist for practical uses of such *de novo* DL models. One is that existing *de novo* DL models have yet to reach satisfactory performances on all known datasets. For example, the best F1 score for the largest dataset, bpRNA (18), remains low, ~0.65, by all DL and traditional models. And this best score is attained by Ufold after incorporation of prior knowledge of RNA structure, specifically with an iterative post-processing module to enforce structural constraints and sparsity. This calls into the question of the learning capacity of *de novo* DL models given the considerable size of 8.6M parameters in the largest model (Ufold). Another related, arguably more critical, issue concerns the generalizability of the *de novo* DL models, i.e., how DL models perform over an unseen/test set of RNA sequences compared with the seen/training set. Substantial performance drops would indicate poor generalizability to which DL models are highly susceptible. To mitigate such

tendency, various strategies have been explored, e.g., SPOT-RNA uses an ensemble of five models and MXfold2 integrates DL-predicted folding scores with thermodynamic regulation. Nonetheless, the generalizability of current DL models strongly depends on the sequence similarities between the seen and unseen datasets (27) and model comparisons are only carried out when the same pair of seen and unseen datasets are used. While this issue is likely inherent to the data-centric statistical learning of all DL models, such non-data agnostic behaviors present major challenges for developing generalizable DL models for RNA secondary structure prediction.

To address these questions, we have investigated the capacity and generalizability of *de novo* DL models with different model sizes and under varied sequence distributions of the seen/unseen datasets. We chose a minimal two-module network architecture without any post processing (except for discretization with the threshold of 0.5) so as to probe the intrinsic behaviors of DL models. We found that a rather small DL model of 16K parameters can achieve decent performances over a medium-sized dataset and that medium-sized models with less than 1M parameters can reach nearly perfect performances and surpass existing DL or traditional models. However, model generalization deteriorates as the sequence similarity between seen and unseen datasets decreases. In order to gain quantitative insight of model generalizability, we determined how model generalization depends on the similarity between the seen and unseen sequences via pairwise sequence alignment. Our observations indicate that *de novo* DL models appear to be largely statistical learners of the correlation between RNA sequences and their secondary structures and we last discuss various pathways to improve model generalizability and develop data-agonistic DL models.

MATERIALS AND METHODS

Two main ingredients of this study are the types of DL network architecture and the sequence distributions of the seen and unseen datasets. We reason that, given an RNA sequence, each nucleotide first explores all its molecular surroundings and then engages in the dynamic process of paring and unpairing before arriving at the most stable configuration. A two-module network

architecture was thus chosen to operate at the sequence and pair levels, respectively. To evaluate model capacity and generalizability, we chose three similarity levels between the seen and unseen datasets as follows. The first level only requires no identical sequences between the seen and unseen sets, i.e., the cross-sequence level. The second level further stipulates that all pairs of sequences in both datasets are below 80% in sequence identity (filtered with the software CD-HIT (28)), referred to as the cross-cluster level. The third level is the most stringent by having the seen and unseen sets from different RNA families, e.g., one from tRNA and the other from 5S rRNA, named as the cross-family level. Brief descriptions of the datasets and the model development are given below and more details are provided in the Suppl. Materials.

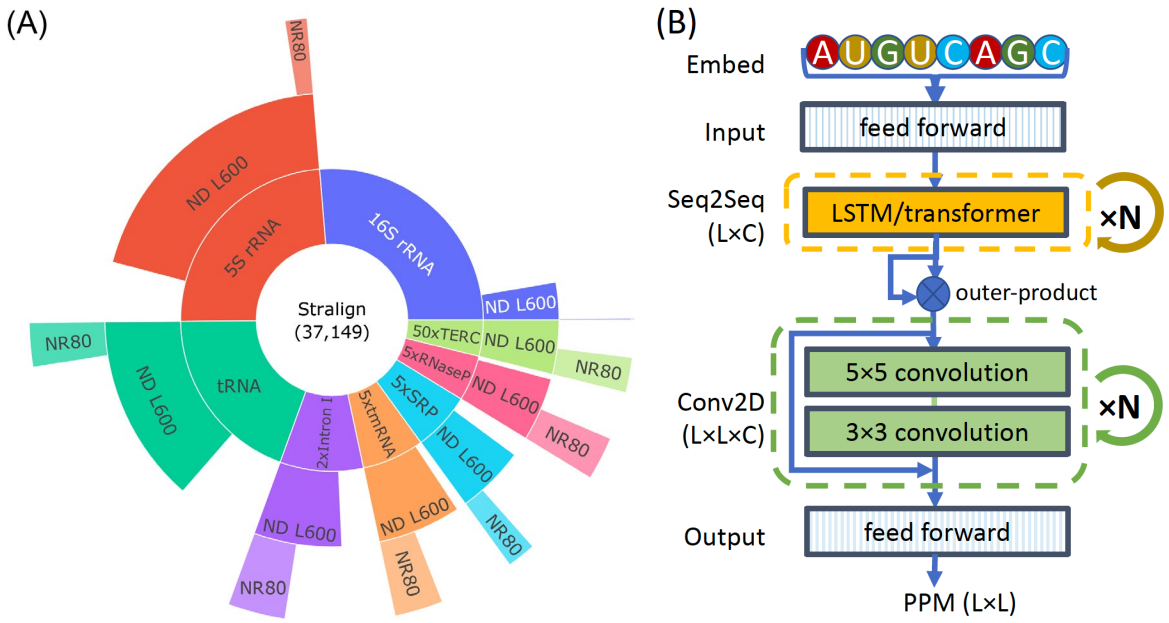

Figure 1. Illustrations of the Stralign dataset and the SeqFold2D architecture. (A) The population and length distributions of eight RNA families in the Stralign dataset at different sequence redundancy levels. The abbreviations of RNA families are, rRNA: ribosomal RNA, tRNA: transfer RNA, Intron I: group I intron, tmRNA: transfer messenger RNA, SRP: signal recognition particle, TERC: telomerase RNA component. The innermost ring shows the population distributions of the RNA families in the original Stralign dataset with a total of 37,149 sequences, noting that the five less represented families (counter-clockwise from Intron I to TERC) are scaled up for visibility and the multiplier N is shown as "N×" in the label (see Suppl. Fig. S1 for the unscaled

version). The middle ring shows the population of each RNA family type at the cross-sequence level after removing duplicated sequences and sequences longer than 600 nucleotides, labelled as “ND L600”. The outermost ring shows RNA family populations at the cross-cluster level after removing redundant sequences with 80% sequence identity or higher, labelled as “NR80”. Note that the number of sequences for the 16S rRNA NR80 set is only 50 and hardly visible. (B) The two-module network diagram of the SeqFold2D models. An input RNA sequence of length L is first one-hot encoded and stacked as an L×12 tensor for k-mer (k=3) representation, followed by a feed-forward layer to yield an L×C tensor. The first main module consists of N stacked blocks of either bidirectional Long-Short-Term-Memory (LSTM) or transformer encoders; the resultant L×C tensor is then transformed into the L×L×C pair representation via outer product, before being fed to the second main module of N stacked blocks of residual 2D convolutional layers. The output block is made up of three feed-forward layers and predicts the pairing probability matrix (PPM) of dimension L×L.

