Published online 24 August 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 **1**
https://doi.org/10.1093/nargab/lqab073

Rapid structure-function insights via hairpin-centric analysis of big RNA structure probing datasets

Pierce Radecki, Rahul Uppuluri and Sharon Aviran*

Biomedical Engineering Department and Genome Center, University of California at Davis, Davis, CA 95616, USA

Received February 16, 2021; Revised July 14, 2021; Editorial Decision July 26, 2021; Accepted August 03, 2021

ABSTRACT

The functions of RNA are often tied to its structure, hence analyzing structure is of significant interest when studying cellular processes. Recently, large-scale structure probing (SP) studies have enabled assessment of global structure-function relationships via standard data summarizations or local folding. Here, we approach structure quantification from a hairpin-centric perspective where putative hairpins are identified in SP datasets and used as a means to capture local structural effects. This has the advantage of rapid processing of big (e.g. transcriptome-wide) data as RNA folding is circumvented, yet it captures more information than simple data summarizations. We reformulate a statistical learning algorithm we previously developed to significantly improve precision of hairpin detection, then introduce a novel nucleotide-wise measure, termed the hairpin-derived structure level (HDSL), which captures local structuredness by accounting for the presence of likely hairpin elements. Applying HDSL to data from recent studies recapitulates, strengthens and expands on their findings which were obtained by more comprehensive folding algorithms, yet our analyses are orders of magnitude faster. These results demonstrate that hairpin detection is a promising avenue for global and rapid structure-function analysis, furthering our understanding of RNA biology and the principal features which drive biological insights from SP data.

INTRODUCTION

RNA structure is driven primarily by the complementarity of nucleotide bases comprising it, which allows for hydrogen bonding between various segments of the molecule. Intramolecular base pairing, combined with the flexible and single-stranded nature of the molecule's backbone, allows for intricate secondary and tertiary structural elements. These structures, as well as their ability to dynamically change between relevant configurations, are known to play central roles in almost every facet of cellular regulation (1–6). Understanding the structures of RNA is therefore important, which has led to an explosion of methods which probe (7–17), computationally predict (18–28) and interpret them in various contexts (1,5,29–35).

Structure probing (SP) experiments currently provide the most practical approach for measuring RNA structures in their natural environment. These experiments work by exposing RNA to chemicals, enzymes, or photons which react differentially with parts of the molecule depending on their structural context (e.g. paired/unpaired nucleotides or ds/ssRNA) (7,8,10–13,36,37). Specific protocols vary, but typically the probing reaction induces changes to the RNA bases or backbone which are detected via sequencing or electrophoresis as mutations or truncations (38,39). The rate of mutation or truncation at a particular nucleotide is used to summarize that nucleotide's reactivity with the probe (40). These data contain critical information on the structural conformation of an RNA, and incorporating them as soft constraints within thermodynamics-based folding algorithms greatly improves their accuracy (18,26,41).

Next-generation sequencing has allowed SP experiments to scale to the level of the whole cell (i.e. transcriptome-wide). Exploration of these data have typically begun with straightforward global-level quantifications and simple comparisons (11,42–46). More recent studies expanded the intricacy of structural analysis to disentangle the dynamic functional roles of RNA structure in fundamental cellular processes (47). For example, Saha *et al.* compared reactivity profiles in the vicinity of spliced introns and retained introns, and found evidence of increased structure upstream and decreased structure downstream of retained introns (48). Yang *et al.* characterized structural impacts on miRNA-mediated mRNA cleaving by computing mean reactivity and mean base-pairing probability profiles around miRNA target sites, which illuminated a strong connection between transcript cleavage and unpaired bases immediately downstream of the miRNA target site (49). Works by Mustoe *et al.* (30) and Mauger *et al.* (50) have linked changes in gene expression within *Escherichia coli* and human cells to the structural dynamics within coding

*To whom correspondence should be addressed. Tel: +1 530 752 6978; Email: saviran@ucdavis.edu

sequences and UTRs as quantified by local median reactivities. A slew of recent works have investigated the role of RNA structure within the interplay between RNA helicases and transcription termination, alternative splicing, translation initiation and translation efficiency (51–54). Twittenhoff *et al.* (55) performed structure probing of *Yersinia pseudotuberculosis* at different temperatures and used averaged reactivity scores to highlight differential structure changes due to temperature in 5′UTRs versus coding regions in addition to using condition-wise reactivity differences to identify temperature-sensitive genes.

A common theme to such studies is the quantification of local 'structuredness' and comparisons of it at global scales. To this end, measures of structure are typically founded on basic statistical summarization of reactivities, sometimes combined with data-directed thermodynamics-based folding algorithms to quantify base-pairing probabilities. Current state-of-the-art algorithms for predicting base-pairing probabilities (and specific RNA structures) are founded on dynamic programming strategies and a nearest neighbor thermodynamic model (NNTM) (56,57). Although relatively efficient, these scale as $O(L^3)$ with the length of an RNA, meaning that complete folding analyses of long RNA transcripts are often computationally infeasible. NNTM-based processing (i.e. RNA folding and computation of base-pairing probabilities) of the massive data associated with recent studies is thus challenging. As a consequence, transcriptome-wide studies have typically utilized ad-hoc folding strategies which attempt to strike a balance between computational overhead and prediction quality by locally folding pre-screened candidate regions or rolling windows of long transcripts. Even with such compromises, *in silico* analyses can take days to complete, depending on the scale of the experiment. The process itself is also susceptible to high error rates especially in molecules with multiple stable conformations (58). It is worth noting that some of the aforementioned experiments relied solely on simple reactivity summarization; nevertheless, even in such situations, detections are typically limited to the most pronounced effects. More sophisticated analysis which accounts for structure in addition to reactivity has the potential to refine such findings and expand on them (59,60). This highlights a need for methods capable of rapidly extracting pertinent structural information from reactivity data.

Motivated by this need, we harnessed *patteRNA*, an NNTM-free method we previously introduced for rapidly mining structural motifs (61,62) to quantify global trends in RNA structure dynamics from SP data. Briefly, the method works in two phases: training and scoring. The training phase learns a hidden Markov model (HMM) of secondary structure and a Gaussian mixture model (GMM) of the reactivity distributions of paired and unpaired nucleotides (see Figure 1A). The learned distributions are used to score sites for their likelihood to harbor any target structural motif (see Figure 1B). *patteRNA* can automatically process data from any type of SP experiment. Although we previously demonstrated that *patteRNA* accurately detects structural motifs in diverse datasets, we found that there was nevertheless room for significant improvement. Namely, there was a need for improved precision of motif detection, particularly pertaining to the vast search space encountered in transcriptome-wide experiments. Additionally, we found that our method, although suitable for comparative analysis of motifs (62), did not provide a clear quantitative framework for making practical and direct structural inferences in large datasets.

In this article, we expand and improve the capabilities of *patteRNA* and demonstrate that motif detection can be used to rapidly quantify RNA structuredness in SP datasets. As a first step, we investigate the properties of hairpin elements in RNA structures and their prevalence among all structural elements, revealing that hairpins readily detectable by *patteRNA* (hairpins without bulges) constitute over 30% of paired nucleotides. We then present an improved unsupervised training approach which yields more accurate motif detection, especially for hairpins, and benchmark it against diverse types of data. Next, we describe a novel measure, the hairpin-derived structure level (HDSL), which uses *patteRNA*'s detected hairpins to quantify the local structure context around nucleotides. We apply HDSL to three recent large-scale SP datasets to demonstrate that our hairpin-driven analysis is (i) capable of recapitulating, strengthening, and expanding on previously detected structural effects and (ii) orders of magnitude faster than comparable NNTM-based routines. Simply put, our method bridges the gap between quick but naïve data summarization and intensive but more sophisticated folding-based analysis to provide rapid structure-aware interpretations. Overall, the results of our work also serve to further our understanding of the ways in which diverse SP datasets can be automatically quantified and interpreted without dependence on the assumptions driving NNTM predictions and the complexities associated with them.

MATERIALS AND METHODS

Data

Details about the datasets used throughout this study are compiled in Table 1. In short, seven datasets were used. Central to the development of our method is the Weeks set, a diverse dataset of 22 non-coding RNAs with high-quality *in vitro* SHAPE data and known structures (∼10 000 nt total) (61). We used this dataset to perform benchmarks as well as to query the structural properties of structured RNAs (i.e. the representation of hairpins within them). Reference structure models were also obtained from the RNA Secondary Structure and Statistical Analysis Database (RNA STRAND) (63) and Rfam (64) to provide a more expansive set of data by which to query hairpin representation and characteristics. The remaining four datasets are recent SP datasets on which we applied *patteRNA* to demonstrate its suitability for obtaining biologically relevant insights in various contexts. This includes transcriptomic data for mRNAs *in vitro* and *in vivo* in *E. coli* (30), *in vitro* and *in vivo* reactivities for the SARS-CoV-2 genome (33), *in vitro* reactivities for the HIV-1 genome for three chemical probes (65), and *in vitro* and *in vivo* transcriptome-wide reactivities for two human cell lines, K562 and HepG2 (32). References for the sources of each dataset are provided in Table 1 with accession numbers included where applicable.

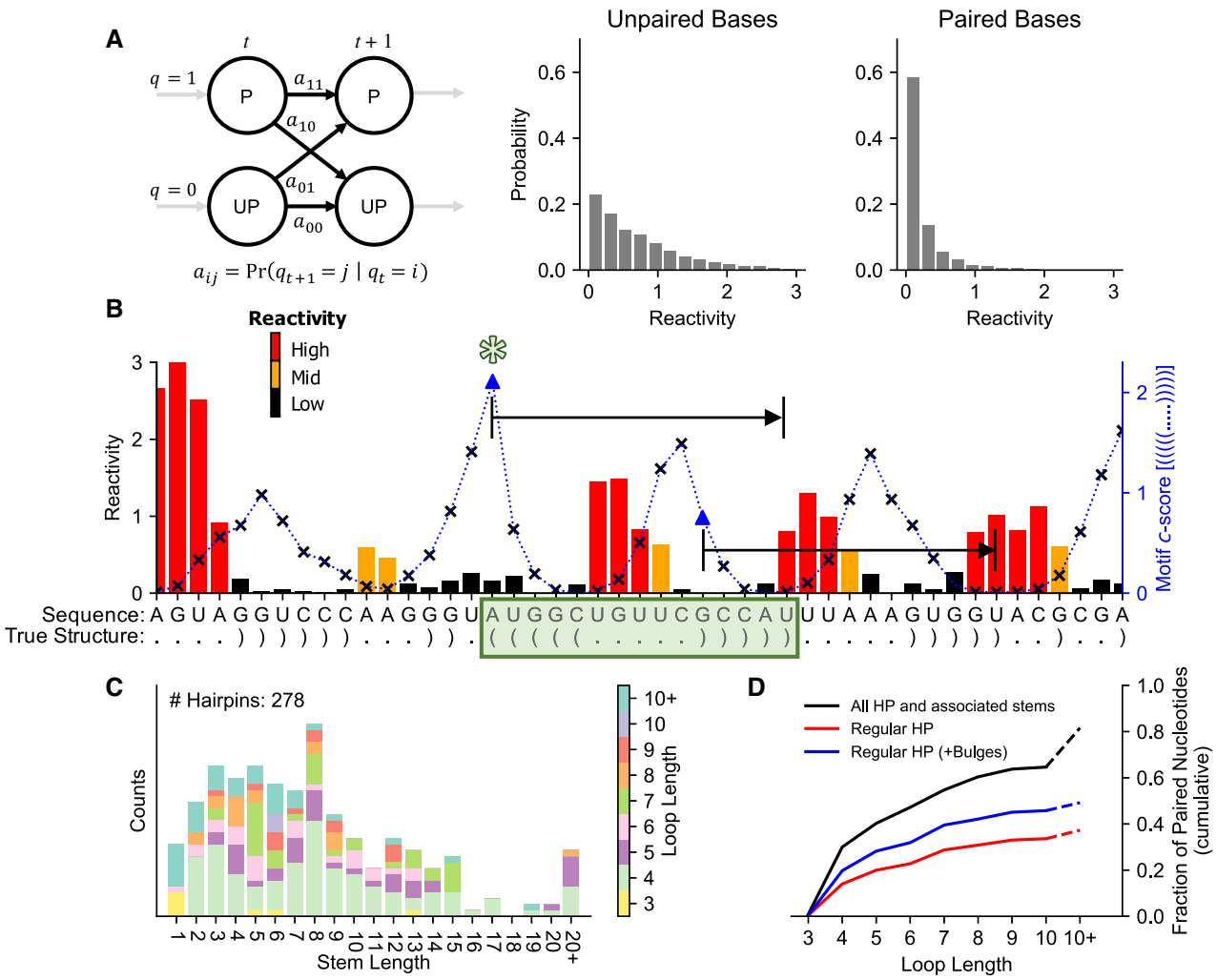

Figure 1. Identification of structural motifs in probing data and representation of hairpins in structures. (**A**) Key components comprising *patteRNA*'s statistical model of probing data. A Hidden Markov Model (HMM) is used to describe the tendency of RNA to transition between paired and unpaired states across adjacent nucleotides ($a_{ij} = \Pr(q_{t+1} = j \mid q_t = i)$), while an emission model of reactivity captures the distributions of SP observations associated with paired and unpaired states, respectively. (**B**) Schematic illustrating reactivity profile (black, yellow, red) for a region against the corresponding *patteRNA* *c*-score profile (blue) when mining for a hairpin with loop length 5 and stem length 5 (dot-bracket: '((((((.....))))))'). The score profile represents the likelihood of the target motif occurring at the site corresponding to using the current nucleotide as the start (left side) of a sliding window. This profile achieves a maximum at the true positive site of the mined hairpin (score indicated with star, site indicated as green box). Locations which satisfy sequence constraints necessary for the base pairs of the motif are denoted by triangle-shaped markers on the score profile, and vice versa for *x*-shaped markers (thus, only sites denoted with triangles are considered by *patteRNA* when scoring). The precise bounds of the sites which satisfy the sequence constraints of the motif are also indicated with black arrows. Data shown are SHAPE-Seq reactivities from the 23S rRNA of *E. coli* (nt 2531–2576) (41). Reactivities are color coded according to their magnitude (high: > 0.7; mid: > 0.3 and ≤ 0.7; low: ≤ 0.3). (**B**) Distribution of hairpin stem and loop lengths in a diverse set of structured RNAs (referred to as the Weeks set; see Materials and Methods). The vast majority of hairpins have stem lengths shorter than 15 nt and loop lengths between 3 and 10 nt. (**C**) Fraction of paired nucleotides in the Weeks set which can be represented as belonging to a regular hairpin (red), a regular hairpin with up to one or two bulges of length 1–5 nt (blue), or any/all type of hairpin and associated stems (black).

