ScanFold 2.0: A rapid approach for identifying potential structured RNA targets in genomes and transcriptomes

Ryan J. Andrews [1], **Warren B. Rouse** [2], **Collin A. O'Leary** [2], **Nicholas J. Booher** [3], **Walter N Moss** Corresp. [2]

[1] Department of Biochemistry, University of Utah, Salt Lake City, UT, United States

[2] The Roy J Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa, United States

[3] Infrastructure and Research IT Services, Iowa State University, Ames, IA, United States

Corresponding Author: Walter N Moss
Email address: wmoss@iastate.edu

A major limiting factor in target discovery for both basic research and therapeutic intervention is the identification of structural and/or functional RNA elements in genomes and transcriptomes. This was the impetus for the original ScanFold algorithm, which provides maps of local RNA structural stability, evidence of sequence-ordered (potentially evolved) structure, and unique model structures comprised of recurring base pairs with the greatest structural bias. A key step in quantifying this propensity for ordered structure is the prediction of secondary structural stability for randomized sequences which, in the original implementation of ScanFold, is explicitly evaluated. This slow process has limited the rapid identification of ordered structures in large genomes/transcriptomes, which we seek to overcome in this current work introducing ScanFold 2.0. In this revised version of ScanFold, we no longer explicitly evaluate randomized sequence folding energy, but rather estimate it using a machine learning approach. This can increase prediction speeds for high randomization numbers by up to 140 times compared to ScanFold 1.0, allowing for the analysis of large sequences, as well as the use of additional folding algorithms that may be computationally expensive. In the testing of ScanFold 2.0, we re-evaluate the Zika, HIV, and SARS-CoV-2 genomes and compare both the consistency of results and the time of each run to ScanFold 1.0. We also re-evaluate the SARS-CoV-2 genome to assess the quality of ScanFold 2.0 predictions vs several biochemical structure probing datasets and compare the results to those of the original ScanFold program.

ScanFold 2.0: A rapid approach for identifying potential structured RNA targets in genomes and transcriptomes

Ryan J. Andrews[1,‡], Warren B. Rouse[2,‡], Collin A. O'Leary[2], Nicholas J. Booher[3], and Walter N. Moss[2,†]

[1]Department of Biochemistry, University of Utah, Salt Lake City, UT 84112, USA.
[2]Roy J. Carver Department of Biophysics, Biochemistry and Molecular Biology, Iowa State University, Ames, IA 50011, USA.
[3]Infrastructure and Research IT Services, Iowa State University, Ames, IA 50011, USA.

Corresponding Author:
Walter N. Moss[2]
Email address: wmoss@iastate.edu

Abstract

A major limiting factor in target discovery for both basic research and therapeutic intervention is the identification of structural and/or functional RNA elements in genomes and transcriptomes. This was the impetus for the original `ScanFold` algorithm, which provides maps of local RNA structural stability, evidence of sequence-ordered (potentially evolved) structure, and unique model structures comprised of recurring base pairs with the greatest structural bias. A key step in quantifying this propensity for ordered structure is the prediction of secondary structural stability for randomized sequences which, in the original implementation of `ScanFold`, is explicitly evaluated. This slow process has limited the rapid identification of ordered structures in large genomes/transcriptomes, which we seek to overcome in this current work introducing `ScanFold 2.0`. In this revised version of `ScanFold`, we no longer explicitly evaluate randomized sequence folding energy, but rather estimate it using a machine learning approach. This can increase prediction speeds for high randomization numbers by up to 140 times compared to `ScanFold 1.0`, allowing for the analysis of large sequences, as well as the use of additional folding algorithms that may be computationally expensive. In the testing of `ScanFold 2.0`, we re-evaluate the Zika, HIV, and SARS-CoV-2 genomes and compare both the consistency of results and the time of each run to `ScanFold 1.0`. We also re-evaluate the SARS-CoV-2 genome to assess the quality of `ScanFold 2.0` predictions vs several biochemical structure probing datasets and compare the results to those of the original `ScanFold` program.

Introduction

Interest in RNA has, arguably, never been higher. RNA plays key regulatory roles in all organisms including human pathogens such as HIV, Zika, and SARS-CoV-2 (Cao et al. 2021; Li et al. 2018; Watts et al. 2009). Furthermore, since both the viral vector and the most efficacious preventative modality for COVID-19 both consist of RNA, interest in RNA as both a therapeutic agent and target is surging (Bhat et al. 2021; Damase et al. 2021). Significantly, in both its biological function and potential for targeting, RNA secondary structure plays key and diverse roles (Andrzejewska et al. 2020; Disney 2019; Hargrove 2020; Meyer et al. 2020; Szabat et al. 2020; Wan et al. 2011). For example, in processes such as RNA splicing and posttranscriptional gene regulation, secondary structures can vary the distances between or accessibility of various regulatory elements in RNA (Andrzejewska et al. 2020; Jiang & Coller 2012; Li et al. 2014) as well as provide specific platforms for recognition by regulatory molecules (e.g., proteins and noncoding RNAs (Law et al. 2006; Sanchez de Groot et al. 2019; Yang et al. 2020)). Secondary structures are also found within long noncoding RNAs (Andrzejewska et al. 2020; Chillon & Marcia 2020; McCown et al. 2019; Somarowthu et al. 2015) and in the coding regions of mRNAs, where there is increasing awareness of their roles in modulating translation and protein folding (Andrzejewska et al. 2020; Faure et al. 2016; Faure et al. 2017; Mauger et al. 2019; Mustoe et al. 2018).

Unsurprisingly, there is great interest in gaining additional structure/function knowledge about RNA (particularly as related to human health) and in therapeutically modulating RNA biology via its secondary structure. Both tasks require the identification of robust structural models of RNA folding which, for large genomes/transcriptomes, is an immense challenge. Despite the

availability of rapid and robust algorithms for RNA secondary structure prediction (Lorenz et al. 2011; Reuter & Mathews 2010; Zuker 2003), novel methods for assessing the phylogenetic impact/significance of structure (Manfredonia et al. 2020; Rivas et al. 2017; Rivas et al. 2020), and tremendous advances in approaches for high-throughput probing of RNA secondary structure (Mitchell et al. 2019; Regulski & Breaker 2008; Smola & Weeks 2018; Strobel et al. 2018; Tomezsko et al. 2021); a major challenge that continues to hamper efforts to understand and target RNA secondary structure is the determination of which fragments form extremely stable, and likely functional structure.