Datasets. With the three similarity levels of sequence distributions in mind, we chose to primarily work with two recent datasets, RNA Stralign curated in 2017 (29) and ArchiveII curated in 2016 (26), both of which are medium-sized, comprehensive databases developed for benchmarking RNA secondary structure predictions. Compared with the only other larger comprehensive collection (bpRNA (18), also used later in this study), RNA family types are conveniently provided by Stralign and ArchiveII, facilitating cross-family examinations. Both datasets have also been used by several DL models (e.g., E2Efold, MXfold2, and Ufold) with pre-trained model parameters available. For training efficiency and consistency with other DL models, RNA sequences longer than 600 nucleotides are excluded. For the cross-sequence level, only duplicated sequences are deleted from each dataset, yielding Stralign ND-L600 and ArchiveII ND-L600 with 20,118 and 3,395 sequences, respectively. CD-HIT is then used to remove sequences with over 80% sequence identity, yielding the Stralign NR80-L600 and ArchiveII NR80-L600 datasets for the cross-cluster study. The ArchiveII NR80-L600 sequences with over 80% identity to Stralign NR80-L600 are further removed to give the ArchiveII-Stralign NR80-L600 dataset. Fig. 1A shows the population distributions of the eight RNA families of the

Stralign dataset at different sequence similarity levels, evidencing vastly different redundancies for different families. For the cross-family study, the Stralign and ArchiveII datasets are combined to create the Strive dataset, which is processed in the same way to obtain its ND-L600 and NR80-L600 subsets. More information on these and other datasets (e.g., bpRNA) is given in the Suppl. Materials .

Network architecture. As shown in Fig. 1B, our DL neural network, named SeqFold2D, comprises two main learning modules flanked by the input and output blocks. As the only input to the SeqFold2D network, each RNA sequence of length L is first one-hot-encoded as an L×4 tensor and then stacked into its k-mer (k=3) representation of L×12. The input block mixes the twelve channels into an L×C tensor with two feed-forward layers. The channel size C is then kept constant for the rest of the network unless noted otherwise. The first main learning module is made up of N repeated blocks of bidirectional Long-Short-Term-Memory (LSTM) or self-attention-based transformer encoder layers to learn richer representations from the whole sequence, which are then transformed into 2D pair representations (L×L×C) via outer-product. The second learning module consists of the same number (N) of repeated blocks of residual 2D convolutional layers to learn pair interactions locally. The output block consists of three feed-forward layers with the channel size C=2 for the final layer. Softmax is then used to yield the continuous PPM of shape L×L. Non-linear activations (LeakyReLU or Swish) are applied before every linear operation with weights and biases, followed by layer normalizations and dropouts. The size of a specific SeqFold2D network is thus determined by two design variables, N (the number of blocks) and C (the channel size). For example, N=1 and C=16 gives ~16K trainable parameters and N=4 and C=64 gives ~960K parameters.

Performance metrics and benchmarking. We use the F1 score as the main metric for evaluating model performances (30). It is the harmonic mean of Precision and Recall and defined as $\frac{2\times Precision\times Recall}{Precison+Recall} = \frac{2\times TP}{2\times TP+FP+FN} = \frac{2\times TP}{L^2+TP-TN}$, where Precision is TP/(TP+FP), Recall TP/(TP+FN), TP the number of true positives, FP false positives, FN false negatives, and TN true negatives. In order to examine the exclusive capabilities of DL models, no post-processing is applied to the continuous PPM except for discretization with the threshold of 0.5, i.e., no grid search, model

ensemble averaging, thermodynamic regulation, or enforcement of RNA structure constraints. In addition to several *de novo* DL models, we further benchmark SeqFold2D models against the following traditional algorithms: Mfold/UNAfold (31), RNAfold (32), RNAstructure(33), LinearFold (34), SimFold (35), CONTRAfold (36), ContextFold(37), and Centroidfold (38). Note that we categorize machine-learning-based hybrid models as traditional algorithms for wording simplicity.

Model development and evaluation of overfitting and generalizability. All SeqFold2D models were implemented with the Paddle framework (https://github.com/PaddlePaddle/Paddle) and trained with the AdamW optimizer in two stages. The first stage uses the binary cross-entropy (CE) loss between the predicted PPM and the ground truth with equal weights for positive and negative labels. Note that E2Efold and Ufold over-weighted the positive labels with a constant of 300 to address the issue of imbalanced negative/positive distributions, which we found to be helpful in the beginning but lead to higher ratios of false positives in the end. Once the CE loss plateaus, the second stage is invoked with the soft F1-score loss used by E2Efold and others. The two-stage training was found to give the best F1 score compared with the use of either loss function alone. For hyperparameter tuning, we carried out limited manual searches for the learning rate, batch size, and dropout rate. No significant differences were found between the uses of LSTM or transformer layers for the first module, though the LSTM layers led to more consistent training outcomes.

We refer to the set of sequences used to develop a specific SeqFold2D model as the seen set that consists of two subsets, training and validation. The training set is used to train model weights and biases and the validation set is used to optimize model hyperparameters such as the number of epochs, weight decay, and learning rate. We refer to the test set of sequences for benchmarking models also as the unseen set, which may have the same or different sequence distributions from the seen set. These different sets of sequences are obtained via random splits stratified by the family type from their corresponding parent dataset.

Model performances on the training (TR), validation (VL), and test (TS) sets can all be different, usually with the best for TR and the worst for TS. In order to take a closer look into model behaviors, we further distinguish two different performance gaps. The first is between the TR and VL sets that are always taken from the same parent dataset in this study. We reason that the TR-VL variance indicates whether a DL model is learning both TR and VL sets (i.e., no overfitting) or additional spurious patterns of the TR set only (i.e., overfitting). The TR-VL variance is thus used as the indicator for model overfitting. The second is between the TR and TS sets that may have similar or different sequence distributions. Hence the TR-TS variance reflects model generalizability at specific levels of sequence similarities and is also referred to as the generalization gap. Importantly, every SeqFold2D model herein was trained to optimize the F1 score on its corresponding VL set, and the model parameters giving the highest VL F1 score were saved as the final model to be benchmarked on the unseen TS set. One alternative is to stop training as soon as the TR-VL variance becomes significant (e.g., 5%), which we chose not to follow given our purpose of investigating model capacity and generalizability.