Note that for RNA STRAND data, the entire collection of structure models was not utilized. STRAND houses 4666 high-quality RNA structures as determined from NMR, X-ray crystallography or comparative sequence analysis. For our work, we heuristically pruned the number of structures significantly (to 797 structures) to account for unequal representation of RNA classes within the database (specifically, the overrepresentation of ribosomal RNA structures). This pruning was achieved by sampling a defined number of structures from each RNA type in the database. The total numbers of original structures within each RNA type, as well as the corresponding numbers of RNA structures sampled, are given in Supplementary Table S1. A simple visualization of the fraction of (i) transcripts, (ii) nucleotides and (iii) hairpins in the pruned data coming from each RNA class is given in Supplementary Figure S1. The numbers used for subsampling were heuristically determined but were guided by the composition of pruned data as observed in visualizations like the one shown in Supplementary Figure S1. We found that the utilized values led to a fairly balanced set of data from the perspective of transcript composition, nucleotide composition and hairpin composition.

Table 1. Summary of datasets used throughout this study

Dataset name	Description	Size	References
Weeks set	22 well-studied RNAs with reference structures and high-quality SHAPE data	11 070 nt	(18,21,41,61–62)
STRAND data	797 diverse RNAs with experimentally determined structures (via NMR, crystallography or comparative sequence analysis) [no probing data]	276 290 nt	This work, (63)
Rfam data	Secondary structure models informed by covariance models for 3935 RNA families [no probing data]	526 608 nt	(64)
Manfredonia data	SARS-CoV-2 genome probed by: • *In vitro* DMS-MaP • *In vitro* DMS-MaP • *In vivo* SHAPE-MaP	3 × 29 903 nt	(33), GSE151327
Siegfried data	HIV-1 genome probed *in vitro* with 1M6, 1M7 and NMIA (SHAPE-MaP)	3 × 9174 nt	(65), SRX554885
Mustoe data	194 *E. coli* mRNA transcripts probed by SHAPE-MaP across three conditions (each condition is the average of two replicates) • Cellfree (*in vitro*) • Incell (*in vivo*) • Kasugamycin (*in vivo* + 10 mg/ml kasugamycin)	3 × 442 421 nt	(30), PRJEB23974
Corley data	*In vivo* and *in vitro* icSHAPE data (as well as fSHAPE data, not included in the dataset size) for RNA transcripts in two human cell lines: K562 and HepG2 (each condition is the average of two replicates)	2 × 40.8 million nt (K562) 2 × 35.4 million nt (HepG2)	(32), GSE149767

Hairpin counting and quantification in known structures

To better understand the representation of hairpins within RNA structures, we parsed sets of reference structures (the Weeks set, STRAND data and Rfam data) and denoted hairpin elements according to three schemes: (i) all hairpins (hairpins and associated stems, with and without bulges), (ii) regular hairpins (hairpins with stem length between 4 and 15 nt and loop length between 3 and 10 nt without bulges or internal loops) and (3) regular hairpins with and without bulges. The specific definitions used for each scheme are as follows (see Supplementary Figure S2 for an example structure with defined hairpin motifs indicated). In all cases, loops which are involved in pseudoknotted base pairing are treated as unpaired loops for the purpose of hairpin identification.

All hairpins (hairpins and associated stems, with and without bulges). Hairpins in reference dot-bracket structures were retrieved by first identifying hairpin-loops and then backtracking to determine the full stem length. Hairpin loops are defined as locations in the dot-bracket structures where a base pair flanks a sequence of unpaired states of any length. Once a hairpin loop is identified, the stem length is determined by walking along the structure in both directions until a branching base pair is encountered (i.e. a ‘)’ to the left of the stem-loop or a ‘(’ to the right). At this point, the stem length is called as the number of nested base pairs before the first branching base pair on either side of the stem. Unpaired bases are ignored while traversing the local structure, so the entire nested scope of stems with bulges and internal loops is included.

Regular hairpins (hairpins without bulges or internal loops). We defined regular hairpins as hairpins having a stem length between 4 and 15 nt and loop length between 3 and 10 nt with no bulges or internal loops within the helix. For these 96 distinct motifs, identifying their locations amounts to simply searching the dot-bracket data for the exact dot-bracket sequence defined for each hairpin size. For example, a regular hairpin with stem length 4 and loop length 4 has dot-bracket sequence (‘((((....))))’).

Regular hairpins with and without bulges. Identifying locations of regular hairpins with up to one or two bulges was performed similarly to the identification procedure used for regular hairpins without bulges. However, due to the combinatorial explosion of qualified motifs when allowing for bulges, we used a regular expression scheme to perform the search. The regular expression has the form ‘({2,10}.{0,5}({3,10}.{3,MAXLOOP}){3,10}.{0,5}){2,10}’ where MAXLOOP is the maximum loop length to include in the search. This regular expression, in order to permit flexibility for the position of bulges along the stem when identifying hairpins with bulges, also matches some motifs with stem lengths longer than 15 nt. As such, any constructed structure patterns with a stem longer than 15 nt through were discarded prior to the search.

Discretized Observation Model (DOM)

The discretized observation model serves as an alternative approach for describing the probabilities of a particular state (unpaired/paired) to yield a particular reactivity value (state emission distributions). Typically, the emission distributions are modeled as continuous distributions, as is the case when *patteRNA* uses a GMM of reactivity. However, the DOM framework instead discretizes reactivities based on percentiles, then constructs probability mass functions (PMFs) over the discrete reactivity classes for the two pairing states. The state PMFs are then learned in an unsupervised fashion by coupling the emission model to an HMM and performing expectation-maximization (EM) optimization of parameters, analogously to the original GMM implementation. Also analogous to the GMM’s number of

Gaussian kernels, the resolution of bins used in the DOM is gradually increased until an optimal model is reached via a minimum in Bayesian information criteria (BIC) (62). Typically, 7–10 bins are deemed optimal.

A more complete description of the mathematical formulation behind the DOM, including initialization and M-step parameter updating, is available in Supplementary Material.

Scoring with *patteRNA*

patteRNA mines structural elements as represented in dot-bracket notation. In the context of *patteRNA*, this representation of a structure is referred to as a target motif. To mine for a motif, *patteRNA* first encodes the structure as a sequence of pairing states (states denoted as $i \in \{0, 1\}$, where 0 is unpaired and 1 is paired), called the target path. Then, all possible locations in the data are scored for the presence of the target path. With sequence constraints enforced, this amounts to all sites in an RNA where the nucleotide sequence permits folding of the target motif via Watson-Crick and Wobble base pairs (sequence constraints can also be manually disabled, and in such situations all windows of length equal to the length of target motif are considered—i.e. a full sliding window approach). Regardless of sequence constraints, the *patteRNA* score for a site (a window of length n beginning at nucleotide m) is defined as the log ratio of joint probabilities between the target path and its inverse path (i.e. the opposite binary sequence) (61). More specifically,

$$\text{score } (z|y) = \log \frac{\Pr(y,\ z|\theta)}{\Pr(y,\ z'|\theta)}.$$

Here, y is the reactivity profile at a site, z is the target binary state path, z' is the inverse path, and θ represents the parameters of a trained GMM/DOM-HMM model. The parameters of the trained model include the transition ($a_{i,j}$ for states i and j) and initial probabilities for paired and unpaired states within the Markov model, as well as an emission model (either a GMM or DOM) that describes the likelihoods of paired and unpaired states to yield specific reactivity values. For a GMM (61), the emission model is parameterized by Gaussian weights, means and variances ($w_{i,k}$, $\mu_{i,k}$, and $\sigma_{i,k}$ respectively, where k corresponds to an individual Gaussian kernel in the learned mixture distributions). For a DOM, the emission model is simply parameterized by the learned discrete probability mass functions of paired and unpaired nucleotides ($p_{i,k}$, where k is a bin in the discretization scheme). A trained GMM/DOM-HMM model enables computation of $b_{i,t}$ (the emission likelihood for state i at nucleotide t) as well as $\alpha_{i,t}$ and $\beta_{i,t}$ (the forward and backward probabilities for state i at nucleotide t, respectively, as computed via the forward-backward algorithm (66)). For the full formulation of emission likelihoods when using GMMs and DOMs, see the Supplementary Material.

The simplified score representation given above can be written in an expanded form by considering the Markov framework used to model pairing state along transcripts. Recall that the forward ($\alpha_{i,t}$) and backward ($\beta_{i,t}$) probabilities are defined as the following, where q represents hidden states underpinning the observed data.

$$\alpha_{i,t} = \Pr(y_1,\ \ldots,\ y_t,\ q_t = i \mid \theta)$$

$$\beta_{i,t} = \Pr(y_{t+1},\ \ldots,\ y_T \mid\ q_t =\ i,\ \theta)$$

The probability of the observed data and a target path of length n starting at nucleotide m can therefore be expanded as

$$\begin{aligned}\Pr(y,\ \ z \mid \theta) &= \Pr(y_1, \ldots,\ y_T,\ z_m,\ \ldots,\ z_{m+n-1} \mid \theta)\\ &= \alpha_{z_m,m}\, a_{z_m,z_{m+1}} b_{z_{m+1},m+1} \cdots\\ &\quad a_{z_{m+n-2},z_{m+n-1}} b_{z_{m+n-1},m+n-1} \beta_{z_{m+n-1},m+n-1}\end{aligned}$$

Using product notation to simplify the transition and emission probabilities between the forward and backward terms, we can then write

$$\Pr(y,\ \ z \mid \theta) = \alpha_{z_m,m} \left(\prod_{t\,=\,m}^{m+n-2} a_{z_t,z_{t+1}} b_{z_{t+1},t+1} \right) \beta_{z_{m+n-1},m+n-1}$$

The final score for the target path z is the log-ratio of this path probability with the inverse path probability. Thus,

$$\text{score } (z \mid y) = \log \left[\frac{\alpha_{z_m,m}}{\alpha_{z'_m,m}} \frac{\beta_{z_{m+n-1},m+n-1}}{\beta_{z'_{m+n-1},m+n-1}} \prod_{t=m}^{m+n-2} \left(\frac{a_{z_t,z_{t+1}} b_{z_{t+1},t+1}}{a_{z'_t,z'_{t+1}} b_{z'_{t+1},t+1}} \right) \right].$$

A score of zero indicates the target path and inverse path are equally likely, and a positive score indicates the target path is more likely (and vice versa). Locations with the highest scores are subsequently deemed most likely to harbor the target motif.

To facilitate the comparative analysis of scores between different motifs and datasets, scores were further processed into c-scores as previously described (62) by normalizing against a null distribution of scores estimated via sampling of scores from locations which violate the sequence compatibility necessary for the motif's base pairs (and therefore can be presumed to not harbor the target motif). The resulting c-scores are the $-\log_{10}$ of a P-value, meaning they are strictly positive and theoretically have no upper bound. That said, a c-score above 2 is intuitively considered a strong indicator of the motif (corresponding to a P-value of 0.01), with c-scores between 0.5 and 2 providing moderate evidence in favor of the motif. Example SP data with real *patteRNA* c-scores superimposed is illustrated in Figure 1B.

Posterior pairing probabilities

patteRNA computes posterior pairing probabilities as described (61). Briefly, a parameterized GMM-HMM or DOM-HMM model is utilized to compute emission likelihoods for each nucleotide, followed by the forward and backward probabilities via the forward-backward algorithm. Posteriors are then computed as the product of the forward and backward probabilities and appropriately scaled such that $P(\text{paired}) + P(\text{unpaired}) = 1$ for each nucleotide. Note that this computation is a special case of the path probabilities used during scoring; posteriors are analogous to path probabilities where the length of the target path is simply a single state, either paired or unpaired.

Hairpin-driven structure level (HDSL)

The hairpin-driven structure level (HDSL) is a nucleotide-wise measure quantifying the local level of structure from SP data. HDSL is initialized using posterior pairing probabilities as computed by *patteRNA*. This profile is then augmented using hairpin *c*-scores calculated by *patteRNA*. For each detected hairpin with *c*-score >0.5, the value $0.2 \cdot$ (*c*-score – 0.5) is added to the profile at all nucleotides covered by the hairpin. After profile augmentation, profiles are clipped to the interval [0, 1], and then profile smoothing is achieved via a 5 nt sliding-window mean followed by a 15 nt sliding-window median to give the final HDSL profile. Analogous approaches using just a sliding mean or just a sliding median were also tested, but we found that the best results were obtained when coupling the two summary statistics together (data not shown).