Early on, it was noted that functional RNA structures have a sequence-ordered stability bias. That is to say, the predicted folding free energy of functional/evolved RNA is lower than that of randomized sequences (Clote et al. 2005; Moss 2018; Qu & Adelson 2012). This bias is quantified via the thermodynamic z-score, which measures the difference in predicted minimum free energy of folding for a native RNA vs. randomized sequence (with the same nucleotide and/or dinucleotide content) and normalizing by the standard deviation. Thus, the z-score indicates the number of standard deviations more or less stable the native secondary structure is vs. that predicted by nucleotide content (i.e., negative values indicate significantly ordered stability) (Andrews et al. 2017; Clote et al. 2005).

`ScanFold 2.0` (SF2) uses the same approaches as `ScanFold 1.0` (SF1) without the need for explicit MFE calculations of randomized sequences to determine thermodynamic z-scores. To bypass the computationally expensive explicit z-score calculations, we have implemented a machine learning approach: Google's publicly available TensorFlow algorithm (Abadi et al. 2016a; Abadi et al. 2016b). TensorFlow was trained using 20 different sequence features including: sequence length, GC percentage, CG ratio, AU ratio, and the frequency of 16 different dinucleotide types. Using these features, both mono- and dinucleotide shuffling models were generated. SF2 uses these models to estimate the randomized MFEs and standard deviations needed to calculate thermodynamic z-scores for all windows. This new version of `ScanFold` still uses the same algorithm to highlight local structural features, `ScanFold-Fold` (Andrews et al. 2020; Andrews et al. 2018), which is now the rate limiting step of the program. This improvement has led to an increase in computational speeds of at least 10x, and in some cases increases of over 100x (**File S1**). This new tool is available for download on GitHub (https://github.com/moss-lab/ScanFold2.0) or through a webserver hosted at: https://mosslabtools.bb.iastate.edu/scanfold2.

Methods

TensorFlow training of z-score model

An overview of the training process can be seen in **Figure 1**. A total of 836,377 representative sequences were generated to be used for training. Sequence lengths were between 60 and 200 nt (based on typical `ScanFold` window sizes (Andrews et al. 2020; Andrews et al. 2018)) in 20 nt increments. To represent as many potential sequence types as possible, dinucleotide frequencies for all 16 dinucleotide types were set to vary between 0 and 45%, averaging ~6.3% across all sequences. Native MFEs, mean of 100 randomized MFEs ($\overline{MFE}$), and their standard deviations (σ) were calculated for all sequences using `RNAfold` version 2.4.18 (Lorenz et al. 2011). Two different randomization procedures were used to train the algorithm: mononucleotide and

dinucleotide shuffling (Andrews et al. 2020; Andrews et al. 2018; Gesell & Washietl 2008). Twenty different training features were also collected for each sequence including: sequence length, GC percent, AU ratio, GC ratio, and all 16 dinucleotide frequencies.

All 20 features were used during training of $\overline{MFE}$ and standard deviation (σ) models. The mean MFE and STD models are Keras sequential, with one preprocessing normalization layer, and two hidden layers: Rectified Linear Unit (ReLu) and sigmoid. `RNAfold` is used to calculate MFEs, while $\overline{MFE}$ and standard deviation (σ) models are invoked separately for z-score calculation **(Eq. 1)**. All training code was run through Google Colab (Bisong 2019) and can be viewed and run directly in the corresponding python notebook (**File S2**).

$$z - score = \frac{MFE - \overline{MFE}}{\sigma} \#(1)$$

Updates to ScanFold 2.0 and integration in the webserver

To make the use of SF2 more user friendly, it has been incorporated into the Moss Lab Tools webserver (https://mosslabtools.bb.iastate.edu/scanfold2). Similar to SF1, any sequence longer than the chosen window size can be uploaded (or pasted) in FASTA format, all parameters can be set by the user, and the scan can be started by clicking the submit button at the bottom of the page. Once the prediction is complete the results are output in an `Integrative Genomics Viewer` (IGV.js) window (Robinson et al. 2020) and made available for download as a zip file.

Testing of ScanFold 2.0 vs ScanFold

SF2 was tested to determine its accuracy and speed compared to that of SF1. Testing was performed on HIV-1, Zika, and SARS-CoV-2 genomes, which had been previously analyzed using SF1 (Andrews et al. 2020; Andrews et al. 2021; Andrews et al. 2018). To ensure that our testing was comprehensive we compared SF2 mono- and dinucleotide shuffling results to those of SF1 mono- and dinucleotide shuffling using 100, 1000, and 10000 randomizations for each genome. The results of all output CT files (i.e. -2, -1, and No Filter z-scores) from both versions of ScanFold were compared using an in-house python script, ct_compare.py ((Andrews et al. 2021); https://github.com/moss-lab/SARS-CoV-2). This comparison allowed us to evaluate the percent of paired nucleotides and the percent similarity or consistency between the output files of both versions of `ScanFold` as well as determine the improvements in speed for each run. Additionally, we were able to compare the outputs from SF1 (mono- vs dinucleotide shuffling and different number of randomizations) and the outputs of SF2 (mono- vs dinucleotide shuffling) to themselves to evaluate their performance using different shuffling methods. In total, 13 different comparisons were completed for each genome. All accuracy and speed results can be found in **File S1**.

ROC Analysis

ROC analysis was performed on `ScanFold-Fold` results for SF1 mono- and dinucleotide shuffling using 100 and 10000 randomizations as well as SF2 mono- and dinucleotide shuffling models following a previously establish protocol (Andrews et al. 2021). Briefly, reactivity value thresholds were sequentially set from the lowest to highest value at 1% intervals (i.e. 0-100% constrained) for various SHAPE and DMS reactivity datasets generated from SARS-CoV-2 probing experiments (Huston et al. 2021; Lan et al. 2021; Manfredonia et al. 2020; Sun et al.