RESULTS

Cross-sequence study: excellent capacity and generalizability of *de novo* DL models

We first use the larger Stralign ND-L600 dataset (Stral-ND in short) to train and test SeqFold2D models of various sizes at the cross-sequence level. The TR, VL, and TS sets are randomly split from Stral-ND at the ratios of 70%, 15%, and 15%, respectively. By varying the number of blocks (N) and the channel size (C), we gradually increase the number of model parameters from ~16K (N=1, C=16) to ~960K (N=4, C=64). Fig. 2A shows the F1 scores for the TR and TS sets from five SeqFold2D models and selected traditional models. Note that other DL models are not included here as they were trained with different dataset configurations to be examined next. As the model size increases, the performances of the SeqFold2D models increase steadily for both TR and TS and surpass all other algorithms by a large margin. Notably, the smallest SeqFold2D model with 16K parameters achieves an F1 score of ~0.8 and the largest model, SeqFold2D-960K, attains a nearly perfect performance (F1~0.985). Meanwhile, the TR-TS generalization

gaps show to be negligible for the SeqFold2D models of sizes up to 420K parameters. A slight drop (~1.5%) in the F1 score can be spotted for the SeqFold2D-960K model (the top bars in Fig. 2A). As the F1-score distributions are non-Gaussian, we apply the Kolmogorov-Smirnov (KS) test and find the TR-TS variance to be indeed statistically significant for the 960K model (P-value around 3e-6) but insignificant for all other SeqFold2D models (P-values greater than 0.1). Nonetheless, the TR-TS generalization gaps are rather small and we further verify that the TR-VL variances show similar behaviors (see Suppl. Fig. S8). Altogether, the SeqFold2D models demonstrate excellent expressive capacities for mapping RNA sequences to their secondary structures and generalize to high degrees of performances at the cross-sequence level (e.g., F1~0.97 on the TS set by SeqFold2D-960K).

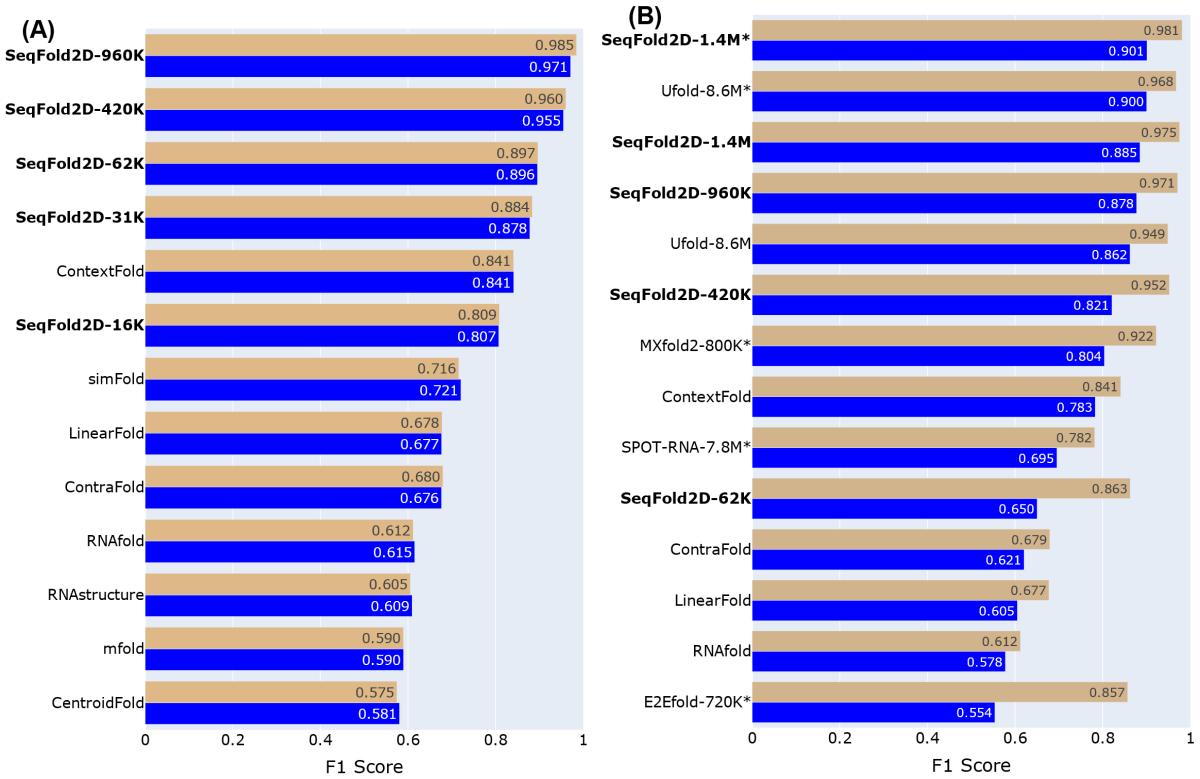

Figure 2. The F1 scores on the TR (tan) and TS (blue) sets by SeqFold2D models and other selected DL and traditional models in two different dataset setups. The number of parameters is shown after the name and the F1 value is given at the end of each bar. (A) Both TR and TS sets are taken from the Stralign ND-L600 dataset. (B) The TR set is taken from the Stralign ND-

L600 dataset and the TS set is the entire ArchiveII ND-L600 dataset. The asterisk after the name indicates the use of post-processing.

In order to compare with other DL models such as MXfold2 and Ufold, we next follow the same dataset setup of using the Stral-ND dataset as the TR and VL sets and the ArchiveII-ND-L600 (Archi-ND in short) as the TS set. Fig. 2B shows the F1 scores from several DL and traditional models, including five SeqFold2D models with 62K to 1.4M parameters. Note that the only available SPOT-RNA model was trained with the bpRNA dataset and its performance should not be directly compared. As previously reported for the Archi-ND dataset (22,23,25), all DL models, including the SeqFold2D models, outperform traditional algorithms substantially and attain F1 scores as high as ~0.9. Notably, both E2Efold and Ufold post-process predicted PPMs by excluding non-canonical base pairs and sharp turns and enforcing sparsity through iterative refinement. In part promoted by the superior performance of Ufold over all SeqFold2D models on the TS set, we examined the effectiveness of such post-processing by comparing the performances of Ufold with and without post-processing, shown as Ufold-8.6M* and Ufold-8.6M in Fig. 2B, respectively. Considerable gains in F1 scores were observed, ~2% for Stral-ND and ~4% for Archi-ND. We subsequently experimented with the same training and post-processing steps for the SeqFold2D-1.4M* model and realized similar performance gains, e.g., producing the best F1 scores for both Stral-ND (0.981) and Archi-ND (0.901) among all models. However, such post processing requires prior knowledge of the test set since the Archi-ND dataset indeed has canonical base pairs only and no sharp turns. We therefore did not use post-processing for all other SeqFold2D models. Overall, the SeqFold2D models show to compare favorably against other *de novo* DL models for both datasets, albeit with fewer parameters.

Another emphatic trend in Fig. 2B is that all DL models, including SeqFold2D, exhibit significant generalization gaps between Stral-ND and Archi-ND, with F1-score falloffs ranging from 8% to nearly 40%. This also holds true for MXfold2 with thermodynamic regularization and SPOT-RNA trained on different datasets. However, we observed very little TR-VL variances in training all SeqFold2D models (Suppl. Fig. S9). We also carried out five-fold cross-validation to rule out the likelihood of fortuitous data splitting. Therefore, the SeqFold2D models are not

overfitting the TR set over the VL set but faithfully describing the entire Stral-ND distribution. Different sequence distributions between Stral-ND and Archi-ND thus remain the most plausible cause for the observed generalization gap, which is corroborated by even larger performance drops for more dissimilar datasets such as bpRNA. On the whole, Stral-ND and Archi-ND have nearly identical RNA family types (the only exception is 23S rRNA in Archi-ND only, but at a mere 0.6%), but with very different population distributions of RNA families (Suppl. Fig. S5). This motivated us to examine the RNA family-specific performances of DL models, reasoning that Archi-ND may happen to have higher fractions of more difficult RNA families.