The parameter values used in profile augmentation (e.g. a slope of 0.2 and a *c*-score threshold of 0.5) were determined by a grid-based optimization scheme seeking to maximize the observed difference between HDSL for nucleotides in well-folded segments of the SARS-CoV-2 genome and HDSL for nucleotides outside of these regions (see Supplementary Figure S3). In this context, well-folded segments were defined as low SHAPE, low Shannon entropy regions as called by Manfredonia *et al.* (33). The SARS-CoV-2 genome was selected for this optimization as it is distinguished from the other datasets by having both regions of high structure and un-structuredness (compared to the Weeks set, which is generally highly structured) in addition to a partially validated preliminary reference structure model (compared to the Mustoe or Corley data, which lack reliable structure models). Note that the results shown in Supplementary Figure S3 demonstrate a large region of HDSL parameterizations which greatly improve the distinction between well-folded and less-folded segments over posteriors alone (see top left cell of each heatmap in Supplementary Figure S3 as approximately representing the use of posteriors alone). In other words, other parameterizations arrived at similar results to the chosen parameterization. Generally speaking, we observed that as the *c*-score threshold is increased, the slope of augmentation must also be increased in order to allow the reduced number of considered sites to sufficiently impact the final HDSL signal. It is also important to note that smoothed pairing probabilities on their own can serve as a meaningful measure of local structure (without augmentation at detected hairpin elements). In tandem to HDSL, we explored the use of smoothed P(paired) (pairing probabilities smoothed via a 5 nt rolling mean and 15 nt rolling median) when quantifying structure trends to better understand the effects of hairpin augmentation on the HDSL approach.

A flow chart illustrating the flow of information as handled by *patteRNA*, including the relationship between HDSL and the training and scoring phases, is included as Figure 2. In summary, HDSL integrates *patteRNA*'s normalized scores (*c*-scores) for hairpins with posterior pairing probabilities to arrive at a nucleotide-wise measure of structuredness. Whereas hairpin *c*-scores (and non-normalized scores) are assigned only at specific sites in the data which satisfy the sequence base pairing requirements of a hairpin motif, HDSL is computed at all nucleotides. This is because all nucleotides are assigned a posterior pairing probability via the GMM/DOM-HMM. Hairpin scores are used to augment this profile to improve its relevance to local structure elements, but regions lacking any strong hairpins scores are still assigned pairing probabilities and as such are assigned HDSL based on those outputs.

patteRNA training and scoring

All *patteRNA* analyses were performed with default training parameters (KL divergence for training set $D_{\mathrm{KL}} = 0.01$, convergence criterion $\varepsilon = 0.0001$, automatic determination of model complexity, k, via Bayesian information criteria) (62). With the exception of benchmarks investigating the effect of log-transforming data, log-transformed data were always used when using a GMM and non-transformed data were used when using a DOM. Scoring for regular hairpins was achieved using the '--hairpins' flag, computation of HDSL profiles was achieved with the '--HDSL' flag, and computation of smoothed P(paired) profiles was achieved with the '--SPP' flag. Sequence constraints were always enforced when mining hairpin motifs.

Computation of statistical performance metrics

The accuracy of *patteRNA* to detect motifs is primarily assessed through the receiver operating characteristic (ROC) and precision-recall (PR) curves. These curves were computed by varying a theoretical *c*-score threshold between called positives and negatives and, at each threshold, computing the true-positive rate (TPR/recall), false positive rate (FPR) and precision (also referred to as positive predictive value, PPV). A site is deemed a positive if all base pairs in the target motif are also present in the corresponding location of the reference structure. These performance profiles are then visualized (ROC: FPR versus TPR, PR: TPR versus PPV) and summarized using the area under the curve (AUC) of the ROC and average precision (AP) of the precision-recall curve. The Scikit-learn Python module (v0.24) was utilized to perform these computations.

Simulated datasets and benchmarks

We generated simulated data for RNAs in the Weeks set by sampling reactivities according to various state distributions schemes (see Table 2). 50 replicates of each scheme were generated for the performance benchmarks using in-house Python scripts. *patteRNA* was then used the train and mine the replicates for regular hairpins using the 'patteRNA ${SHAPE} ${OUTPUT} -f ${FASTA} [--GMM or --DOM] --hairpins' command. The '-l' flag was added to use log-transformed data where applicable; training was performed independently for each replicate. Overall performance for a scheme was summarized as the mean of average precisions for the 50 replicates.

Averaging and integrating HDSL over mRNA coding sequences

We delineated the regions surrounding the 432 genes in the Mustoe data into four groups: (i) start site; ±30 nt around

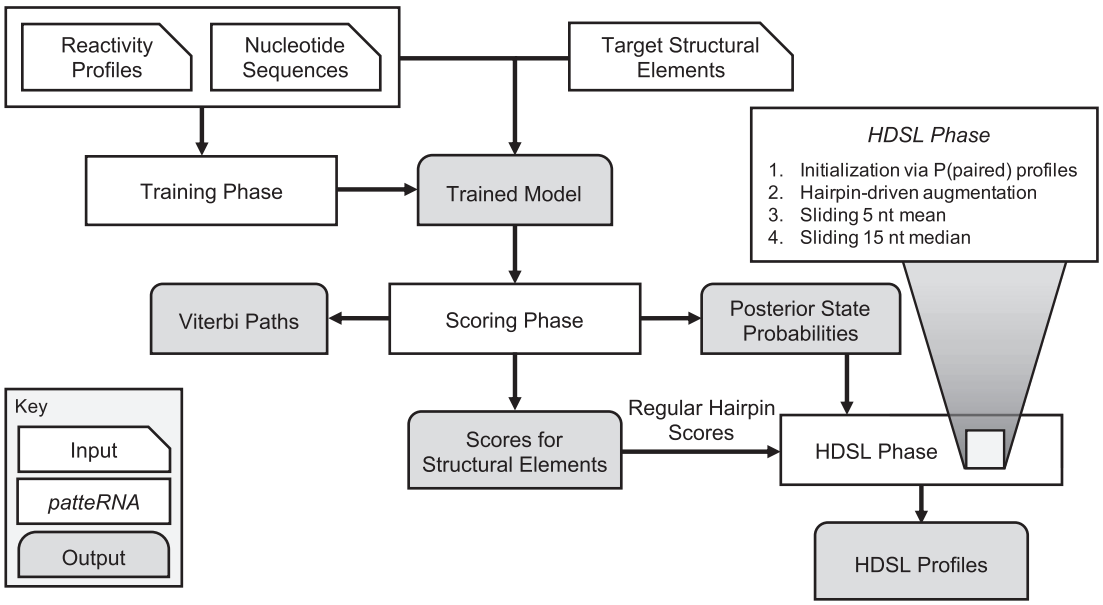

Figure 2. Overall flow of data and computing behind *patteRNA* and hairpin-derived structure level (HDSL). The measure is initialized as the pairing probability profiles, which are then augmented by boosting values at sites covered by highly scored hairpins (see Materials and Methods section). The subsequent profile is clipped to the interval [0, 1] and local smoothing is achieved with sliding window mean and sliding window median approaches with windows of size of 5 and 15 nt, respectively.

Table 2. Parameters of state distributions used to generate artificial data on the Weeks set. GEV: generalized extreme value

Scheme name	Paired distribution	Unpaired distribution
Heitsch distributions (69)	*Helix-end*: GEV($\mu = 0.09, \sigma = 0.114, \xi = -0.821$) *Stacked*: GEV($\mu = 0.04, \sigma = 0.040, \xi = -0.763$)	Exponential distribution with $\lambda = 1.468$
Gaussian / Gaussian (poor)	Gaussian distribution with $\mu = 0, \sigma = 1$	Gaussian distribution with $\mu = 0.5, \sigma = 1$
Gaussian / Gaussian (medium)	Gaussian distribution with $\mu = 0, \sigma = 1$	Gaussian distribution with $\mu = 1, \sigma = 1$
Gaussian / Gaussian (high)	Gaussian distribution with $\mu = 0, \sigma = 1$	Gaussian distribution with $\mu = 2, \sigma = 1$
Exponential / Gaussian	Exponential distribution with $\lambda = 2$	Gaussian distribution with $\mu = 2, \sigma = 1$
Exponential / Exponential	Exponential distribution with $\lambda = 2$	Exponential distribution with $\lambda = 1/2$

AUG, (ii) 5′UTR; −70 to −31 nt from AUG, (iii) 3' UTR; +1 to +40 from STOP codon and (iv) coding sequences; +31 nt from AUG to the STOP codon. For the start site, 5′UTR, and 3′UTR, HDSL averages were taken at each aligned position as these groups each have a constant length. For situations where all regions might not exist for a gene, aligned HDSL profiles were included in the analysis as far as the nucleotide sequence allowed, and remaining positions were treated as missing values and omitted from subsequent averaging. For instance, if the 5′UTR was 50 nt (i.e. <70 nt), those 50 nt were aligned with the corresponding locations and the missing 20 nt upstream were treated as missing values. For coding sequences (which inherently have a non-constant distribution of lengths), the profiles were interpolated to a vector of length 300 to allow for aligned averaging relative to the beginning and end of the window. About 99% confidence intervals were computed using the Wald formulation (mean HDSL $\pm$ 2.576 $\cdot$ SE).

Footprinting SHAPE (fSHAPE) analysis

The Corley data comprise *in vitro* and *in vivo* icSHAPE reactivities and footprinting SHAPE (fSHAPE) scores for over 10 000 transcripts from two human cell lines, K562 and HepG2 cells, respectively (32). For each cell type, reactivities and fSHAPE scores were taken as the average of two replicates for each condition following the work by Corley *et al.*. Averaged reactivities from each condition were independently processed by *patteRNA* to train a model, mine for regular hairpins, and compute HDSL and smoothed P(paired) profiles. Structuredness profiles for each condition-cell type were cross-referenced with the fSHAPE data using Python scripts. Observed distributions were compiled based on three fSHAPE data groups, as defined by Corley *et al.*: high (fSHAPE $\geq$ 2), moderate (2 > fSHAPE $\geq$ 0) and low fSHAPE (fSHAPE < 0).

Hairpin mining performance of NNTM partition function approach

We benchmarked the performance of partition function approaches to detect hairpins in the Weeks set by using the 'RNAsubopt' command from ViennaRNA to generate 1000 structures for each transcript in the Weeks set, using that transcript's SHAPE data as soft constraints ('RNAsubopt -p 1000 --shape ${SHAPE_FILE} <

${SEQUENCE}'). For each regular hairpin in the generated structural ensemble, a 'score' was assigned as the fraction of structures in the structural ensemble which contain the base pairs comprising that hairpin. Predicted hairpins and their scores were processed into a receiver operating characteristic and precision-recall curve as done for *patteRNA*'s hairpin scores (see *Computation of Statistical Performance Metrics*).

Local folding calculations

To measure the required compute time to process SP datasets with a local partition function workflow, windowed partition function calculations were performed using the 'RNAfold -p' command from ViennaRNA (19). Three schemes were utilized: windows of length 3000 nt, spaced 300 nt apart; windows of length 2000 nt, spaced 150 nt apart; and windows of length 150, spaced 15 nt apart. In each case, sequences within each window were parsed using custom Python scripts and then processed sequentially with *RNAfold*. Only the time required to run *RNAfold* commands was measured in timing benchmarks (no integration of windowed outputs or post-processing were accounted for). *RNALfold* benchmarks were performed using the default arguments of the command to process all sequences in the Corley data sequentially. All timing comparisons in this study were performed on an AMD Ryzen 9 5900X CPU running Ubuntu 20.04 LTS.

RESULTS

Overview of *patteRNA* mining

To mine structure elements from SP data, *patteRNA* first learns the statistical properties of the data via the training phase. The purpose of this procedure is to estimate the distributions of reactivities associated with paired and unpaired nucleotides, respectively, as well as the HMM's transition probabilities between paired and unpaired nucleotides (Figure 1A). Training is unsupervised and has been shown to accommodate diverse data distributions (see Ledda *et al.* (61) for a complete description). With the dataset characterized via its statistical model, *patteRNA* can then mine for structural motifs.

Figure 1B demonstrates key concepts related to *patteRNA*'s motif mining. When mining a particular structural element (i.e. the target), sites which satisfy the sequence constraints necessary for the target's secondary structure are scored for their probing data's consistency with its pairing state sequence (61,62). Sites which do not satisfy sequence constraints can also be scored; however, these sites are almost certainly all negatives and can therefore be discarded (the only exception being the possibility of non-canonical base pairs). Sequence constraints provide an important filtering step at the start of a search; therefore, they were always enforced when utilizing *patteRNA* in this work. Sites which harbor the target motif presumably have SP data consistent with the desired state sequence and therefore score highly. *patteRNA*'s overall objective is to identify sites harboring particular structural elements, such as hairpins, as accurately as possible.

Hairpins comprise a significant portion of structural elements

To assess the plausibility of a hairpin-centric approach in making general assessments of structure, we examined a diverse dataset of 22 RNAs with known structures (∼10 000 nt) (61) to quantify the distribution of hairpins present as well as the proportion of base pairs contained within hairpins. We refer to this dataset as 'the Weeks set.' Analyzing the 278 distinct hairpins in the Weeks set reveals that a majority fall within a narrow range of stem and loop lengths (Figure 1C). Specifically, hairpins most frequently have loop lengths between 3 and 10 nt, and stem lengths 15 nt or less. In other words, although their properties are diverse, there is a range of stem and loop sizes which represents a majority of hairpins (83%). Later in the study will we leverage these characteristic properties to focus our searches on this most representative subset of hairpins.