2021). The -1 z-score CT files from SF1 and SF2 were then cross referenced to these reactivity datasets and used to find the true positive rate (TPR) and false positive rates (FPR) for each comparison. In this analysis, the TPR and FPR are represented by equations 2 and 3 below:

$$TPR = \frac{TP}{(TP + FN)} \#(2)$$

$$FPR = \frac{FP}{(FP + TN)} \#(3)$$

The true positive (TP) is defined as being *paired* in the given CT file and *paired* at the defined reactivity threshold, the false negative (FN) is *paired* in the CT file and *unpaired* at the reactivity threshold. The false positive (FP) is *unpaired* in the CT file and *paired* at the reactivity threshold, and the true negative (TN) is *unpaired* in the CT file and *unpaired* at the given reactivity threshold. When the threshold is set to 0%, TPR and FPR will be equal to zero, and when the reactivity threshold is set to 100%, TPR and FPR will be equal to one. If a given RNA secondary structure model is truly random, when compared to increasing reactivity thresholds from a probing data set, then the TPR and FPR should increase proportionately yielding a linear trend in the plot. However, if the RNA secondary structure model agrees with the reactivity data set, the TPR should initially rise faster than the FPR, creating a larger area under the curve (AUC) and producing a curve on the plot. In this way, we can quantitatively assess and compare each model's ability to fit the data via their respective AUCs. All the ROC and AUC analysis can be found in **File S3**.

Results and Discussion

Comparing time and accuracy of ScanFold 2.0 vs ScanFold 1.0

SF2 requires significantly less time than SF1 using only 100 randomizations, with increases in speed being even greater when compared to SF1 using 1000 and 10000 explicitly shuffled RNA sequences for z-score calculations. In both cases, increasing sequence length does increase the time needed, but this effect is seen to a lesser degree in SF2. When comparing the times, SF1 using 100 randomizations with mononucleotide shuffling takes 8.70 hrs, 1.02 hrs, and 1.75 hrs to complete all predictions for SARS, HIV, and Zika respectively (**Table 1**). SF2 on the other hand reduces these times to 2.64 hrs, 0.27 hrs, and 0.35 hrs for SARS, HIV, and Zika respectively (**Table 2**). This decrease in time for SF2 is greater for higher randomization numbers and dinucleotide shuffling (**Table 1** and **Table 2**). For SF2, the scanning step is now the fastest step in the process, taking only 0.27 hrs, 0.07 hrs, and 0.09 hrs for SARS, HIV, and Zika respectively (**Table 2**). Importantly, increased speed does not come at the cost of reduced accuracy.

Gross comparisons of the percent of predicted pairs by SF1 and SF2 using 100, 1000, and 10000 randomizations with mononucleotide shuffling displays an average difference of 2.00% (0.03% to 4.5%) between all z-score cutoffs across the three genomes analyzed, regardless of the number of randomizations. HIV-1 is the most consistent between versions, displaying less than a 1.25% difference in -2 z-score pairs, 3.2% difference in -1 z-score pairs, and 0.5% difference in all pairs (no filter) across all randomizations (**File S1**). In a similar analysis, it is also seen that the percent similarity or consistency of paired and unpaired nucleotides between SF1 and SF2 using

mononucleotide shuffling is quite high, with the average difference being only 4.01% (1.11% to 6.29%) between all z-score cutoffs across the three genomes analyzed (**File S1**). Here, HIV-1 shows some of the best results with only the no filter cutoff reaching a 6.24% difference, and z-score cutoffs of -2 and -1 being only 1.42% and 4.7% different, respectively (**Fig. 2**).

The same analyses were carried out between SF1 and SF2 using dinucleotide shuffling. Comparing the percent of predicted paired nucleotides using 100, 1000, and 10000 randomizations with dinucleotide shuffling displays an average difference of 5.26% (0.57% to 10.26%) between all z-score cutoffs across the three genomes analyzed. HIV-1 showed the least variance with a 4.38% difference in -2 z-score pairs, an 8.72% difference in -1 z-score pairs, and a 1.85% difference in all (no filter) pairs across all randomizations (**File S1**). The percent similarity or consistency in the paired and unpaired nucleotides between SF1 and SF2 using dinucleotide shuffling is again quite high, especially for structures within the significant z-score cutoffs of -2 and -1, with the average difference being 10.42% (4.71% to 20.64%) between all z-score cutoffs across the three genomes analyzed (**File S1**). Here, HIV-1 shows some of the best results with only the no filter cutoff reaching a 20.64% difference, and z-score cutoffs of -2 and -1 being only 4.82% and 10.16% different respectively (**Fig. 2**). Notably, when comparing the predictions to biochemical probing data all approaches showed consistency with experimental results (**Fig. 3**).

Mono vs Di nucleotide shuffling of ScanFold 2.0

When comparing SF1 and SF2 results for mononucleotide shuffling there is an average difference in percent paired of 2.00% (0.03% to 4.5%) and in the majority of cases SF2 is predicting more pairs than SF1. For all results other than HIV and SARS all pairs (no filter), SF2 consistently predicts more pairs than SF1. When comparing SF1 and SF2 results for dinucleotide shuffling, there is an average difference of 5.26% (0.57% to 10.26%) and similar to mononucleotide shuffling, all results other than Zika no filter (all pairs), show that SF2 is always predicting slightly more pairs. These small differences serve as evidence that SF1 and SF2 are producing almost an identical number of pairs when the same shuffling method is used (**File S1**).

When comparing the results of SF1 *mono*nucleotide shuffling to SF1 *di*nucleotide shuffling, on average mononucleotide shuffling finds more pairs than dinucleotide shuffling, but this does not always hold true—as is the case with all iterations of Zika results for all pairs (no filter; **Fig. S1**). The smallest difference in SF1 results is seen in Zika all pairs where dinucleotide shuffling finds 0.72% more pairs than mononucleotide, and the largest difference is seen in Zika -1 z-score pairs where mononucleotide shuffling predicts 8.65% more pairs than dinucleotide (**Table S1** and **Fig. S1**). SF2 comparisons show a split between which shuffling method predicts more pairs. In the case of Zika, the same trend seen for SF1 holds true for SF2, with mononucleotide shuffling finding more pairs than dinucleotide shuffling for all cutoffs other than all pairs. For HIV, SF2 dinucleotide shuffling finds more pairs than mononucleotide shuffling at all z-score cutoffs, but for SARS dinucleotide shuffling finds more pairs than mononucleotide shuffling only at the -2 z-score cutoff. Here, the smallest difference in SF2 is seen in the SARS results for all pairs where mononucleotide shuffling finds 0.36% more pairs than dinucleotide, and the largest difference is seen in Zika results for -1 z-score pairs where mononucleotide shuffling finds 3.13% more pairs than dinucleotide (**Table S1** and **Fig. S1**). These small variations between the shuffling methods provide further evidence that SF1 and SF2 are performing similarly in identifying ordered structure, and that the shuffling technique used does not influence the results to a high degree.