Fig. 3 shows the F1 scores per RNA family on Stral-ND and Archi-ND for the SeqFold2D-1.4M model. We indeed observe divergent family-specific performances, for example, F1~0.998 for tRNA and F1~0.764 for telomerase RNA (TERC). The general trend is the better performances for the more populous families as expected from the statistical nature of DL model training. However, instead of similar family-wise performances, large drops for the same RNA family are observed for several families, with the largest being ~41% for 16S rRNA and a significant ~5% for tRNA (0.998 to 0.952). We further verified that another DL model (Ufold-8.6M*) manifests the same qualitative family-wise performance differences (Suppl. Fig. S10). It is worth noting that some families (e.g., tmRNA and TERC) show almost no difference and our subsequent analyses indicate their sequences are highly redundant across Stral-ND and Archi-ND. In addition, the SeqFold2D-1.4M model gives nearly identical F1 scores between the Stral-ND TR and VL sets for all RNA families (Suppl. Fig. S11), ruling out the possibility that some RNA families are overfit and some are not. Altogether, these lead to the somewhat surprising observation that the *de novo* DL models are not guaranteed to generalize within the same RNA family type that is supposedly made up of closely related sequences.

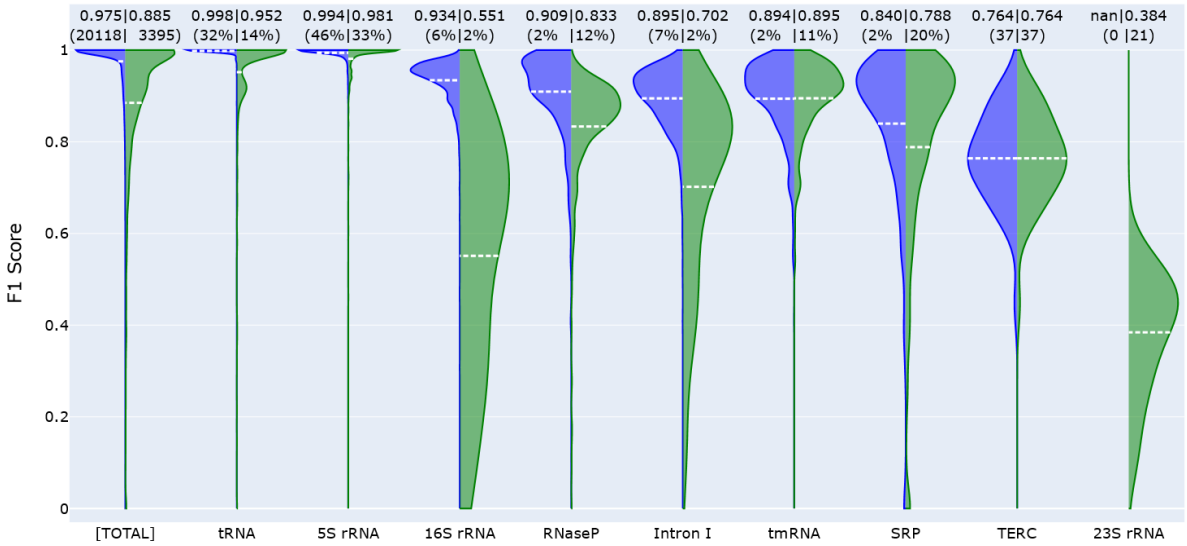

Figure 3. The F1 scores on the seen set (Stralign ND-L600, left in blue) and the unseen/test set (ArchiveII ND-L600, right in green) by the SeqFold2D-1.4M model. The leftmost pair of violins show the F1 scores for the whole seen and unseen set and the following violin pairs show the F1 scores for each constituent RNA family. The averaged F1 scores are shown at the top and also as dashed lines (white) in the corresponding violins. The values in the parentheses above the violins are the sequence counts in actual numbers (for the whole set or families with less than 1% population shares) or in percentages (for families with more than 1% population shares). Note that 23S rRNA only exists in the ArchiveII ND-L600 dataset and is thus shown as nan for the Stralign ND-L600 dataset.

Cross-cluster study: propensity to overfitting and degraded generalizability

The sequence distributions within the same RNA family can be highly uneven when only identical sequences are removed (i.e., at the cross-sequence level). One way to mitigate such imbalance is to cluster similar sequences and remove all but one redundant sequence from each cluster. We applied this procedure using CD-HIT with the cluster similarity cutoff of 80%, the lowest allowed by CD-HIT, and obtained the non-redundant Stral-NR80 (3,122 RNAs) and Archi-Stral-NR80 (433 RNAs) datasets as aforementioned. Out of curiosity, we verified that all inter-family sequences are below 80% similarity level for all datasets. With Stral-NR80 as the

seen (TR and VL) set and Archi-Stral-NR80 as the unseen (TS) set, several SeqFold2D models of sizes 400K-1.4M were trained for this cross-cluster study. As both intra- and inter-family sequence distributions are drastically changed by the de-redundancy procedure, model performances exhibit broad changes at the ensemble and family-specific levels. On the whole, all SeqFold2D models yield noticeably lower F1 scores on both seen and unseen datasets compared with the cross-sequence study, which can be attributed to the fact that the removed sequences are redundant and preferentially well-fitted. For each RNA family, its performance gain/loss is generally correlated with the increase/decrease of its population share in the dataset, consistent with the effects of observational bias on DL model development.

However, this cross-cluster study witnesses the emergence of significant TR-VL variances (i.e., overfitting) and further increases of the TR-TS variances (i.e., degraded generalizability) at both ensemble and family-specific levels for all SeqFold2D models, as shown in Suppl. Figs. S12&13. The TR-VL and TR-TS variances can be reduced with stronger regularization methods (e.g., larger dropout and weight decay rates). We carried out limited explorations but observed poorer overall performances for all datasets. Hence the SeqFold2D models presented here were developed with the same model hyperparameters as in the cross-sequence study. Rather curiously, the TR-VL variances typically appear when the F1 scores reach ~0.8 and further training improves the F1 scores for both datasets, but at a faster rate for the TR set. It is thus suggested that, in the early training phase, the directions with the deepest gradient descent are given by sequence-structure correlations common to the TR and VL sets. The gradients become shallower as the training progresses and the models start fitting spurious patterns in the TR set only while continuing to improve on the VL set. Consequently, even larger TR-TS variances are observed (Suppl. Fig. S11) given the more dissimilar sequence distributions between Stral-NR80 and Archi-Stral-NR80 than between the TR and VL sets taken from the same Stral-NR80 dataset. At the RNA family level, Fig. 4 shows the TR-TS variances for the SeqFold2D-1.4M model for individual families, evidencing substantial generalization gaps for all families existing in both datasets. Overall, despite the excellent performances on the TR and VL sets (with the best F1 scores of 0.956 and 0.905, respectively, see Suppl. Fig. S12), the performances on the TS set are rather unsatisfactory on the absolute scale (the best F1 score of 0.762) and some RNA families

have F1 scores close to or far below 0.5 (e.g., 16S and 23S rRNAs). These strongly suggest that that *de novo* DL models developed and tested with non-redundant sequences may generalize rather poorly over other non-redundant sequences.