Our results also illustrate that hairpins comprise a large fraction of structural elements. We first focused on hairpins with no bulges or internal loops (i.e. unpaired stretches flanked by some number of base pairs), which we call regular hairpins, and found that around 35% of paired nucleotides reside in such structures (Figure 1D). If you also consider hairpins with up to two bulges each with length up to 5 nt, this coverage increases to over 50%. This suggests that, although hairpins are only a subset of RNA structural elements, they are indeed the most prevalent, and therefore identifying them in SP data could provide a strong quantification of general structural trends.

Understanding that the Weeks set is a small sample of structures to draw conclusions from, we repeated this hairpin counting and quantification on a diverse set of 797 reference structures from the STRAND database (63) and 3935 reference consensus structures for RNA families in Rfam (64), representing a more complete profile of structured RNA properties. The distributions of hairpins in these datasets are shown in Supplementary Figure S4 and recapitulate the observations from the Weeks set. The STRAND data suggest that regular hairpins specifically comprise a slightly larger fraction (40%) of structural elements than is seen in the Weeks set (35%), while the Rfam data suggest this fraction is slightly less (30%). We noted that the Rfam data were slightly biased by an overrepresentation of microRNA families, typically comprised by long (>20 nt) stem-loops. As such, Supplementary Figure S4 also shows the representation of hairpins in Rfam when microRNAs are removed. In this case, we observe that hairpin trends align closely to what is observed with STRAND and the Weeks set, with approximately 35 to 40% of paired nucleotides residing in regular hairpins.

One can further expand the definition of a hairpin to also include the associated stems that extend from a hairpin element up to the first nucleotide that base-pairs outside of the nested context of this element (see Supplementary Figure S2 for examples). We refer to these helices as external stems and note that such motifs are prevalent in structured RNAs. Figure 1C shows that relaxing the definition of a hairpin to include external stems leads to over 80% coverage of paired nucleotides, with the remaining ∼20% of base pairs described by longer-range interactions—e.g., internal stems (see dashed red frame in Supplementary Figure S2)

and pseudoknots. Although external stems are nevertheless outside the scope of the *patteRNA*-based analysis that follows, this high coverage indicates that a large majority of RNA structure can be represented as simple motifs with local base pairing. Moreover, it's important to note that virtually all types of canonical RNA structure motifs necessarily exist in the context of hairpin elements—internal stems, multibranch junctions, etc., only exist in the presence of hierarchical domains which all terminate in a hairpin-like fashion.

In the context of *patteRNA*, we note that there are practical limitations on the types of searches that can be performed. Specifically, although structures comprised by internal loops, bulges and external stems are within the permitted scope of minable motifs described solely by local base pairing, the automated identification of such motifs in SP data is computationally burdensome. This is due to the combinatorial explosion of considered motifs associated with allowing for flexibility in the position and size of internal loops and bulges. For instance, regular hairpins are comprised by 96 distinct motifs (12 stem lengths and 8 loop lengths), but regular hairpins with bulges (as defined in this work) are comprised by a set of motifs with size >20 000 due to the many possible bulge locations and sizes within each regular hairpin motif. Allowing for the presence of various internal loops further increases the space of motifs by orders of magnitude. Although permitted by *patteRNA*, such more comprehensive searches scale poorly to transcriptome-wide applications. As such, the analyses that follow generally focus on mining and assessment of regular hairpins.

Simplified reactivity model improves accuracy of motif detection

In an attempt to improve *patteRNA*'s performance, we investigated alternative statistical models of reactivity and their downstream effects on scoring accuracy. While the GMM approach performs well, especially at the task of approximating the underlying state distributions, we encountered issues in motif scoring. Namely, reactivities from the tails of the overall data distribution would be strongly predicted to be paired or unpaired. This isn't an inherent problem, as the most extreme reactivities should theoretically be the best candidates for confident prediction. However, these reactivities present problems during scoring as they have the propensity to dominate the score for sites they fall into. In other words, a single extreme reactivity consistent with the target state sequence could yield a high score for a site, even if data within that site is otherwise inconsistent with the target (and vice versa). Generally speaking, for SP data such as SHAPE, the most extreme reactivities are only about 3–5 times more likely to be in one state over the other (67), yet the GMM often arrives at likelihood ratios 10 or 100 times larger than this empirical ratio. Such predictions have negative consequences on the interpretation of scores.

Motivated by these issues, we devised a simplified framework for unsupervised learning of the state reactivity distributions. It entails a discretized observation model (DOM) which substitutes for the GMM component of the statistical model (i.e. the emission probabilities), resulting in a DOM-HMM model of SP data. The DOM entails modeling reactivities as a discrete distribution where they are binned into classes based on percentiles. During training, pseudocounts are estimated for each class (E-step) and then utilized in the M-step to infer the discrete reactivity distribution for paired and unpaired states. A schematical comparison of the GMM and DOM approaches is shown in Figure 3A (see Materials and Methods and Supplementary Material for a complete mathematical formulation).

We benchmarked the capacity of *patteRNA* to identify regular hairpins in the Weeks set via the GMM and DOM. We assessed their discriminatory power primarily via the receiver operating characteristic (ROC) and precision-recall curve (PRC), which are shown in Figure 3B. Our results indicate that the DOM approach improves both the area-under-the-curve (AUC) of the ROC and the average precision (AP) of the PRC. Although the improvement to AUC appears minor, average precision was increased from 0.48 with a GMM to 0.64 with a DOM. Precision is a crucial performance metric in structure motif mining where the vast majority of scored sites are negatives (even with sequence constraints applied), so the improvements seen in the DOM are important through this perspective. Notably, precision at the highest scores is much better in the DOM compared to the GMM, which is susceptible to numerous negatives at the highest hairpin scores despite decent precision at moderate scores. This is evidenced by the large fluctuations in precision at low levels of recall for the GMM (see the top left of precision-recall plot in Figure 3B). The DOM approach, on the other hand, is far more reliable for returning positive hits at the highest scores. Figure 3B also includes a benchmark for data-directed NNTM folding algorithms which shows that *patteRNA* is, although improved via the DOM, generally unable to match the precision of RNA folding. Notably, NNTM folding was performed with an ensemble-based approach, which, although much slower, outperforms a single MFE calculation (61).

Importantly, the presented results show overall performance on the collection of all regular hairpins, which is comprised predominantly by motifs with shorter stems. Shorter stems present a challenge to *patteRNA*, as fewer base pairs render sequence constraints less effective in controlling the number of negative sites considered in the analysis. When comparing performance on individual motifs, however, we find that *patteRNA* matches the precision of NNTM-ensemble methods for longer stems. In some cases, such as hairpins with stem length 6 and loop length 7, it even surpasses the performance of the NNTM approach (see Supplementary Figure S5). We also observe a universal trend for the DOM to outperform the GMM at the motif-level, further validating its superior performance.

Not only does the DOM improve precision, but the model itself is described by fewer parameters and trains faster than a GMM. As seen in Figure 3C, faster training is achieved in two distinct ways. First, the DOM generally requires fewer EM iterations to converge. Second, EM iterations are significantly faster. The latter is presumably due to the DOM's simpler M-step formulation, which reduces to simple counting as opposed to the GMM which requires multiplication and squaring to update the means and variances of each Gaussian kernel.

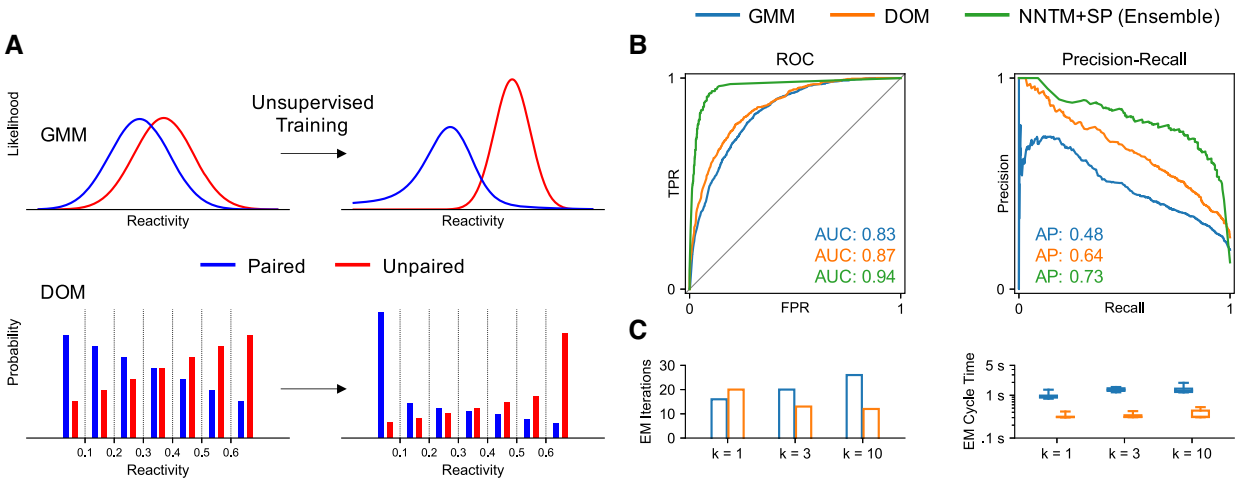

Figure 3. A discretized observation model (DOM) of reactivity improves hairpin detection precision when compared to a Gaussian mixture model (GMM). (**A**) Schematic illustration of GMM and DOM approaches in the content of *patteRNA*'s unsupervised learning scheme. The DOM is founded upon a percentile-based discretization of reactivities which yields a discrete emission probability scheme. The discretization scheme it itself optimized during training based on Bayesian information criteria (BIC) of models using progressively smaller bins. (**B**) Receiver operating characteristic curves and precision-recall curves when mining regular hairpins in a reference dataset ('the Weeks set,' see text) with *patteRNA* using either GMM (blue) or DOM (orange) approaches, or when using data-driven NNTM-based folding (green). (**C**) Timing benchmarks of unsupervised training via GMM and DOM on the Weeks set. Shown are the number of EM iterations required for convergence on the Weeks set and time required for a single EM iteration. Five repetitions were used when measuring EM cycle times.

Given the rapidly evolving field of structure probing and disparate statistical properties of SP datasets (47), we also investigated whether the benefits from the DOM generalize to other data distributions. Different probes have different quality (47,68), different conditions yield different quality (47), and the quality of probes is constantly improving (69); therefore, adaptability of methods is crucial. Benchmark datasets like the Weeks set are not currently available for the plethora of probes used, so we resorted to simulations. We constructed several artificial datasets and benchmarked *patteRNA*'s performance via the GMM or DOM approaches. We sampled reactivities for the underlying structures in the Weeks set according to various state distributions, including empirically fitted distribution models from Sükösd *et al.* (70), referred to as the Heitsch distributions, as well as a collection of mock distributions with varying classification power (i.e. various degrees of separation between the state distributions). For each scheme, 50 replicates were created, and we benchmarked performance against both the regular and log-transformed data. We note that the fidelity of the GMM is dependent on the Gaussianity of the data, presenting a weakness of this approach as the decision to log-transform can have a major impact on scoring efficacy.

The results of the benchmarks are shown in Table 3. Generally speaking, the DOM matches or exceeds the performance of the GMM. Depending on the data properties, the DOM's performance gain ranges from minute to transformative. In only one of the benchmarks did the GMM outperform the DOM (poor quality Gaussian/Gaussian data), and only by a small margin. This specific outcome might be explained by the DOM's simplification of SP data which effectively clips extreme reactivities when discretizing the data. In datasets of poor quality, the most extreme reactivities likely provide the only opportunity for reliable inference on pairing state, so it's possible that the relatively coarse discretization scheme reduces the information content of the data. Regardless, it's worth noting that data of such poor quality is uncommon, especially in light of on-going improvements to experimental protocols and probe quality (8,10,69,71). Our results also demonstrate the adaptability of the DOM and its robustness to non-Gaussian data, which render the method broadly applicable. When using the DOM, log-transforming is largely irrelevant to model performance, as the discretization scheme is founded on data percentiles. The lone exception to this rule is when handling reactivities below zero, which are necessarily binned together if data is log-transformed.

Overall, these results demonstrate the benefit of the DOM approach in more efficiently and effectively mining structures from SP data. Note, however, that the GMM still provides a specific utility when one's objective is to arrive at continuous models of the state reactivity distributions (e.g. to use for simulations, or for data inspection). *patteRNA* includes both implementations such that the respective approach can be used depending on the intended use-case.

Summarizing structuredness in RNAs from hairpin detection

As hairpins comprise a large fraction of structural elements, we sought to utilize *patteRNA* to quantitatively summarize local 'structuredness'. Due to the plethora of cellular processes affected by RNA structures, there are numerous contexts in which summarizing local structure is important. To name a few examples, one might wish to find structural domains and druggable pockets in viral genomes (29,33,72), quantify connections between mRNA structure and gene regulation (43,48,51–54,61,73–75), identify transcriptome-wide where RNA is differentially affected by particular stimuli (32,76), or compare structure between conditions

Table 3. Average precisions of *patteRNA* for hairpin mining when utilizing a Gaussian mixture model (GMM) or discretized observation model (DOM) of reactivity against various artificial data schemes (see Table 2). For all benchmarks, average precision was averaged over 10 replicates. Bold entries highlight the best performing approaches for each scheme. AP: average precision.