As additional evidence that the shuffling method does not have a large impact on results, we analyzed the percent consistency in pairing between SF1 and SF2 using 100, 1000, and 10000 randomizations with both shuffling methods. Here, we observe that SF2 mono- and dinucleotide results are generally consistent (within 5-10%) with that of SF1 mono- and dinucleotide results across all three genomes, with HIV demonstrating the most consistency (**Figure 2**). The general trend among the three genomes shows the more stringent -2 and -1 z-score predictions are always within 10-12% consistency of each other regardless of shuffling or randomization, while the no filter pairings often show more variation (**File S1**). All comparisons seem to show no significant benefit of using dinucleotide over mononucleotide shuffling as the percent consistency between these methods in both SF1 and SF2 predictions are on average 7.53% different (1.85% to 18.27%) and when looking at just SF2 using both methods, predictions are on average 4.79% different (1.96% to 9%) (**File S1**). The differences associated with SF1 and SF2 mononucleotide and dinucleotide shuffling can most likely be equated to the differences in z-scores (**Figure S2** and **File S4**). The box and whisker plot in **Figure S2** shows that for SF2, the average z-scores are consistently lower for both shuffling methods compared to that of SF1, and the differences in z-scores between the two shuffling methods is also much smaller for SF2 (average difference of -0.019) compared to that of SF1 (average difference of -0.363) (**Table S2** and **File S4**). The lower overall z-score of SF2 is potentially causing the differences in percent paired and percent similarity or consistency that is seen between the shuffling methods when comparing SF1 and SF2. Regardless of the differences in percent paired, percent similarity or consistency, and z-score the results of mononucleotide and dinucleotide shuffling for SF2 are similar to SF1 as shown by the agreement of biochemical probing data (**Fig. 3**).

ROC Analysis of SARS-CoV-2

As another layer of validation, we followed an established protocol (Andrews et al. 2021) to perform a receiver operator characteristic (ROC) analysis on the SARS genome predictions. We compared SF1 and SF2 results using 100-10000 randomizations with both shuffling methods to six different SHAPE and DMS biochemical probing datasets (Huston et al. 2021; Lan et al. 2021; Manfredonia et al. 2020; Sun et al. 2021). Here, the effect of increasing the stringency of reactivity cutoffs, which considers whether a site is to be paired in the model, provides a measure of the consistency of probing data compared to `ScanFold` models (see Material and Methods and (Andrews et al. 2021)). We initially compared the SF1 results using both shuffling methods with 100 and 10000 randomizations and the SF2 results using both shuffling methods to the *Lan et al.* in vitro DMS data. The ROC analysis showed that all SF1 and SF2 results clustered into the same curve with almost identical area under the curve (AUC) values (**Figure 3A**). The ROC analysis of SF1 and SF2 results using 100 randomizations and both shuffling methods was repeated on all six probing datasets. SF2 predictions match the curves of both the previous analysis and all SF1 results obtained in this study (**Figure 3B**). After calculating the area under the curve (AUC) for each set of results, all were found to be above 0.5, indicating global consistency of the data with SF1 and SF2 results. AUC values for SF2 ranged from a minimum value of 0.629 for comparison of SF2 dinucleotide to *in vivo* SHAPE dataset (Huston *et al.*) to a maximum value of 0.780 for comparison of SF2 mononucleotide to *in vivo* DMS dataset (Lan *et al.*). No large differences were observed when comparing any of the AUC values between SF1 or SF2 and the respective datasets. These findings indicate that, similar to SF1, SF2 is detecting the most robust local elements that do not vary between experimental conditions.

Conclusion

SF2 produces effectively indistinguishable results to that of SF1 in a fraction of the time. Based on our results, we see that SF2 using the dinucleotide shuffling model tends to produce results more similar to mononucleotide than SF1; however, both SF1 and SF2 results are generally similar to each other. ROC analysis using several SHAPE and DMS datasets against SF1 and SF2 predictions also suggests that, regardless of the model, SF2 detects robust structural elements that persist between experimental conditions. Here, we have demonstrated that the improved SF2 algorithm performs similarly to, but in a fraction of the time as SF1. We hope that this improved speed can provide the RNA community with a fast, accurate, and user-friendly tool that will help in finding potentially functional structures across any gene or transcript of interest and drive forward RNA research.

Acknowledgements

Thank you to the Iowa State University Research IT group for their support over the course of this project and members of the Moss Lab for their input.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, and Zheng X. 2016a. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. p arXiv:1603.04467.

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, and Zheng X. 2016b. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. Savannah, GA, USA: USENIX Association. p 265–283.

Andrews RJ, Baber L, and Moss WN. 2017. RNAStructuromeDB: A genome-wide database for RNA structural inference. *Sci Rep* 7:17269. 10.1038/s41598-017-17510-y

Andrews RJ, Baber L, and Moss WN. 2020. Mapping the RNA structural landscape of viral genomes. *Methods* 183:57-67. 10.1016/j.ymeth.2019.11.001

Andrews RJ, O'Leary CA, Tompkins VS, Peterson JM, Haniff HS, Williams C, Disney MD, and Moss WN. 2021. A map of the SARS-CoV-2 RNA structurome. *NAR Genom Bioinform* 3:lqab043. 10.1093/nargab/lqab043

Andrews RJ, Roche J, and Moss WN. 2018. ScanFold: an approach for genome-wide discovery of local RNA structural elements-applications to Zika virus and HIV. *PeerJ* 6:e6136. 10.7717/peerj.6136

Andrzejewska A, Zawadzka M, and Pachulska-Wieczorek K. 2020. On the Way to Understanding the Interplay between the RNA Structure and Functions in Cells: A Genome-Wide Perspective. *Int J Mol Sci* 21. 10.3390/ijms21186770

Bhat B, Karve S, and Anderson DG. 2021. mRNA therapeutics: beyond vaccine applications. *Trends Mol Med* 27:923-924. 10.1016/j.molmed.2021.05.004

Bisong E. 2019. Google Colaboratory. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley, CA: Apress, 59-64.