To verify that the observations are not due to the relatively smaller model sizes or other insufficiencies of the SeqFold2D architecture, we trained several SeqFold2D models on the bpRNA TR0 and VL0 datasets and tested them with the bpRNA TS0 datasets. The three bpRNA sets (TR0, VL0, and TS0, a total of 13,419 structures) are all below 80% sequence similarity filtered with CD-HIT and much larger than the Stral-NR80 dataset (3,122 structures). When benchmarked against other models (Suppl. Fig. S14), the smallest SeqFold2D model with 960K parameters achieves an F1 score of 0.635 on the bpRNA TS0 dataset, only inferior to the Ufold-8.6M* model with the score of 0.654. Larger SeqFold2D models generally lead to better performances, though with only marginal gains for the VL0 and TS0 datasets. Still, the largest SeqFold2D-3.5M model (N=7 and C=96) attains an F1 score of 0.665 on the bpRNA TS0 set, the highest reported to date. Despite the still relatively poor scores on the TS0 set, rather high F1 scores are observed on the TR set with over 10K non-redundant sequences, 0.812 and 0.903 by SeqFold2D-960K and SeqFold2D-3.5M, respectively. This again testifies to the remarkable expressive capacity of DL models. However, even greater levels of TR-VL and TR-TS variances are also observed with the bpRNA datasets whose sequence distributions are expected to be more diverse. Similar to the model training with Stral-NR80, TR-VL variances only appear in later training phases, though with an onset at a much smaller F1 value of ~0.55. Continued training increases the TR-VL variances but also the performances over the VL set (e.g., the best F1 scores of 0.657 and 0.686 for SeqFold2D-960K and SeqFold2D-3.5M, respectively). Note that the SeqFold2D models show ever worse generalizability for the bpRNA-NEW dataset and we plan to use data augmentations techniques demonstrated by Ufold to improve generalizability in future work. Importantly, the experimentation with the bpRNA dataset corroborates the challenges of *de novo* DL models in generalizing over non-redundant sequences.

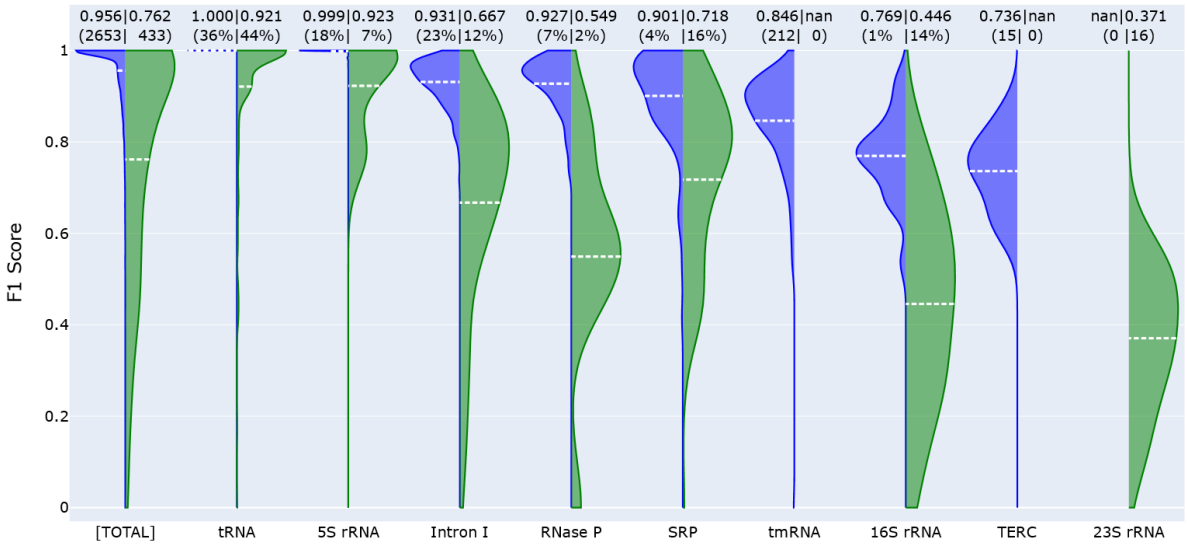

Figure 4. The F1 scores on the TR set from Stralign NR80-L600 (left, blue) and the TS set Archi-Stral-NR80 (right, green) by the SeqFold2D-1.4M model. The specific annotations follow that in Fig. 3. Note that tmRNA and TERC are absent in Archi-Stral-NR80 due to their high redundancies between the Stralign and ArchiveII datasets and that 23S rRNA only exists in the Archi-Stral-NR80 dataset.

Pairwise sequence alignment: decisive roles of sequence similarity in model generalization

Our observations so far indicate that the mode performance for a test sequence is strongly correlated with its similarity to the training set. In the case of high similarity as in the cross-sequence study, none or very little performance gaps can be obtained for non-identical sequences. On the other hand, the low sequence similarity in the cross-cluster study results in large overfitting and generalization gaps. Here we aim to probe the correlation between model generalization and sequence similarity quantitatively. With the model performance quantified by the F1 score, we carried out pairwise sequence alignment (PSA) as the means to assess the similarity between a test sequence and the training set, as described below. The SeqFold2D-1.4M model with Stral-NR80 as the unseen and Archi-Stral-NR80 as the seen datasets is analyzed here because of their relatively small numbers of sequences with broad distributions of F1 scores.

For a given test sequence, it is first aligned against every sequence in the seen set (Stral-NR80, 3122 sequences in total) to produce 3122 PSAs. A percentage sequence identity (PSI) score is then calculated from each PSA as the number of identical aligning nucleotides divided by the average length of the sequence pair. Note that we experimented with alternative PSI definitions (e.g., using the shorter length of the sequence pair) and observed qualitatively consistent behaviors. As the training set consists of highly diverse sequences across eight RNA families, the 3122 PSI scores are broadly distributed. Now with the pairwise PSI values of the test sequence to the entire seen set, the question we next ask is, can we use the subset of seen sequences above a certain PSI value to inform the model performance for the test sequence? One extreme case would be setting the PSI threshold to 1.0 (i.e., using a seen sequence identical to the test sequence), where the same performances are granted. Lowering the PSI threshold gradually includes more dissimilar sequences from the seen set and the informative power is expected to decrease. As such, the dependence of the informative power on the PSI threshold can provide quantitative insights into the model generalizability. Specifically, for the subset of seen sequences above a given PSI threshold, we average their F1 scores weighted by the PSI values to obtain the F1-seen score as the surrogate of their informative power. The F1-seen score is then compared with the actual F1 score for the test sequence (F1-unseen) at different PSI thresholds.