	Mean AP			
Data Scheme	GMM	GMM (log data)	DOM	DOM (log data)
Heitsch Distributions	0.43	0.58	**0.63**	**0.63**
Gaussian / Gaussian (poor)	0.32	**0.36**	0.34	0.34
Gaussian / Gaussian (medium)	0.48	0.48	**0.49**	**0.49**
Gaussian / Gaussian (high)	**0.70**	0.62	**0.70**	**0.70**
Exponential / Gaussian	0.58	0.55	**0.71**	**0.71**
Exponential / Exponential	0.52	**0.57**	**0.57**	**0.57**

and/or logical regions of genomes (1,30,77). The most popular approach for quantifying structuredness relies on a combination of two metrics: local reactivity and local Shannon entropy. Local reactivity is generally computed via a rolling mean or median with windows ranging 25–500 nt, while local Shannon entropy derives from base-pairing probabilities computed via NNTM folding routines. The combination of these two metrics yields regions which are largely unreactive (i.e. base paired) and stable (i.e. tending to adopt one conformation). We note that each metric by itself is generally insufficient in this context, as low reactivity regions sometimes include regions which see multiple competing conformations (but are nevertheless highly paired), and low Shannon entropy can also be observed for regions which are preferentially single stranded.

To integrate *patteRNA*'s results into a quantification of structuredness, we propose a nucleotide-wise measure we term the hairpin-derived structure level, or HDSL. At the highest level, HDSL combines *patteRNA*'s computed base-pairing probabilities with information from searches for regular hairpins. This allows us to consider the locations of stable hairpins in addition to the overall pairing propensity of regions, the former of which typically does not account for all structured regions (e.g. internal and external stems, stems with bulges or stems with non-canonical base pairing). Briefly, the posterior pairing probabilities are used as a starting point. We use posteriors as a basis because such a probabilistic interpretation provides a calibrated representation of reactivities that intrinsically handles outliers and enables quantitative comparisons between different SP datasets (78). They are then amplified at nucleotides covered by highly scored hairpins, depending on the hairpin *c*-score—the higher a hairpin is scored, the larger the boost. We refer to this step as the augmentation phase. The profile is clipped to [0, 1] and then locally smoothed by taking a 5 nt rolling mean followed by a 15 nt rolling median (see Figure 2 and Materials and Methods for a complete description). In summary, HDSL integrates posterior pairing probabilities with the locations of detected regular hairpins to arrive at a nucleotide-wise measure of structuredness that is mindful of local structure elements. Whereas *c*-scores quantify the likelihood for specific sites in the data to harbor a specific structure motif, HDSL is computed at all nucleotides and considers a representative collection of 96 regular hairpins simultaneously. This is because all nucleotides are assigned a posterior pairing probability via the GMM/DOM-HMM, and as such, all nucleotides can be assigned HDSL. This is distinct from *c*-scores which are only assigned at sites in the data which satisfy the sequence constraints necessary for the considered targets. We explored the properties of HDSL and validated its utility as an indicator of local structure by applying it to recent datasets that were previously used to assess local structuredness in diverse contexts. Moreover, we also explored the use of locally smoothed pairing probabilities, referred to as smoothed P(paired), as a measure of structuredness derived in the same manner as HDSL but without the augmentation phase. Examining the differences between these measures serves to highlight the contribution of hairpin augmentation when summarizing structuredness with HDSL.

HDSL versus smoothed P(paired) demonstration

The HDSL approach is demonstrated on a representative region of SARS-CoV-2 in Figure 4. At the foundation of the measure are posterior pairing probabilities computed by *patteRNA*, which for the demonstration in Figure 4 were obtained from analysis of *in vivo* SHAPE-MaP data (33). These profiles are then augmented (increased) at the locations of detected regular hairpins (Figure 4A). This step serves to reinforce the structuredness of regions containing hairpin elements, which typically appear as a high-low-high signal from the perspective of raw pairing probabilities. After augmentation, the profile is smoothed via a local mean (5 nt) followed by local median (15 nt) approach (Figure 4B). The structure model of this region as proposed by Manfredonia *et al.* is shown in Figure 4C.

There are three predicted hairpins within this region, and they are all scored highly by *patteRNA*. Therefore, pairing profiles are significantly amplified at nucleotides in these areas (indicated by blue boxes in Figure 4). This has the effect of elevating unpaired nucleotides in hairpin loops from low P(paired) (which contributes 'unstructuredness' to the region from the perspective of local pairing) to a higher magnitude in the augmented profile. The downstream effect of augmentation on quantifying local structure is observed when comparing HDSL to smoothed P(paired) (pairing profiles smoothed analogously to HDSL). As seen in Figure 4B, which shows smoothed versions of the profiles in Figure 4A, examining pairing probabilities alone can obfuscate interpretations of structuredness. This is because loops within stable hairpins are readily predicted to be unpaired. As such, they reduce the local average pairing probability profile (see gray smoothed P(paired) profiles within blue boxes). HDSL rectifies this by overriding the 'low structure' of hairpin loops during the augmentation phase by applying an increase to the pairing profile of the hairpin before smoothing.

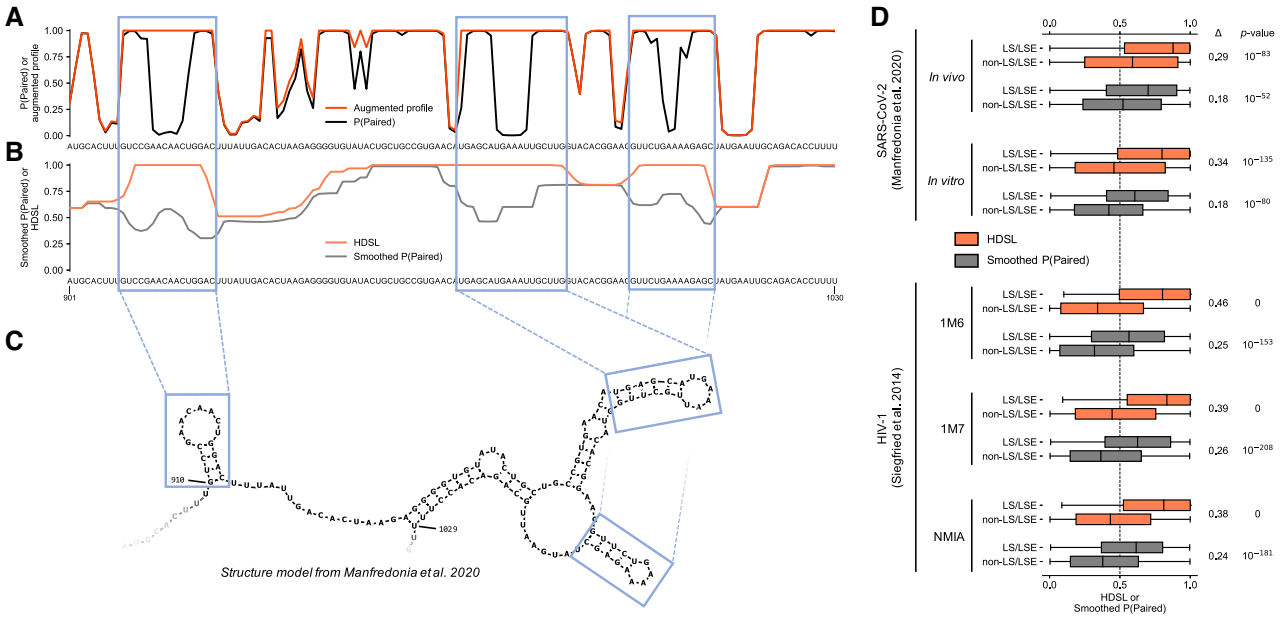

Figure 4. HDSL approach and association with structured regions in viral RNA genomes. (**A**) Pairing probabilities for a representative region of SARS-CoV-2 as computed by *patteRNA* superimposed with the profiles after augmentation with detected hairpins. (**B**) Locally smoothed profiles from (A); local smoothing is achieved with a 5 nt rolling mean followed by a 15 nt rolling median. Smoothed P(paired) correlates weakly with structured elements of the region due to the unpaired nature of hairpin loops. Conversely, HDSL portrays a more relevant picture of local structuredness by accounting for the fact that some unpaired regions (i.e. hairpin loops) are within a stable structural context. (**C**) Secondary structure model of the region as proposed by Manfredonia *et al.* with hairpins detected by *patteRNA* indicated in blue boxes. (**D**) Discrimination between structured regions (low SHAPE, low Shannon entropy or 'LS/LSE') and non-structured regions via HDSL and smoothed P(paired) for five data replicates across SARS-CoV-2 (33) and HIV-1 (65). Both measures correlate strongly with structured regions, although the augmentation step underpinning HDSL drives a stronger difference between the considered regions. This is evidenced by larger median differences (Δ) and smaller *P*-values (Mann–Whitney *U* test) between the considered regions when using HDSL over smoothed P(paired).

patteRNA's hairpin detection scheme is imperfect, which can yield situations where an applied augmentation improperly models the local structure. For example, *patteRNA* may falsely predict a hairpin that does not stably fold. This would lead to an unjustified boost in HDSL for its nucleotides. This can be seen in Figure 4A, where between the first and second hairpin there are several plausible hairpins scored moderately ($0.5 < c$-score < 1.5), strong enough to be included in the augmentation step. The pairing profile of the area is subsequently augmented, although only slightly. The magnitude of the boost depends on the *c*-score of the hairpin, so moderately scored hairpins tend to drive only small boosts. This has the effect of restricting major augmentations to nucleotides in the most confidently detected hairpins (where *patteRNA* has the highest precision). Alternatively, a true hairpin may be missed, leading to non-augmentation at a stable structural element. In the absence of any detected hairpins for a region, HDSL amounts to a smoothed pairing probability profile. Thus, even when true hairpins are missed by the scoring phase, pairing probabilities 'fill in the gaps' and measure structuredness on their own. Nevertheless, such regions would remain susceptible to the issue associated with stable unpaired loops when quantifying structure via P(paired) alone.

HDSL verification

We used HDSL to measure structuredness across the SARS-CoV-2 and HIV-1 genomes as probed by Manfredonia *et al.* (33) and Siegfried *et al.* (65), respectively. We chose these two genomes as they are characterized by structured and unstructured regions as determined via low SHAPE and low Shannon entropy, referred to as LS/LSE regions. They also have high-quality probing data from different conditions and reagents as well as state-of-the-art structure models. In total, these data provide five profiles across two viral genomes, enabling a robust investigation of how HDSL and smoothed P(paired) correlate with structured regions.

Using both HDSL and smoothed P(paired), we calculated the difference in the medians of the measures between LS/LSE regions and non-LS/LSE regions (Δ) and computed *P*-values between the two distributions (Mann–Whitney *U* tests) to quantify the overall discriminatory power. The results of our analysis demonstrate that HDSL associates more strongly with structured regions than smoothed P(paired) alone (Figure 4D). LS/LSE regions unsurprisingly have higher average smoothed P(paired) than non-LS/LSE regions, as one criterion in determining the former was low SHAPE. The discrimination between LS/LSE and non-LS/LSE regions is increased, however, when utilizing HDSL. In all analyses, we observed that the difference between the median of the two groups was larger with HDSL and the corresponding *P*-value smaller. This demonstrates that the augmentation of pairing profiles with detected hairpins can improve discrimination between structured and non-structured regions.

We continued our analysis of the Manfredonia data by inspecting *in vivo* and *in vitro* HDSL profiles in more detail. First, we characterized the consistency of *patteRNA*'s detected hairpins with the structure model proposed by Manfredonia *et al.* We took the published structure model as ground-truth, searched for all predicted regular hairpins, and quantified the accuracy of *patteRNA* via the ROC curve (Figure 5A) and PRC (Figure 5B). Our results reveal a very strong agreement between detected and predicted hairpins, as evidenced by AUCs around 0.89 and APs of 0.70. Next, we inspected HDSL profiles around the 5′UTR and observed trends consistent with currently accepted structure models (see Figure 5C) (33,79–82). Namely, HDSL is high at known stable stem-loops, such as SL2, SL4, SL5A/C, SL7 and SL8. A weaker signal is found at SL6, which also shows differential structuredness between *in vitro* and *in vivo* data. Comparative analysis, *in vivo* RNA–RNA interactions (80), and multiple probing datasets (33,79) support the presence of this element. However, mutagenesis studies on a related coronavirus, murine coronavirus (MHV), demonstrated that disrupting this stem loop did not significantly affect virus viability (83). Given that SL6 is within ORF1ab, it is possible that the element is transient in nature. That said, NMR experiments concluded SL6 stably forms and additionally measured a significantly larger internal loop than was predicted with *in silico* structure models (82). The internal loop, also identified as a major binding site for the N protein, appears to be responsible for high reactivities and the observed differential structuredness of SL6 between *in vitro* and *in vivo* data. Similarly, for SL3, although comparative sequence analysis and NNTM-based folding with *in vitro* data suggest the presence of this stem-loop, *in vivo* data does not agree with its presence (33,79). NMR investigations concluded that the stability of the element is strongly influenced by ionic conditions (82), and studies on RNA–RNA interactions suggest that this stem loop is unfolded *in vivo* to facilitate genome cyclization, as the region is involved in a long-range interaction with the 3′UTR (80). As such, differential structuredness between *in vitro* and *in vivo* conditions is consistent with current understandings of the stem-loop element. Finally, we observe relatively low HDSL for SL5B, an element confirmed via RNA–RNA interactions (80) and NMR (82). NMR studies, however, suggest that the upper part of the stem is destabilized at physiological temperatures by the presence of SL5C. The presence of a bulge and high reactivities near the apical loop of SL5B subsequently result in attenuated HDSL observations around this element, as the structure scores poorly for the regular hairpin motifs considered by *patteRNA* when summarizing structuredness. Although a complete analysis of the SARS-CoV-2 genome is beyond the scope of this study, full HDSL profiles for the two conditions are included in Supplementary Figure S6.