Cao C, Cai Z, Xiao X, Rao J, Chen J, Hu N, Yang M, Xing X, Wang Y, Li M, Zhou B, Wang X, Wang J, and Xue Y. 2021. The architecture of the SARS-CoV-2 RNA genome inside virion. *Nat Commun* 12:3917. 10.1038/s41467-021-22785-x

Chillon I, and Marcia M. 2020. The molecular structure of long non-coding RNAs: emerging patterns and functional implications. *Crit Rev Biochem Mol Biol* 55:662-690. 10.1080/10409238.2020.1828259

Clote P, Ferre F, Kranakis E, and Krizanc D. 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* 11:578-591. 10.1261/rna.7220505

Damase TR, Sukhovershin R, Boada C, Taraballi F, Pettigrew RI, and Cooke JP. 2021. The Limitless Future of RNA Therapeutics. *Front Bioeng Biotechnol* 9:628137. 10.3389/fbioe.2021.628137

Disney MD. 2019. Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine. *J Am Chem Soc* 141:6776-6790. 10.1021/jacs.8b13419

Faure G, Ogurtsov AY, Shabalina SA, and Koonin EV. 2016. Role of mRNA structure in the control of protein folding. *Nucleic Acids Res* 44:10898-10911. 10.1093/nar/gkw671

Faure G, Ogurtsov AY, Shabalina SA, and Koonin EV. 2017. Adaptation of mRNA structure to control protein folding. *RNA Biol* 14:1649-1654. 10.1080/15476286.2017.1349047

Gesell T, and Washietl S. 2008. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* 9:248. 10.1186/1471-2105-9-248

Hargrove AE. 2020. Small molecule-RNA targeting: starting with the fundamentals. *Chem Commun (Camb)* 56:14744-14756. 10.1039/d0cc06796b

Huston NC, Wan H, Strine MS, de Cesaris Araujo Tavares R, Wilen CB, and Pyle AM. 2021. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol Cell* 81:584-598 e585. 10.1016/j.molcel.2020.12.041

Jiang P, and Coller H. 2012. Functional interactions between microRNAs and RNA binding proteins. *Microrna* 1:70-79. 10.2174/2211536611201010070

Lan TCT, Allan MF, Malsick LE, Khandwala S, Nyeo SSY, Sun Y, Guo JU, Bathe M, Griffiths A, and Rouskin S. 2021. Insights into the secondary structural ensembles of the full SARS-CoV-2 RNA genome in infected cells. *bioRxiv*:2020.2006.2029.178343. 10.1101/2020.06.29.178343

Law MJ, Rice AJ, Lin P, and Laird-Offringa IA. 2006. The role of RNA structure in the interaction of U1A protein with U1 hairpin II RNA. *RNA* 12:1168-1178. 10.1261/rna.75206

Li P, Wei Y, Mei M, Tang L, Sun L, Huang W, Zhou J, Zou C, Zhang S, Qin CF, Jiang T, Dai J, Tan X, and Zhang QC. 2018. Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity. *Cell Host Microbe* 24:875-886 e875. 10.1016/j.chom.2018.10.011

Li X, Kazan H, Lipshitz HD, and Morris QD. 2014. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA* 5:111-130. 10.1002/wrna.1201

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, and Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* 6:26. 10.1186/1748-7188-6-26

Manfredonia I, Nithin C, Ponce-Salvatierra A, Ghosh P, Wirecki TK, Marinus T, Ogando NS, Snijder EJ, van Hemert MJ, Bujnicki JM, and Incarnato D. 2020. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res* 48:12436-12452. 10.1093/nar/gkaa1053

Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, Link K, Khatwani N, Reynders J, Moore MJ, and McFadyen IJ. 2019. mRNA structure regulates protein expression through changes in functional half-life. *Proc Natl Acad Sci U S A* 116:24075-24083. 10.1073/pnas.1908052116

McCown PJ, Wang MC, Jaeger L, and Brown JA. 2019. Secondary Structural Model of Human MALAT1 Reveals Multiple Structure-Function Relationships. *Int J Mol Sci* 20. 10.3390/ijms20225610

Meyer SM, Williams CC, Akahori Y, Tanaka T, Aikawa H, Tong Y, Childs-Disney JL, and Disney MD. 2020. Small molecule recognition of disease-relevant RNA structures. *Chem Soc Rev* 49:7167-7199. 10.1039/d0cs00560f

Mitchell D, 3rd, Assmann SM, and Bevilacqua PC. 2019. Probing RNA structure in vivo. *Curr Opin Struct Biol* 59:151-158. 10.1016/j.sbi.2019.07.008

Moss WN. 2018. The ensemble diversity of non-coding RNA structure is lower than random sequence. *Noncoding RNA Res* 3:100-107. 10.1016/j.ncrna.2018.04.005

Mustoe AM, Corley M, Laederach A, and Weeks KM. 2018. Messenger RNA Structure Regulates Translation Initiation: A Mechanism Exploited from Bacteria to Humans. *Biochemistry* 57:3537-3539. 10.1021/acs.biochem.8b00395

Qu Z, and Adelson DL. 2012. Evolutionary conservation and functional roles of ncRNA. *Front Genet* 3:205. 10.3389/fgene.2012.00205

Regulski EE, and Breaker RR. 2008. In-line probing analysis of riboswitches. *Methods Mol Biol* 419:53-67. 10.1007/978-1-59745-033-1_4

Reuter JS, and Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11:129. 10.1186/1471-2105-11-129

Rivas E, Clements J, and Eddy SR. 2017. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods* 14:45-48. 10.1038/nmeth.4066

Rivas E, Clements J, and Eddy SR. 2020. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics* 36:3072-3076. 10.1093/bioinformatics/btaa080

Robinson JT, Thorvaldsdóttir H, Turner D, and Mesirov JP. 2020. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *bioRxiv*:2020.2005.2003.075499. 10.1101/2020.05.03.075499

Sanchez de Groot N, Armaos A, Grana-Montes R, Alriquet M, Calloni G, Vabulas RM, and Tartaglia GG. 2019. RNA structure drives interaction with proteins. *Nat Commun* 10:3246. 10.1038/s41467-019-10923-5

Smola MJ, and Weeks KM. 2018. In-cell RNA structure probing with SHAPE-MaP. *Nat Protoc* 13:1181-1195. 10.1038/nprot.2018.010