The entire process described above is repeated for all test sequences in the Archi-Stral-NR80 dataset (433 in total). As a result, for each PSI threshold, we obtain 433 pairs of F1-seen and F1-unseen values and examine their statistical correlations with rubrics such as the Pearson correlation coefficient (PCC). In order to rule out coincidental statistics, three different PSA programs were used: Foldalign (39), LaRA2 (40), and LocaRNA (41). Figs. 5A&B show the ratios and PPC values between the F1-test and F1-seen scores as a function of the PSI threshold. Consistent trends are given by the three PSA programs, supporting the robustness of observed dependencies. It is worth noting that all three programs find seen sequences with PSI values higher than 0.8, somewhat unexpected because of the below 80% redundancy between Stral-NR80 and Archi-Stral-NR80 filtered by CD-HIT. We speculate that this is caused by algorithmic differences as CD-HIT employs a greedy incremental clustering method that first estimates

sequence similarity via short-world counting (word length of five used in this study as recommended for 80% cutoff). This may lead to missed matches by CD-HIT as the PSA programs always carry out actual sequence alignment. It is thus suggested that additional de-redundancy steps (e.g., with BlastN (42), Infernal (43)) may be warranted for situations where sequence non-redundancies are critical. For our purpose, it is reassuring that the F1 ratios and PCC values both approach the asymptotic value of 1.0 as the PSI threshold approaches 1.0. Thus no further de-redundancy was carried out.

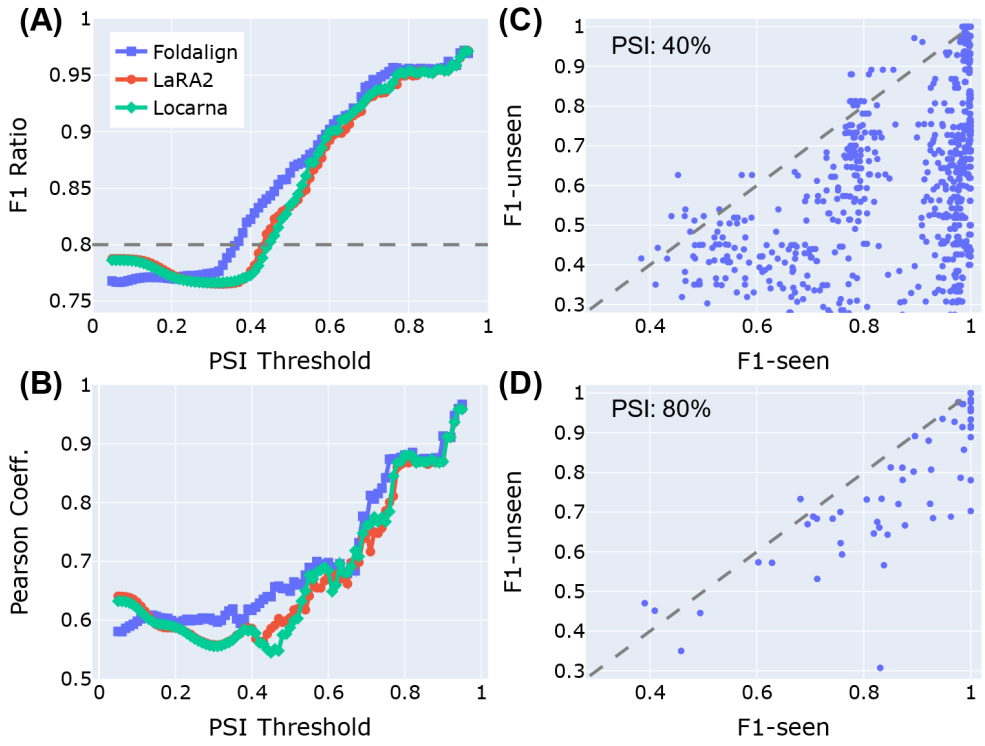

Figure 5. Illustrations of the correlations between the F1-unseen and F1-seen scores. (A) The F1-unseen over F1-seen ratios as a function the PSI thresholds. The horizontal dashed line marks the F1 ratio between the entire unseen and seen datasets. (B) The PCC values as a function of the PSI thresholds. (C) The distributions of the F1-unseen and F1-seen scores at the nominal PSI threshold of 40% by Foldalign. (D) The distributions of the F1-unseen and F1-seen scores at the nominal PSI threshold of 80% by Foldalign. Note that it is common to find no seen sequences for a test sequence above a high PSI threshold, leading to many null F1-seen values that are absent in (D).

The main observation from Fig. 5 is the rapid declines of both F1 ratios and PPC values with the decrease of sequence similarity. In the direction of decreasing PSI thresholds, both F1 ratios and PCCs start out in the high 90%s, affirming the excellent generalizability over non-identical but similar sequences observed in the cross-sequence study. The F1 ratios and PCCs then quickly drop as the PSI threshold is lowered, indicating a fast decay of generalizability over increasingly dissimilar sequences. The F1 ratio approaches its ensemble value (i.e., the F1 ratio between the entire unseen and seen datasets) around PSI ~ 0.4, in spite of the existence of significant PCC values of ~0.6 for all curves. Representative distributions of F1-seen and F1-unseen values at two PSI thresholds are shown in Figs. 5C&D. Overall, considering both F1 ratios and PCC values, the informative power of the seen set falls below PCC ~ 0.8 at PSI ~0.7 and diminishes completely at PSI ~0.4. While the exact PSI transition points likely depend on the specific datasets used, the PSI definition, etc., our observations provide quantitative insights on how the generalizability of *de novo* DL models depends on the sequence distributions of the seen and unseen datasets.

Cross-family study: inability to generalize over unseen RNA families

An even more stringent examination of model generalizability is to train DL models with one RNA family and test with other types. While all *de novo* DL models are expected to fail such a cross-family test, we experimented with a SeqFold2D-400K model trained on the most abundant family type, tRNA. To maximize the number of tRNA sequences, the Stralign and ArchiveII datasets were combined into a single Strive dataset with a total of nine families (see Suppl. Fig. S5). Training was relatively straightforward given the high similarity levels of the tRNA sequences. To limit overfitting, we stopped training when the TR-VL variance became significant around the F1 score of 0.975. As shown in Suppl. Fig. S15, the SeqFold2D-400K model fails completely on all other seven RNA families, yielding the F1 scores ranging from 0.03 to 0.1. We did not pursue the use of other families for training in part due to their relatively smaller numbers. We like to note a recent proposal of a different cross-family test by only excluding the test family type from the seen set (27). It was shown that *de novo* DL models performed poorly on such cross-family tests, consistent with our observations.