Lastly, we investigated the association between Shannon entropy and the following: SHAPE reactivities, pairing probabilities from *patteRNA*, and HDSL (Supplementary Figure S7). Our results show that reactivity is loosely correlated with Shannon entropy, yet pairing probabilities correlate slightly better despite a sizable collection of nucleotides with low P(paired) and low entropy. However, HDSL shows an even stronger correlation, suggesting that it reflects some aspects of structural stability better than the former measures. Finally, our results on the SARS-CoV-2 genome indicate that HDSL profiles retain sufficient resolution to capitulate locations of specific structural elements (e.g. individual stem-loops in the 5′UTR), boding for the plausible use of our measure to assist in more detailed analyses of regions in addition to quantifying local structuredness.

The application of HDSL on these data allows for the unique opportunity to benchmark it against previously characterized transcripts with both structured and unstructured regions. In that context, we remark that HDSL was developed with the intention of assisting in global structure quantifications and comparisons rather than a tool for *de novo* detection of structured regions. Nevertheless, our results suggest it could also provide utility for *de novo* applications. In such cases, structured regions could be detected by defining criteria based on high HDSL that persists across long spans of nucleotides (e.g. over 50 nt). Structured elements of the SARS-CoV-2 and HIV-1 genomes are typically associated with long stretches of HDSL >0.8. We recommend thresholds around this value when seeking to identify structured regions. When quantifying changes in structure, however, the use of HDSL is more flexible. Depending on the specific application and degree of structure in the RNAs being studied, the magnitude of HDSL should be considered in addition to any relative changes in it across differing cellular conditions or logical transcript regions. Comparative analyses of HDSL are demonstrated in the following sections.

Trends in detected hairpins recapitulate known mRNA dynamics in *E. coli*

We analyzed the set of 197 mRNA transcripts (comprising 432 genes) in *E. coli* probed *in vitro*, *in vivo*, and *in vivo* + kasugamycin with SHAPE-MaP by Mustoe *et al.* (30). In addition to Mustoe *et al.*'s analysis, previous studies have demonstrated that mRNAs fold differentially in cells compared to *in vitro* (50,73,77,84). *In vivo* mRNAs have been observed to be less structured than their *in vitro* counterparts, with the magnitude of structural changes correlated with translation (31,85). These effects have been observed most strongly in the context of the 5′UTR and CDS of highly expressed genes. Conversely, structural changes have also been observed around the 3′UTR, but evidence demonstrating both a decrease (84) and increase (85) in structures has been published in the literature, possibly correlating to the degree of post-transcriptional regulation of transcript decay (85). We applied HDSL to Mustoe *et al.*'s data and investigated to what degree our measure reveals structural changes along mRNA transcripts in a prokaryotic organism like *E. coli*.

The results of our analysis are compiled in Figure 6. In Figure 6A, we compare averaged HDSL profiles over the 432 genes included in the study between *in vitro* and *in vivo* conditions. The averaged HDSL profiles are delineated into three groups: nucleotides near the start site (AUG $\pm$ 30 nt), nucleotides within the coding sequence (at least 31 nt downstream of AUG) and nucleotides in UTRs (5′UTR: 31–70 nt upstream of AUG; 3′UTR: first 40 nt after STOP). Our results demonstrate that, as expected, UTRs are generally the most structured regions of the transcripts. They also

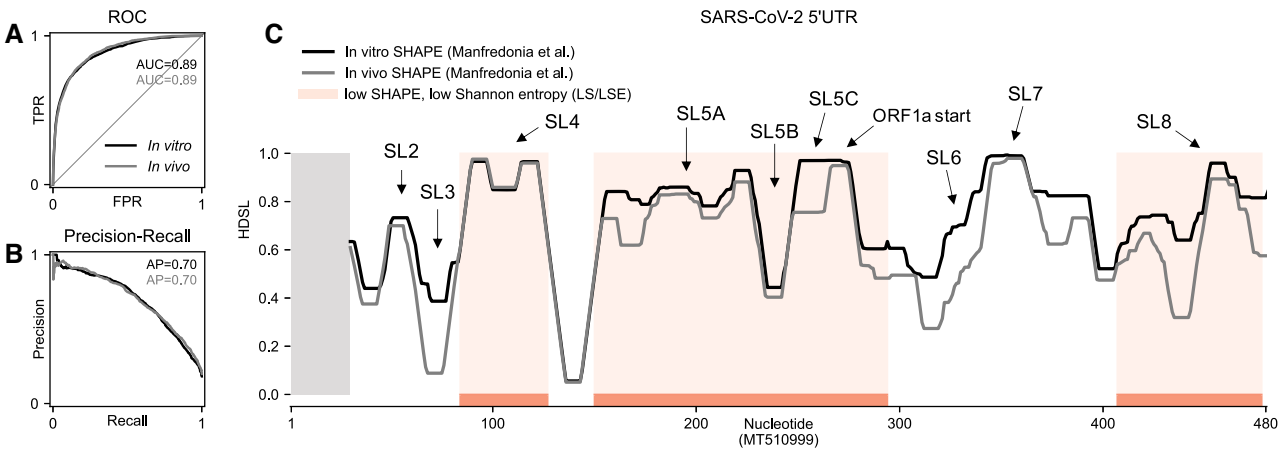

Figure 5. HDSL demonstrates correlated and differential structuredness between *in vitro* and *in vivo* SHAPE experiments on SARS-CoV-2 by Manfredonia *et al.* (33). (**A** and **B**) Receiver operating characteristic curves and precision-recall curves for *patteRNA*'s detected hairpins. Overall. *patteRNA* readily detects hairpins in the genome with moderate to strong precision. (**C**) HDSL profiles for the 5′UTR of SARS-CoV-2 *in vitro* and *in vivo* with low SHAPE, low Shannon entropy (LS/LSE) regions (called by Manfredonia *et al.*) indicated in red. Gray regions indicate no data.

show a strong intrinsic effect for mRNA to be relatively less structured around the start codon in both conditions. Moreover, *in vivo* data show that factors in this condition work to further unfold structures around the start site, as HDSL is significantly lower around the start codon *in vivo* than *in vitro*. Interestingly, we did not detect a strong signal for structures in coding sequences (AUG+31 nt onward) to be de-structured overall when accounting for the region around the start codon separately. It is worth noting that the reduction of HDSL around the start of coding sequences in the *in vivo* condition is only detected if the area around the start codon is delineated separately from the UTRs and CDS. Figure 6B shows the global HDSL trends in logical mRNA regions when (i) delineating start sites from UTRs and CDS and (ii) delineating based solely on CDS/UTR boundaries. Our results indicate that HDSL is significantly different between the conditions only in the region proximal to start codons. This contrasts with the original analysis by Mustoe *et al.* which did not consider start sites separately (i.e. considered only CDS versus non-CDS), concluding that coding sequences are relatively less structured in cells based on a slight increase in reactivities *in vivo* versus *in vitro* for nucleotides in CDS (demonstrated via reactivity scatterplot comparison of the two conditions and a fitted linear model slope >1). Our analysis suggests that global changes to reactivity profiles within CDS between conditions are not significant, yet effects specific to the start codon region are significant. These effects are likely partially responsible for previous inferences on *in vivo* structure dynamics. Notably, the specific relevance of structure around this region of mRNA transcripts has been observed and recognized as important in several other studies on organisms of varying genetic complexity (45,46,50,86).

To further substantiate the effects we observed, we checked the similarity of *patteRNA*'s detected hairpins for each pairwise comparison of the three conditions included in the original study. Ideally, in the absence of significant structural remodeling between two conditions, we expect to find the same hairpins in both. On the other hand, if two conditions are substantially different, we expect to see larger differences in the hairpins detected by *patteRNA*. Searching for the aforementioned set of regular hairpins (see *Hairpins Comprise a Significant Portion of Structural Elements*) and using a *c*-score threshold of 1 to indicate a 'detected' hairpin, we computed the fraction of hairpins reproducible in both conditions of each comparison (Figure 6C). We see that *in vivo* and *in vivo* + kasugamycin have the highest level of hairpin conservation (<10% of detected hairpins are not present in both conditions, meaning >90% similarity in detected hairpins). This high similarity serves as a basic quality control measure, as the *in vivo* + kasugamycin condition, although affected by changes to translation initiation, is nevertheless highly similar to the *in vivo* condition. On the contrary, comparing *in vivo* to *in vitro* data shows that 20% of detected hairpins are unique to one condition. The very high level of similarity between *in vivo* and *in vivo* + kasugamycin reaffirms that the differences observed in Figure 6A between *in vivo* and *in vitro* reflect real differential effects, rather than the impact of biological variation or artifacts from *patteRNA*'s imperfect hairpin detection scheme.

To further investigate the differences between the conditions around start codons, we visualized the condition-wise correlation of HDSL for all nucleotides within this region (Figure 6D). We detected a tendency in this area for the most structured regions *in vitro* to remain structured *in vivo* (see top right of distribution, which is tightly concentrated around the diagonal). The density of HDSL in Figure 6D does reveal a tendency for HDSL to be reduced in the *in vivo* condition, but mostly for regions with moderate HDSL *in vitro*. Thus, the overall de-structuring effect from Figure 6A appears to be driven by unfolding of moderately structured regions. Figure 6E compares the HDSL distribution between *in vitro* and *in vivo* at the adenosine residue of the start codon. There is a noticeable reduction in HDSL in the *in vivo* condition ($P < 1 \times 10^{-60}$, Wilcoxon signed-rank test), presumably driven by translation and possibly other cellular effects destabilizing mRNA structure, as discussed above. There is also a noticeable re-

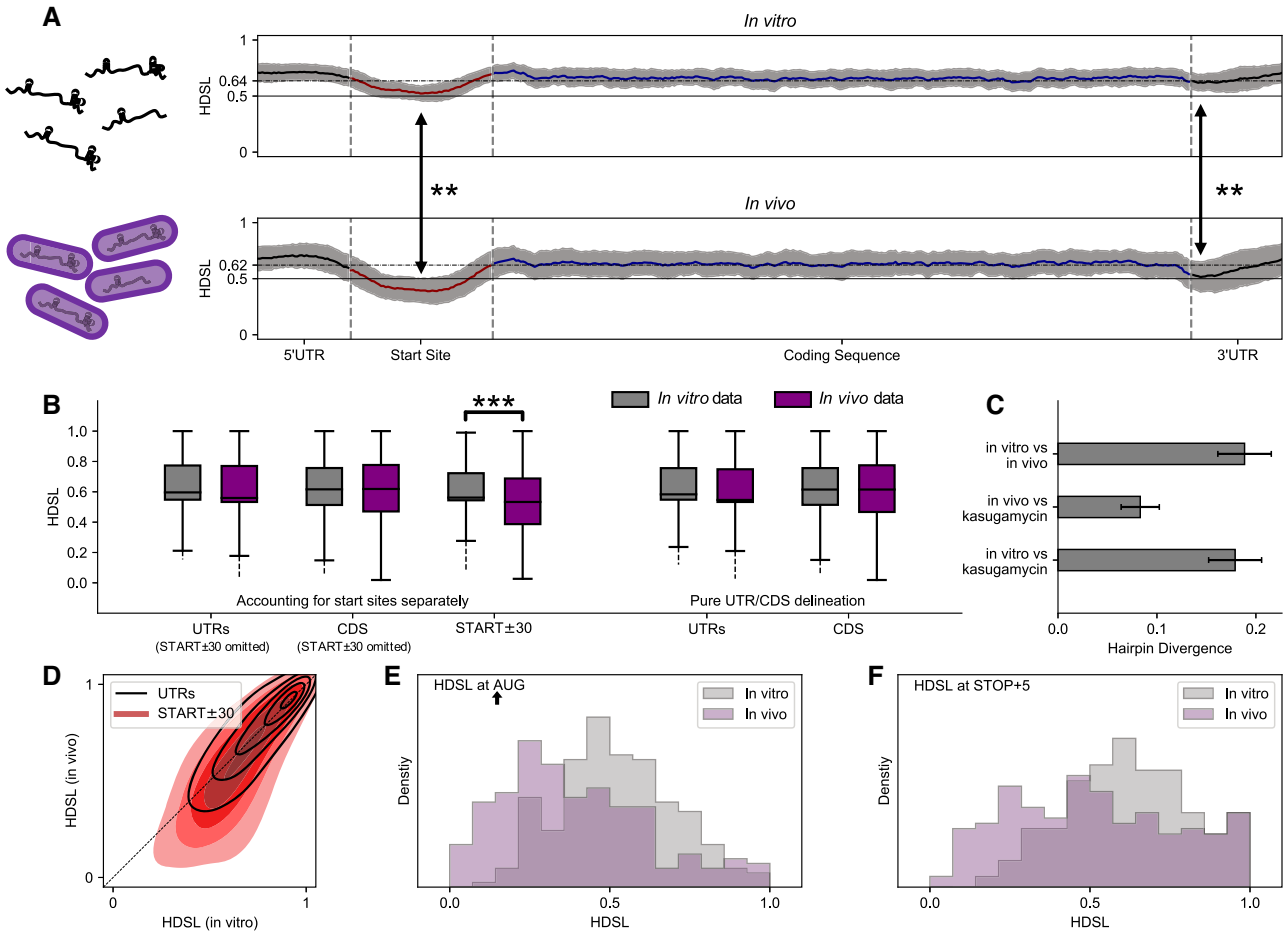

Figure 6. Hairpin-derived structure level (HDSL) demonstrates regional differences in structure changes between *in vivo* and *in vitro* structures for mRNA transcripts in *E. coli* (probed by Mustoe *et al.* (30)). (**A**) Averaged HDSL profiles across all genes ($N = 432$) for nucleotides around the start codon (±30 nt, red), within the coding sequence (AUG+31 to STOP), and 5'/3'UTRs (black). Gray area indicates the 99% CI of mean HDSL (Wald interval, see Materials and Methods). Dot-dashed lines indicate mean HDSL over all nucleotides in each condition. (**B**) HDSL trends between *in vitro* and *in vivo* conditions when delineating mRNA regions by UTRs and CDS (right) versus accounting for the region around the start codon separately (left). Delineating the region around start codons separately from CDS and UTRs reveals a signal occluded by the other delineation scheme. (**C**) Hairpin divergence (fraction of *patteRNA*-detected hairpins unique to one condition) for the three pairwise comparisons between *in vivo*, *in vitro*, and *in vivo* + kasugamycin conditions. Error bars represent the exact binomial (Clopper-Pearson) 99% CI. (**D**) 2D density plot of HDSL between the two conditions shown in (A) indicates a bias for weakly structured regions *in vitro* to become more unstructured *in vivo*. (**E**) Histograms of HDSL at the adenosine of start codons for both conditions in (A). (**F**) Histograms of HDSL at the fifth nucleotide after the STOP codon for both conditions in (A). ** indicates $P < 1 \times 10^{-60}$ (Wilcoxon signed-rank test); *** indicates $P < 1 \times 10^{-100}$ (Mann–Whitney U test).