Somarowthu S, Legiewicz M, Chillon I, Marcia M, Liu F, and Pyle AM. 2015. HOTAIR forms an intricate and modular secondary structure. *Mol Cell* 58:353-361. 10.1016/j.molcel.2015.03.006

Strobel EJ, Yu AM, and Lucks JB. 2018. High-throughput determination of RNA structures. *Nat Rev Genet* 19:615-634. 10.1038/s41576-018-0034-x

Sun L, Li P, Ju X, Rao J, Huang W, Ren L, Zhang S, Xiong T, Xu K, Zhou X, Gong M, Miska E, Ding Q, Wang J, and Zhang QC. 2021. In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs. *Cell* 184:1865-1883 e1820. 10.1016/j.cell.2021.02.008

Szabat M, Lorent D, Czapik T, Tomaszewska M, Kierzek E, and Kierzek R. 2020. RNA Secondary Structure as a First Step for Rational Design of the Oligonucleotides towards Inhibition of Influenza A Virus Replication. *Pathogens* 9. 10.3390/pathogens9110925

Tomezsko P, Swaminathan H, and Rouskin S. 2021. DMS-MaPseq for Genome-Wide or Targeted RNA Structure Probing In Vitro and In Vivo. *Methods Mol Biol* 2254:219-238. 10.1007/978-1-0716-1158-6_13

Wan Y, Kertesz M, Spitale RC, Segal E, and Chang HY. 2011. Understanding the transcriptome through RNA structure. *Nat Rev Genet* 12:641-655. 10.1038/nrg3049

Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr., Swanstrom R, Burch CL, and Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460:711-716. 10.1038/nature08237

406 Yang M, Woolfenden HC, Zhang Y, Fang X, Liu Q, Vigh ML, Cheema J, Yang X, Norris M, Yu S, Carbonell A,
407 Brodersen P, Wang J, and Ding Y. 2020. Intact RNA structurome reveals mRNA structure-
408 mediated regulation of miRNA cleavage in vivo. *Nucleic Acids Res* 48:8767-8781.
409 10.1093/nar/gkaa577
410 Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*
411 31:3406-3415. 10.1093/nar/gkg595

412

Figure 1

Schematic of ScanFold 2.0 training procedure.

Representative sequences were generated for a range of lengths (between 60 and 200 nt) and dinucleotide frequencies. These sequences were shuffled and analyzed using RNAfold to determine their MFEs, mean MFEs and respective standard deviations. Mean MFEs and standard deviations were then combined with 18 sequence composition features to comprise all 20 training features. These 20 features were used to generate mean MFE and standard deviation models.

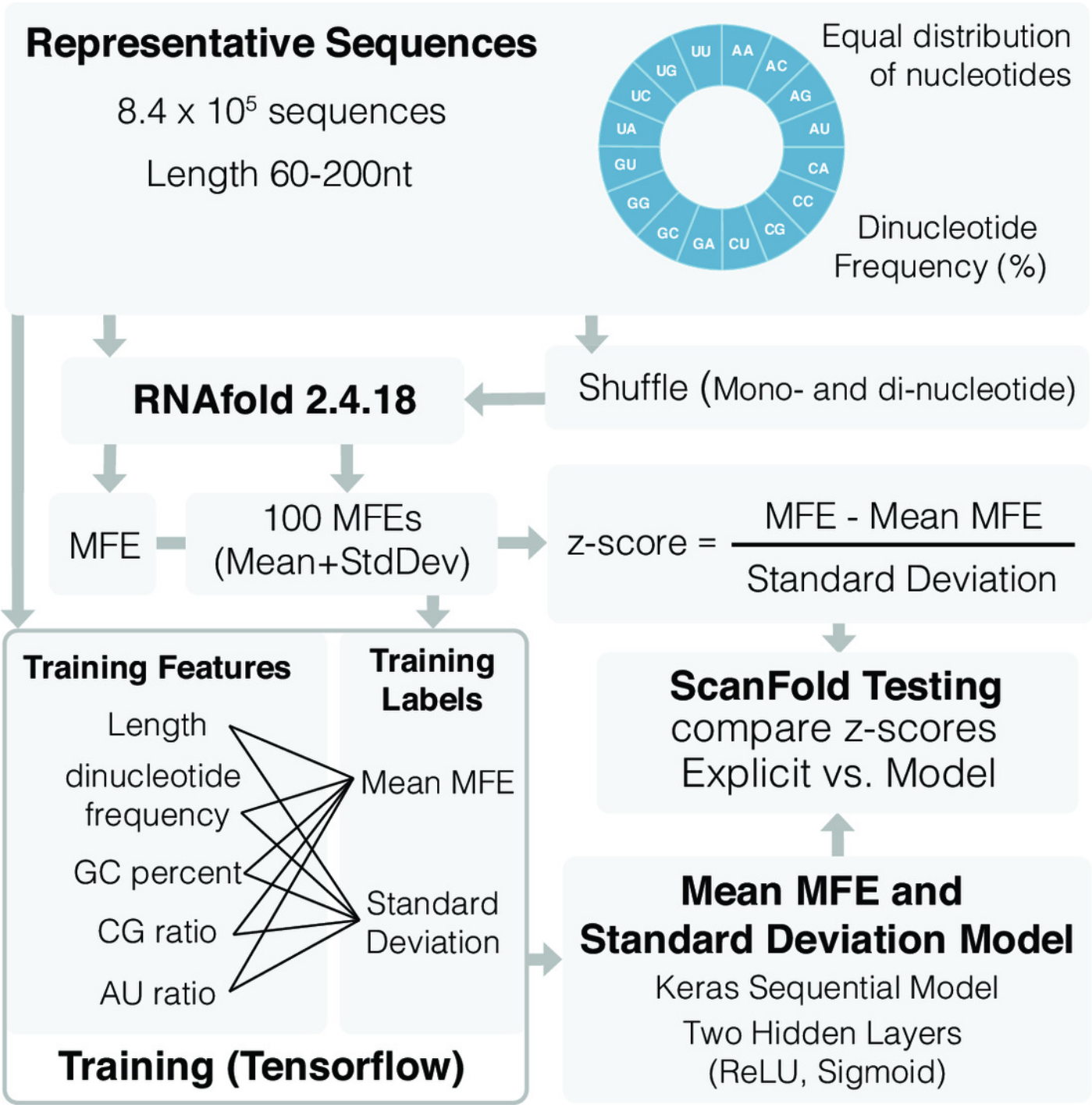
Representative Sequences
8.4 x 10^5 sequences
Length 60-200nt
Equal distribution of nucleotides
Dinucleotide Frequency (%)
AA
AC
AG
AU
CA
CC
CG
CU
GA
GC
GG
GU
UA
UC
UG
UU
RNAfold 2.4.18
Shuffle (Mono- and di-nucleotide)
MFE
100 MFEs (Mean+StdDev)
z-score = MFE - Mean MFE / Standard Deviation
Training Features
Length
dinucleotide frequency
GC percent
CG ratio
AU ratio
Training Labels
Mean MFE
Standard Deviation
Training (Tensorflow)
ScanFold Testing
compare z-scores
Explicit vs. Model
Mean MFE and Standard Deviation Model
Keras Sequential Model
Two Hidden Layers
(ReLU, Sigmoid)