CONCONCLUSION AND DISCUSSION

In this study, we set out to investigate the expressive capacity and generalizability of *de novo* DL models by varying their sizes and the sequence distributions of seen and unseen datasets. To this end, we design the SeqFold2D architecture with a minimal two-module structure and without any post-processing steps such as model averaging, grid search, and enforcement of RNA structural constraints. The SeqFold2D models demonstrate remarkable learning power of the relationships between RNA sequences and secondary structures by achieving excellent performances for all training sets, e.g., the SeqFold2D-960K model attains the F1 scores of 0.985, 0.971, and 0.812 for the Stral-ND, Stral-NR80, and bpRNA TR0 sets, respectively. This is especially notable for the bpRNA TR0 dataset comprising over 10K non-redundant sequences. For the unseen/test sequences, the SeqFold2D models outperform all other *de novo* DL models and traditional algorithms, despite often with much fewer parameters and without post-processing. However, while the SeqFold2D models are able to generalize well over unseen sequences at the cross-sequence level, significant performance gaps appear at the cross-cluster level, manifested as sizeable TR-VL variances (overfitting) and even larger TR-TS variances (poor generalizability). With the help of pairwise sequence alignment, we quantitatively analyze how model generalizability depends on the sequence similarity between the seen and unseen sequences and reveal a rapid decline in performance for increasingly dissimilar sequences. Altogether, these observations suggest that the class of *de novo* DL models are largely statistical learners of sequence and structure pattern mappings and that their applicability in the real world is determined by the sequence distributions of the seen and unseen/test datasets.

Our study can be extended in several important directions. First, the SeqFold2D models were tuned to maximize the performances on the VL set rather than to minimize TR-VL or TR-TS variances (e.g., via early stopping or stronger regularization). We found this strategy also led to best performances on the TS set, which can be argued to be optimal for the purpose of benchmarking. In essence, the models were fitting both general and spurious patterns at the same time in later training phases and hence certain levels of overfitting are beneficial to performance. We also do not fully understand why a SeqFold2D model of the same size would

overfit the larger bpRNA TR0 more than the Stralign NR80 dataset. Nonetheless, model hyperparameters can be further explored to better balance performances and generalizability. Second, we primarily focus on the F1 score as the sole metric for model performance and have yet to analyze other metrics such as Precision, Recall, pseudo-knot prediction, and long-range contact prediction that have been discussed by other DL models. Detailed analyses of the SeqFold2D model predictions are work in progress. Lastly, it would be highly desirable to interpret how the *de novo* DL models predict secondary structures. Existing DL models encompass a wide range of network architectures, e.g., U-net by Ufold, 2D-LSTM by SPOT-RNA, and 1D-LSTM by MXfold2. We experimented with both 1D-LSTM and self-attention-based transformer layers and found similar performances. Various techniques from the machine learning field can be used to inspect the hidden 1D and 2D representations and probe the inner workings of these diverse DL models (44).

Furthermore, continued developments of such *de novo* models, holding the advantage of requiring no more information other than nucleotide sequences, are much needed to improve both performance and generalizability. For example, with or without post-processing, no DL models have accomplished satisfactory performances on the bpRNA TS0 set, with the highest F1 score being 0.665 by the SeqFold2D-3.5M model in spite of an F1 score of 0.903 on the bpRNA TR0 set. Importantly, this study indicates that the statistical nature of *de novo* DL models presents the fundamental challenge for generalizability. As such, the ideal solution is to increase the number and distribution of training structures. Direct experimental determinations of RNA secondary structures are however slow and costly and high-throughput measurements such as PARS (45) and SHAPE (46) usually yield nucleotide-level activity profiles rather than base pairing partners. Note that high-throughput experimental data can serve as valuable constraints for structure prediction (47,48) . Given that comparative sequence analysis is the main method of data curation for large structure databases, renewed efforts with more sophisticated pipelines such as RNAcmap (49) and rMSA (https://github.com/kad-ecoli/rMSA2) could also be viable. On the other hand, before such structure data become available, wide ranges of resources and techniques in the fields of biology and machine learning may be

exploited to develop highly accurate and generalizable *de novo* DL models, some of which are discussed below.

First, several directions can be pursued to enrich the input sequence information beyond one-hot identifiers. Chemical and physical properties of the nucleotides can be embedded to capture their intrinsic features particularly about the different bases, though one potential caveat is that nucleotides are much less diverse compared with amino acids. Moreover, unsupervised pre-training of known RNA sequences with natural language models such as BERT (50) may offer a powerful pathway for representation learning of RNA sequences. One example of this kind is the recent DNABERT model pre-trained on human genome sequences (51). In the case of RNA, non-coding RNA sequences in the RNAcentral database (52) can be used for pre-training such that richer representations can be learned from all non-coding RNAs. Second, avenues can be explored to infuse the mechanisms of RNA structure and folding (3) into the DL architecture. At the sequence and pair-interaction levels, one remarkable example is the EvoFormer module in the Alphafold2 architecture (53) that facilitates cross-level information exchanges and builds in inductive biases such as triangular updates. A recent work, DeepFoldRNA, demonstrated that an EvoFormer-like architecture also works well for RNA (54). This however requires the use of multiple-sequence alignment and it remains to be seen whether good effects can be obtained without co-evolutionary information. Alternatively, it may be possible to enforce learning biases in the form of auxiliary loss functions similar to the Physics-Informed Neural Networks (PINN) (55). For example, the structural constraint of no base multiples can be enforced by a term $ReLU(\sum_{j=1}^{L} ThresholdedReLU\left(PPM_{ij}, 0.5\right) - 1.0)$ and sharp turns can be penalized through a masked loss of the diagonal elements. Another important aspect of RNA secondary structures is the topology (56). A *posteriori* method is to increase the weights of loop opening and closing base pairs in the loss function. Lastly, model outputs can go beyond RNA secondary structures. Additional structural properties can be predicted at the nucleotide level, such as nucleotide-wise solvent accessibility or activity profiles that can be measured independently (e.g., via SHAPE). Thermodynamic quantities at the sequence level such as free energies and melting temperatures may also be predicted. Perhaps one most practical model prediction would be the confidence levels of the predicted

secondary structures, like the pTMscore by Alphafold2. As such, we experimented with the addition of a third module to predict the F1 score which however also showed poor generalization power despite superior performances on the seen sequences.

In closing, RNA secondary structures are important biochemical and physical features of all RNAs, coding or non-coding, and they often play critical roles in the biological functions of RNA molecules. The challenges of predicting RNA secondary structures *de novo* have inspired the development of a number of DL models that demonstrate unprecedented expressive power as well as heavy reliance on broadly distributed training data. This study provides quantitative analyses of the model capacity and generalizability in the context of different sequence distributions. And various pathways for future advances are discussed so as to catalyze the development of next-generation *de novo* DL models for RNA secondary structure prediction.

ACKNOWLEGEMENT

We thank Dr. Chengxin Zhang for the suggestion of LaRA2 for pairwise sequence alignment and the GW High Performance Computing Facility for the access of CPU and GPU nodes.