duction in HDSL near the start of the 3′UTR (Figure 6F, $P < 1 \times 10^{-60}$, Wilcoxon signed-rank test), although this effect disappears on average for nucleotides farther away from the end of the coding sequence (see Figure 6A). Overall, our results demonstrate that HDSL can rapidly measure local structure and gives results consistent with prior analyses.

Finally, it is worth mentioning that similar results were also obtained by using smoothed P(paired) to assess structure without augmentation (see Supplementary Figure S8). However, HDSL highlights specific changes around the start codon more strongly than smoothed P(paired). Specifically, we found a very low level of augmentation in the *in vivo* condition around start codons, especially compared to a larger augmentation *in vitro* around this area. In other words, although both smoothed P(paired) and HDSL capitulated that there is a reduction in structure *in vivo* around the start codon of genes, the condition-wise effect was highlighted more strongly with HDSL. Quantitatively, this is observed as a significantly smaller P-value comparing HDSL around start codons ($P < 1 \times 10^{-100}$, Mann–Whitney U test) than the P-value when comparing with smoothed P(paired) ($P < 1 \times 10^{-55}$, Mann–Whitney U test).

RBPs Bind RNA at structured regions

Corley *et al.* (32) devised a novel experimental procedure called fSHAPE which can detect RNA nucleotides engaging in hydrogen bonding with RNA-binding proteins (RBPs). fSHAPE works by chemically probing RNA transcripts in the presence and absence of native binding factors, then quantifying the degree of modification change between the two conditions. Nucleotides bound by RBP would presumably be more reactive in the absence of binding fac-

tors, which translates to a high fSHAPE score. Integrating fSHAPE information with standard reactivity profiles therefore allows one to examine the structural context of RBP binding sites. In this regard, Corley *et al.* performed icSHAPE in tandem with fSHAPE to perform such analyses transcriptome-wide on human cell lines (K562, HepG2 and HeLa). Their work showed that nucleotides with high fSHAPE scores tend to fall in areas with relatively low Shannon entropy when compared to the regions flanking them, allowing them to conclude that RBP tend to associate with RNA in the general context of stable structured regions.

We sought to use HDSL to address the same question, namely, is there a structural context characteristic to RBP binding? To this end, we processed their icSHAPE data with *patteRNA*, mined for regular hairpins, and computed HDSL profiles. We first investigated what association exists, if any, between high fSHAPE nucleotides and pairing probabilities as computed by *patteRNA*'s DOM-HMM. Simply put, we found that nucleotides with high fSHAPE (fSHAPE ≥ 2) are almost unanimously unpaired (Figure 7A), while nucleotides with low fSHAPE follow a distribution encompassing both states yet biased towards paired states ($P < 10^{-307}$ for all low/high fSHAPE comparisons in Figure 7A, Mann–Whitney U test). The association of high fSHAPE with unpaired nucleotides recapitulates what Corley *et al.* demonstrated with pairing probabilities computed via partition function approaches. It also verifies that *patteRNA* appropriately models the reactivity data and reiterates that fSHAPE is designed to detect RBP footprints predominantly at unpaired nucleotides.

However, despite the ubiquitous accessibility observed at single nucleotides with high fSHAPE, when one expands the considered context to the nucleotides' local neighborhood (i.e. via smoothed P(paired) or HDSL analysis), one observes evidence suggesting a structured context of RBP footprints. This is weakly demonstrated with smoothed P(paired) (Figure 7B), which is marginally higher at high fSHAPE nucleotides in 3 of 4 comparisons performed. In contrast, the structural context of RBP interactions is more readily seen from the perspective of HDSL (Figure 7C). With HDSL, we observe significantly more local structure around nucleotides with high fSHAPE compared to nucleotides with low fSHAPE (Figure 7C). This result is consistent with results from NNTM analyses performed by Corley *et al.*, whose interpretation again depended on the computation of Shannon entropy. Our results were achieved without any folding steps and are more statistically significant ($P < 10^{-307}$ for all low/high fSHAPE comparisons in Figure 7C, Mann–Whitney U test) than originally demonstrated. They were also generated orders of magnitude faster than a comparable NNTM approach, as we will show next. We note that current approaches for summarizing local structuredness from SP data alone, specifically local median reactivity, are generally insufficient for reaching this conclusion (see Supplementary Figure S9). This highlights the capability of our method to extract more information from big SP datasets without relying on the additional assumptions and computational overhead of thermodynamic modeling.

patteRNA processes large data rapidly

An especially appealing property of *patteRNA* is its ability to process big datasets rapidly. To demonstrate its speed in the context of existing methods, we timed our analyses and compared to partition function-based assessment of structure. To this end, we processed the Weeks set, SARS-CoV-2 genome, Mustoe data, and Corley data with three sliding-window partition function analyses of varying computational overhead: partition function calculations with windows of length 3000 nt, spaced 300 nt apart; windows of length 2000 nt, spaced 150 nt apart, and windows of length 150 nt, spaced 15 nt apart. The results of the benchmarks are in Supplementary Figure S10. We observe that *patteRNA* is orders of magnitude faster than sliding-window partition function analysis for massive datasets (e.g., SP data on human transcriptomes). Specifically, *patteRNA* processed the largest dataset included in this study, the Corley data, in <1 h when using a single-threaded implementation (compared to roughly 1 and 7 days for partition function calculation via 150 nt and 2000 nt windows, respectively; 3000 nt window calculations on the Corley data were not performed as they could not be completed in reasonable timeframe). Additionally, our method is natively parallelized, and benchmarks using 12 threads allow *patteRNA* to process such data in <10 min. Analogous parallelization of partition function-based approaches on large batches of RNA transcripts is relatively simple in theory, but not natively provided 'out-of-the-box' for ViennaRNA (meaning it's up to the user to program their own parallelized calls to the relevant methods). An alternative RNA folding package, RNAstructure (21), does provide scalable parallelization out-of-the-box, but the core folding implementation is about one to two orders of magnitude slower than ViennaRNA. The method was therefore not included in our comparison.

We also compared our method to *RNALfold* (87), an optimized routine within the ViennaRNA package designed to rapidly scan long RNAs for locally stable structural elements. As expected, we found that this method is capable of processing large data significantly faster than the sliding-window partition function approaches, yet it is nevertheless outpaced by *patteRNA*. Moreover, this method only returns structural elements with sufficiently low free energy ('significantly low' energies judged via an SVM) and, to the best of our knowledge, has not been well-benchmarked against reference structures. Furthermore, *RNALfold* does not attempt to integrate its results to summarize local structuredness, which is key to the type of comparative analyses performed in this study and a central theme of a broad range of recent SP-based studies (32,47,84,85). Nevertheless, this method arrives at a more specific and comprehensive description of local structures (i.e., it can *de-novo* identify stems with bulges and internal loops), whereas *patteRNA*'s analyses here focus specifically on hairpin elements. We note that the incorporation of such local folding routines would likely improve the efficacy of future methods aiming to summarize local structure in large SP datasets, and our results show promising evidence that localized folding can be incorporated without major sacrifices to computational speed.

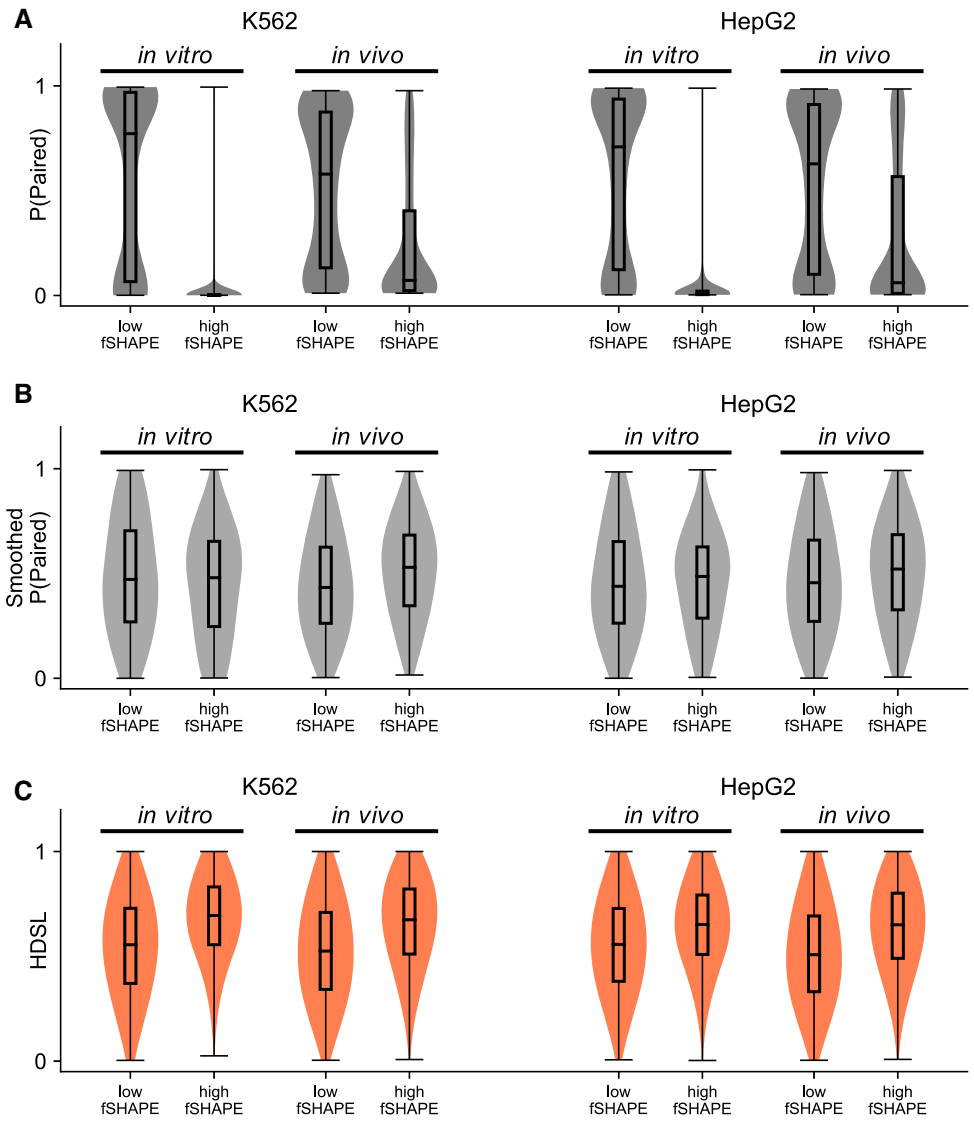

Figure 7. *patteRNA* demonstrates a strong association between RNA structure and RBP binding sites in human cell lines probed as by Corley *et al.* (32). (**A**) Paired probability box-violin plots (determined from icSHAPE reactivities via *patteRNA*'s DOM-HMM) for nucleotides with low fSHAPE (fSHAPE < 0) and high fSHAPE (fSHAPE ≥ 2). Within each of the two cell lines, K562 and HepG2, results are presented for both *in vitro* and *in vivo* SHAPE data. (**B**) Smoothed P(paired) box-violin plots for nucleotides under the same conditions as (A). (**C**) HDSL box-violin plots for nucleotides under the same conditions as (A). Although reactivities indicate that nucleotides with evidence of RBP binding (i.e., nucleotides with high fSHAPE) are remarkedly accessible and therefore likely unpaired, HDSL demonstrates that these reactive nucleotides more frequently occur in the general context of structured regions when compared to nucleotides with low fSHAPE. Similar results are observed when utilizing smoothed P(paired) to assess structuredness, though the results achieved with HDSL portray a clearer association between RBP footprints (high fSHAPE score) and structured regions. $P < 10^{-307}$ for all low/high fSHAPE comparisons in panels (A) and (C) (Mann–Whitney *U* test).