Figure 2

SF1 and SF2 comparisons of HIV results.

Comparison of SF1 and SF2 percent similarity in paired and unpaired nucleotides using mono and dinucleotide shuffling with 100, 1000, and 10000 randomizations. A) HIV percent similarity in -2 z-score results. B) HIV percent similarity in -1 z-score results. All comparison were done using SF1 results as the reference and SF2 results as the target for comparison.

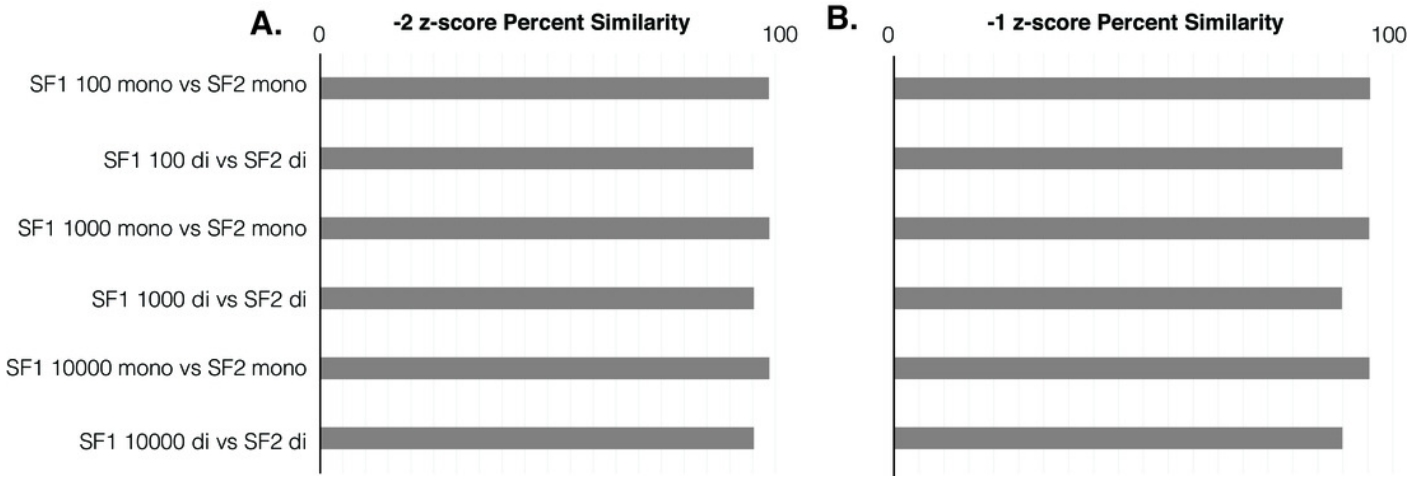

Figure 3

ROC analysis of SF1 and SF2 results.

ROC analysis of six different in vivo and in vitro SHAPE and DMS biochemical probing dataset of the SARS-CoV-2 genome. A) Plot of the initial ROC analysis curve with the AUC for SF1 using mono and dinucleotide shuffling at 100 and 10000 randomization and SF2 results using mono and dinucleotide shuffling for *Lan et al.* DMS in vivo dataset. SF1 mononucleotide with 100 randomizations in blue (AUC = 0.776), SF1 mononucleotide with 10000 randomizations in orange (AUC = 0.773), SF1 dinucleotide with 100 randomizations in gray (AUC = 0.759), SF1 dinucleotide with 10000 randomizations in yellow (AUC = 0.758), SF2 mononucleotide in black (AUC = 0.780), and SF1 dinucleotide in green (AUC = 0.773). B) Plot of the ROC analysis with the AUC for SF1 using mono and dinucleotide shuffling at 100 randomizations and SF2 results using mono and dinucleotide shuffling for all probing datasets. All SF1 and SF2 results for *Lan et al.* DMS in vivo in blue (AUC = 0.759 – 0.780), *Manfredonia et al.* DMS in vitro in yellow (AUC = 0.722 – 0.741), *Sun et al.* SHAPE in vivo in green (AUC = 0.725 – 0.748), *Manfredonia et al.* SHAPE in vitro in orange (AUC = 0.677 – 0.691), *Manfredonia et al.* SHAPE in vivo in gray (AUC = 0.660 – 0.678), and *Huston et al.* SHAPE in vivo in black (AUC = 0.622 – 0.633).