Bibliography

1. Higgs, P.G. (2000) RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.*, **33**, 199-253.
2. Fallmann, J., Will, S., Engelhardt, J., Gruning, B., Backofen, R. and Stadler, P.F. (2017) Recent advances in RNA folding. *J. Biotechnol.*, **261**, 97-104.
3. Chen, S.J. (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annu Rev Biophys*, **37**, 197-214.
4. Cech, T.R. and Steitz, J.A. (2014) The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, **157**, 77-94.
5. Diederichs, S. (2014) The four dimensions of noncoding RNA conservation. *Trends Genet.*, **30**, 121-123.
6. Bevilacqua, P.C., Ritchey, L.E., Su, Z. and Assmann, S.M. (2016) In Bonini, N. M. (ed.), *Annual Review of Genetics, Vol 50*, Vol. 50, pp. 235-266.
7. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. and Weissman, J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701-705.
8. Mauger, D.M., Cabral, B.J., Presnyak, V., Su, S.V., Reid, D.W., Goodman, B., Link, K., Khatwani, N., Reynders, J., Moore, M.J. *et al.* (2019) mRNA structure regulates protein expression through changes in functional half-life. *Proc Natl Acad Sci U S A*, **116**, 24075-24083.

9. Ermolenko, D.N. and Mathews, D.H. (2021) Making ends meet: new functions of mRNA secondary structure. *Wiley Interdisciplinary Reviews: RNA*, **12**, e1611.
10. Seetin, M.G. and Mathews, D.H. (2012) RNA structure prediction: an overview of methods. *Methods Mol Biol*, **905**, 99-122.
11. Deigan, K.E., Li, T.W., Mathews, D.H. and Weeks, K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A*, **106**, 97-102.
12. Eddy, S.R. (2014) Computational Analysis of Conserved RNA Secondary Structure in Transcriptomes and Genomes. *Annual Review of Biophysics*, **43**, 433-456.
13. Zhao, Q., Zhao, Z., Fan, X., Yuan, Z., Mao, Q. and Yao, Y. (2021) Review of machine learning methods for RNA secondary structure prediction. *PLoS Comput Biol*, **17**, e1009291.
14. Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499-512.
15. Andronescu, M., Condon, A., Turner, D.H. and Mathews, D.H. (2014) In Gorodkin, J. and Ruzzo, W. L. (eds.), *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*. Humana Press, Totowa, NJ, pp. 45-70.
16. Rivas, E. (2013) The four ingredients of single-sequence RNA secondary structure prediction. A unifying perspective. *RNA Biol*, **10**, 1185-1196.
17. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133-148.
18. Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L. and Hendrix, D. (2018) bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res.*, **46**, 5381-5394.
19. Gutell, R.R., Lee, J.C. and Cannone, J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, **12**, 301-310.
20. Wang, Y., Liu, Y., Wang, S., Liu, Z., Gao, Y., Zhang, H. and Dong, L. (2020) ATTfold: RNA Secondary Structure Prediction With Pseudoknots Based on Attention Mechanism. *Frontiers in Genetics*, **11**.
21. Wang, L., Liu, Y., Zhong, X., Liu, H., Lu, C., Li, C. and Zhang, H. (2019) DMfold: A Novel Method to Predict RNA Secondary Structure With Pseudoknots Based on Deep Learning and Improved Base Pair Maximization Principle. *Front Genet*, **10**, 143.
22. Chen, X., Li, Y., Umarov, R., Gao, X. and Song, L. (2020) RNA Secondary Structure Prediction By Learning Unrolled Algorithms.
23. Sato, K., Akiyama, M. and Sakakibara, Y. (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun*, **12**, 941.
24. Singh, J., Hanson, J., Paliwal, K. and Zhou, Y. (2019) RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat Commun*, **10**, 5407.
25. Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q. and Xie, X. (2022) UFold: fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.*, **50**, e14.
26. Sloma, M.F. and Mathews, D.H. (2016) Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA*, **22**, 1808-1818.
27. Szikszai, M., Wise, M., Datta, A., Ward, M. and Mathews, D.H. (2022) Deep learning models for RNA secondary structure prediction (probably) do not generalise across families. *bioRxiv*, 2022.2003.2021.485135.

28. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152.
29. Tan, Z., Fu, Y., Sharma, G. and Mathews, D.H. (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.*, **45**, 11570-11581.
30. Mathews, D.H. (2019) How to benchmark RNA secondary structure prediction accuracy. *Methods*, **162-163**, 60-67.
31. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406-3415.
32. Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.
33. Reuter, J.S. and Mathews, D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
34. Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D.A. and Mathews, D.H. (2019) LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics*, **35**, i295-i304.
35. Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H. and Murphy, K.P. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, **23**, i19-28.
36. Do, C.B., Woods, D.A. and Batzoglou, S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90-98.
37. Zakov, S., Goldberg, Y., Elhadad, M. and Ziv-Ukelson, M. (2011) Rich parameterization improves RNA structure prediction. *J. Comput. Biol.*, **18**, 1525-1542.
38. Sato, K., Hamada, M., Asai, K. and Mituyama, T. (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic acids research*, **37**, W277-W280.
39. Sundfeld, D., Havgaard, J.H., de Melo, A.C.M.A. and Gorodkin, J. (2015) Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics*, **32**, 1238-1240.
40. Winkler, J., Urgese, G., Ficarra, E. and Reinert, K. (2022) LaRA 2: parallel and vectorized program for sequence–structure alignment of RNA sequences. *BMC Bioinformatics*, **23**, 18.
41. Will, S., Joshi, T., Hofacker, I.L., Stadler, P.F. and Backofen, R. (2012) LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, **18**, 900-914.
42. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
43. Nawrocki, E.P., Kolbe, D.L. and Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335-1337.
44. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. (2019) Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, **116**, 22071-22080.
45. Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y. and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103-107.

46. Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A. and Arkin, A.P. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences*, **108**, 11063-11068.
47. Delli Ponti, R., Marti, S., Armaos, A. and Tartaglia, Gian G. (2016) A high-throughput approach to profile RNA structure. *Nucleic Acids Research*, **45**, e35-e35.
48. Wayment-Steele, H.K., Kladwang, W., Strom, A.I., Lee, J., Treuille, A., Participants, E. and Das, R. (2021) RNA secondary structure packages evaluated and improved by high-throughput experiments. *bioRxiv*, 2020.2005.2029.124511.
49. Zhang, T., Singh, J., Litfin, T., Zhan, J., Paliwal, K. and Zhou, Y. (2021) RNAcmap: a fully automatic pipeline for predicting contact maps of RNAs by evolutionary coupling analysis. *Bioinformatics*, **37**, 3494-3500.
50. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
51. Ji, Y., Zhou, Z., Liu, H. and Davuluri, R.V. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, **37**, 2112-2120.
52. The, R.C., Petrov, A.I., Kay, S.J.E., Kalvari, I., Howe, K.L., Gray, K.A., Bruford, E.A., Kersey, P.J., Cochrane, G., Finn, R.D. *et al.* (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.*, **45**, D128-D134.
53. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.
54. Pearce, R., Omenn, G.S. and Zhang, Y. (2022) De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. *bioRxiv*, 2022.2005.2015.491755.
55. Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S. and Yang, L. (2021) Physics-informed machine learning. *Nature Reviews Physics*, **3**, 422-440.
56. Zhao, Y., Wang, J., Zeng, C. and Xiao, Y. (2018) Evaluation of RNA secondary structure prediction for both base-pairing and topology. *Biophysics Reports*, **4**, 123-132.