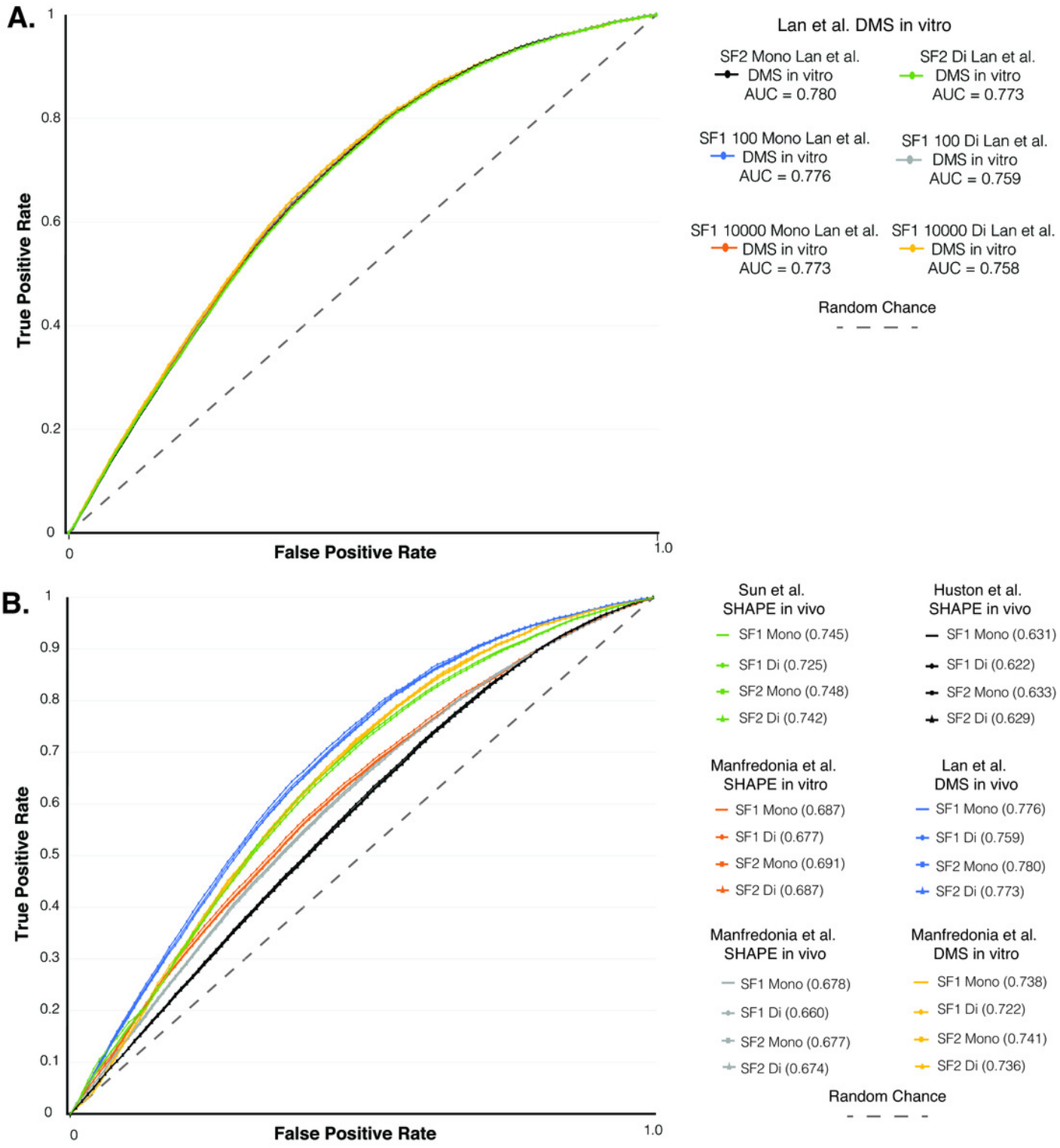
A.
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
0
1.0
Lan et al. DMS in vitro
SF2 Mono Lan et al. DMS in vitro AUC = 0.780
SF2 Di Lan et al. DMS in vitro AUC = 0.773
SF1 100 Mono Lan et al. DMS in vitro AUC = 0.776
SF1 100 Di Lan et al. DMS in vitro AUC = 0.759
SF1 10000 Mono Lan et al. DMS in vitro AUC = 0.773
SF1 10000 Di Lan et al. DMS in vitro AUC = 0.758
Random Chance
B.
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1.0
Sun et al. SHAPE in vivo
SF1 Mono (0.745)
SF1 Di (0.725)
SF2 Mono (0.748)
SF2 Di (0.742)
Huston et al. SHAPE in vivo
SF1 Mono (0.631)
SF1 Di (0.622)
SF2 Mono (0.633)
SF2 Di (0.629)
Manfredonia et al. SHAPE in vitro
SF1 Mono (0.687)
SF1 Di (0.677)
SF2 Mono (0.691)
SF2 Di (0.687)
Lan et al. DMS in vivo
SF1 Mono (0.776)
SF1 Di (0.759)
SF2 Mono (0.780)
SF2 Di (0.773)
Manfredonia et al. SHAPE in vivo
SF1 Mono (0.678)
SF1 Di (0.660)
SF2 Mono (0.677)
SF2 Di (0.674)
Manfredonia et al. DMS in vitro
SF1 Mono (0.738)
SF1 Di (0.722)
SF2 Mono (0.741)
SF2 Di (0.736)
Random Chance

Table 1(on next page)

Time required for SF1 runs using different shuffling methods and number of randomizations to finish.

The time required to finish runs for both versions of ScanFold were evaluated using different shuffling methods and number of randomizations. All times are reported in hours.

1

2

	Total Time 100 Rnds (hrs)	Total Time 1000 Rnds (hrs)	Total Time 10000 Rnds (hrs)
SARS SF1 Mono	8.70	21.28	164.17
HIV SF1 Mono	1.02	4.58	32.85
ZIKA SF1 Mono	1.75	4.15	36.55
SARS SF1 Di	7.50	22.07	134.00
HIV SF1 Di	0.95	4.48	35.58
ZIKA SF1 Di	1.25	4.67	38.53

Table 2(on next page)

Time required for each step of SF2 to run, total SF2 run time, and increase in SF2 speeds compared to SF1.

The time required to finish SF2 scanning step, folding step, and both steps were evaluated using different shuffling methods. Increase in speed was calculated by dividing SF1 total run time for each shuffling technique at each number of randomizations by SF2 total run time. All times are reported in hours.

1

2

3

	Scan Time	Fold Time	Total Time	Speed Increase 100 Rnds	Speed Increase 1000 Rnds	Speed Increase 10000 Rnds
SARS SF2 Mono	0.27	2.37	2.64	3.30x	8.06x	62.19x
HIV-1 SF2 Mono	0.07	0.20	0.27	3.78x	16.96x	121.67x
ZIKA SF2 Mono	0.09	0.27	0.35	5.00x	11.86x	104.43x
SARS SF2 Di	0.33	1.67	2.00	3.75x	11.04x	67.00x
HIV SF2 Di	0.07	0.17	0.24	3.96x	18.67x	148.25x
ZIKA SF2 Di	0.09	0.23	0.32	3.91x	14.59x	120.41x