DISCUSSION

RNA structure probing experiments are rapidly evolving in terms of their design, scale and quality. This evolution is accompanied by a need for versatile and scalable methods capable of extracting information from diverse and massive SP data. *patteRNA* is one such tool which was developed to rapidly extract insights from such data. Here, we have demonstrated reformulation of the *patteRNA* framework which increases its speed, adaptability and precision, enabling it to scale well to data containing millions or billions of nucleotides. Moreover, we have shown that RNA structure can be rapidly quantified and compared in various contexts by detecting the signatures of hairpin elements. Our work expands the repertoire of analyses which *patteRNA* is capable of and demonstrates the power of simpler schemes when interpreting reactivity information. As seen with our benchmarks using a DOM approach, relatively low-resolution discretization schemes (akin to those used to highlight low/medium/high reactivities when visualizing SP data) are valuable when quantifying and mining motifs. We also demonstrated that structuredness quantifications can benefit from an assessment of the locations of stable hairpin elements. HDSL correlated strongly with

structured regions and recapitulated structure trends in RNA genomes and mRNA transcripts. We also found that *patteRNA*'s pairing probabilities alone, when judiciously smoothed, can be a useful measure of structuredness, and for this reason *patteRNA* gives users the option to compute smoothed P(paired) or HDSL. Although we showed that HDSL better highlights certain structure trends than smoothed P(paired), we remark that smoothed P(paired) does not depend on a hairpin search and therefore can be computed even more rapidly than HDSL. As such, some users may find this option suitable in situations where data is large or where a more interpretable measure of base pairing is desired.

In the context of RNA structure determination, we note that *patteRNA* is not envisioned as a competing method or replacement to traditional NNTM-based approaches. Rather, we view the method as a tool to be used in tandem to RNA folding. As seen in Figure 3, NNTM-based ensemble methods provide a far more accurate prediction of specific structures and are capable of assessing the entire structure landscape including bulges, internal loops and internal stems. The analyses via *patteRNA* shown here, on the other hand, intentionally compromise on the type of structures considered in the analysis in order to maximize the speed and scalability of the approach. This is evidenced by the relatively lower sensitivity of our method when compared to NNTM-based partition function analyses (Figure 3B). It's worth noting, however, that HDSL handles the low sensitivity of hairpin detection by utilizing posterior pairing probabilities to quantify structure in regions where no highly scored hairpins are found. In other words, structured regions which house no detected hairpins are still likely to see high HDSL assuming local reactivities are moderately low. It's also worth mentioning that, although overall sensitivity on the representative set of hairpins benchmarked was relatively lower than NNTM-based ensemble approaches, benchmarks for individual motifs (Supplementary Figure S5) reveal that *patteRNA*'s *c*-scores are capable of matching and outperforming partition function analyses for hairpin motifs with longer stems. In summary, although HDSL considers a partial landscape of detected hairpins as provided by *c*-scores, the formulation is driven primarily by the most confident hairpin predictions, resulting in a measure of structure significantly more correlated to Shannon entropy than local reactivities or pairing probabilities alone (Supplementary Figure S7). Nevertheless, the sensitivity of hairpin detections underpinning the method leaves room for improvement, for example, by combining simple thermodynamic assessments of local structure (88). As a consequence of these compromises, *patteRNA* is most useful when assessing structure properties in large-scale data. For instance, as we demonstrated, it could be utilized to quantify macroscopic structural trends related to specific regions, or it could be used to identify regions of RNA which see differential structuredness associated to some factor, which might then be followed by more intensive RNA folding approaches (e.g. partition function computation). In this way, *patteRNA* helps mitigate the computational limitations of such methods, especially for those who do not have advanced computing hardware at their disposal. Finally, although analyses in this study generally focus on using *patteRNA* to derive information on structuredness via hairpins, the method itself is fundamentally a versatile structure-mining algorithm which has been demonstrated to effectively search for putative functional motifs across in transcriptome-wide data (61).

Our analysis of the SARS-CoV-2 5′UTR is distinguished from the others by a comparison of HDSL with specific structures that have been validated in a plethora of ways, including NMR spectroscopy (82). We remarked on a great correspondence of HDSL peaks and stable structural elements, indicating that HDSL captures more than just local structure—it retains information on specific motifs with high resolution. This observation is important in the context of our analysis of Corley *et al.*'s fSHAPE data. Namely, the increase in HDSL at sites with high fSHAPE (Figure 7C) suggests the possibility that RBP frequently associate not only in the context of stable structured regions, but specifically in the context of hairpin-like elements. RBP which recognize sequence motifs in hairpin-loops have previously been identified (89,90), but our results demonstrate the plausibility that the association between hairpin elements and RBP is more prevalent than previously thought. This is not entirely unexpected, as RBP are known to bind both dsRNA and ssRNA in a manner that correlates with the structure of the protein (91). Moreover, RBP binding ssRNA are observed to associate at unpaired bases stemming from RNA helix irregularities (e.g. bulges and internal loops) (92), also placing them in the context of hairpin elements. Recent studies have further documented that structured RNAs interact with a larger number of proteins than less structured RNAs (91). Our result further strengthens the utility of *patteRNA* in mining biologically relevant structures transcriptome-wide.

Looking ahead to future development of rapid analysis of SP data, *patteRNA* is well-suited to adapt to evolving probing technologies and datasets. That being said, its current implementation does come with several limitations. First, motif mining depends on the definition of specific secondary structures, which limits its application to situations where a specific structure or small collection of similar structures can be defined. For motifs like hairpins, this means that considering situations where a bulge or internal loop may or may not be present complicates analyses due to the combinatorial explosion of unique secondary structures needed to define all possible hairpin architectures through loop size, bulge size and bulge position. *patteRNA* is already capable of exhaustively mining such motifs, but such analyses come at the cost of significant computational overhead, generally working against the utility of the method. A more efficient approach for motif mining which naturally considers alternative similar structures within a region could theoretically address some parts of this limitation. Secondly, although the circumvention of RNA folding enables rapid computational analyses, it also handicaps the accuracy of the approach, as the energetic favorability of sequences within stems and loops is ignored. The incorporation of an optimized local folding routine could likely assist in this regard, although the coupling of such models into a statistical model like *patteRNA* is non-trivial. Nevertheless, methods like *RNALfold* (87) bode for the potential incorporation of NNTM-derived information without sacrific-

ing on speed and scalability. Regardless of these limitations, however, *patteRNA* remains a viable computational method for the rapid assessment and quantification of structural trends in the largest SP datasets.

DATA AVAILABILITY

The latest version of *patteRNA*, version 2.0, was used for all analyses in this study. *patteRNA* is an open-source Python 3 module and is freely available at www.github.com/AviranLab/patteRNA under the BSD-2 license. Python scripts for generating simulated datasets, computing statistical benchmarks (e.g. ROC and PRC), and post-processing of HDSL profiles related to genes in the Mustoe data are available in Supplementary File S1. The original datasets used in this study are all publicly available from the indicated references in Table 1.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank the reviewers and Associate Editor for their attentive and constructive feedback on the manuscript.

FUNDING

Internal funds.
Conflict of interest statement. None declared.

REFERENCES

1. Ganser,L.R., Kelly,M.L., Herschlag,D. and Al-Hashimi,H.M. (2019) The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.*, **20**, 474–489.
2. Mustoe,A.M., Brooks,C.L. and Al-Hashimi,H.M. (2014) Hierarchy of RNA functional dynamics. *Annu. Rev. Biochem.*, **83**, 441–466.
3. Fica,S.M. and Nagai,K. (2017) Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nat. Struct. Mol. Biol.*, **24**, 791–799.
4. Dallaire,P., Tan,H., Szulwach,K., Ma,C., Jin,P. and Major,F. (2016) Structural dynamics control the MicroRNA maturation pathway. *Nucleic Acids Res.*, **44**, 9956–9964.
5. Serganov,A. and Patel,D.J. (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.*, **8**, 776–790.
6. Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, **12**, 861–874.
7. Spitale,R.C., Crisalli,P., Flynn,R.A., Torre,E.A., Kool,E.T. and Chang,H.Y. (2013) RNA SHAPE analysis in living cells. *Nat. Chem. Biol.*, **9**, 18–20.
8. Smola,M.J., Rice,G.M., Busan,S., Siegfried,N.A. and Weeks,K.M. (2015) Selective 2' hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc.*, **10**, 1643–1669.
9. Zhang,K., Li,S., Kappel,K., Pintilie,G., Su,Z., Mou,T.-C., Schmid,M.F., Das,R. and Chiu,W. (2019) Cryo-EM structure of a 40 kDa SAM-IV riboswitch RNA at 3.7 Å resolution. *Nat. Commun.*, **10**, 5511.
10. Wang,P.Y., Sexton,A.N., Culligan,W.J. and Simon,M.D. (2019) Carbodiimide reagents for the chemical probing of RNA structure in cells. *RNA*, **25**, 135–146.
11. 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). *Proc. Natl. Acad. Sci. USA*, **108**, 11063–11068.
12. 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.
13. Ziv,O., Gabryelska,M.M., Lun,A.T.L., Gebert,L.F.R., Sheu-Gruttadauria,J., Meredith,L.W., Liu,Z.Y., Kwok,C.K., Qin,C.F., MacRae,I.J. *et al.* (2018) COMRADES determines in vivo RNA structures and interactions. *Nat. Methods*, **15**, 785–788.
14. Ghut,J., Aw,A., Lim,S.W., Wang,J.X., Lambert,F.R.P., Tan,W.T., Shen,Y., Zhang,Y., Kaewsapsak,P., Li,C. *et al.* (2020) With nanopore long reads. *Nat. Biotechnol.*, **39**, 336–346.
15. Cheng,C.Y., Kladwang,W., Yesselman,J. and Das,R. (2017) RNA structure inference through chemical mapping after accidental or intentional mutations. *Proc. Natl. Acad. Sci. USA*, **114**, 9876–9881.
16. Holbrook,S.R. and Kim,S.-H. (1997) RNA crystallography. *Biopolymers*, **44**, 3–21.
17. Fürtig,B., Richter,C., Wöhnert,J. and Schwalbe,H. (2003) NMR spectroscopy of RNA. *ChemBioChem*, **4**, 936–962.
18. Hajdin,C.E., Bellaousov,S., Huggins,W., Leonard,C.W., Mathews,D.H. and Weeks,K.M. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. USA*, **110**, 5498–5503.
19. Hjelm,B.E., Rollins,B., Morgan,L., Sequeira,A., Mamdani,F., Pereira,F., Damas,J., Webb,M.G., Weber,M.D., Schatzberg,A.F. *et al.* (2019) Splice-Break: exploiting an RNA-seq splice junction algorithm to discover mitochondrial DNA deletion breakpoints and analyses of psychiatric disorders. *Nucleic Acids Res.*, **47**, 26.
20. Lavender,C.A., Lorenz,R., Zhang,G., Tamayo,R., Hofacker,I.L. and Weeks,K.M. (2015) Model-Free RNA sequence and structure alignment informed by SHAPE probing reveals a conserved alternate secondary structure for 16S rRNA. *PLOS Comput. Biol.*, **11**, e1004126.
21. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinform.*, **11**, 129.
22. 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.
23. Miao,Z. and Westhof,E. (2017) RNA structure: advances and assessment of 3D structure prediction. *Annu. Rev. Biophys.*, **46**, 483–503.
24. Ponce-Salvatierra,A., Astha, Merdas,K., Nithin,C., Ghosh,P., Mukherjee,S. and Bujnicki,J.M. (2019) Computational modeling of RNA 3D structure based on experimental data. *Biosci. Rep.*, **39**, BSR20180430.
25. Li,H. and Aviran,S. (2018) Statistical modeling of RNA structure profiling experiments enables parsimonious reconstruction of structure landscapes. *Nat. Commun.*, **9**, 606.
26. Lorenz,R., Luntzer,D., Hofacker,I.L., Stadler,P.F. and Wolfinger,M.T. (2016) SHAPE directed RNA folding. *Bioinformatics*, **32**, 145–147.
27. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
28. Dowell,R.D. and Eddy,S.R. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinform.*, **7**, 400.
29. Pirakitikulr,N., Kohlway,A., Lindenbach,B.D. and Pyle,A.M. (2016) The coding region of the HCV genome contains a network of regulatory RNA structures. *Mol. Cell*, **62**, 111–120.
30. Mustoe,A.M., Busan,S., Rice,G.M., Hajdin,C.E., Peterson,B.K., Ruda,V.M., Kubica,N., Nutiu,R., Baryza,J.L. and Weeks,K.M. (2018) Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing. *Cell*, **173**, 181–195.
31. Kramer,M.C. and Gregory,B.D. (2018) Does RNA secondary structure drive translation or vice versa? *Nat. Struct. Mol. Biol.*, **25**, 641–643.
32. Corley,M., Flynn,R.A., Lee,B., Blue,S.M., Chang,H.Y. and Yeo,G.W. (2020) Footprinting SHAPE-eCLIP reveals Transcriptome-wide hydrogen bonds at RNA-Protein interfaces. *Mol. Cell*, **80**, 903–914.
33. Manfredonia,I., Nithin,C., Ponce-Salvatierra,A., Ghosh,P., Wirecki,T.K., Marinus,T., Ogando,N.S., Snijder,E.J., van Hemert,M.J., Bujnicki,J.M. *et al.* (2020) Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.*, **48**, 12436–12452.

34. Ochsenreiter,R., Hofacker,I.L. and Wolfinger,M.T. (2019) Functional RNA structures in the 3′UTR of tick-borne, insect-specific and no-known-vector flaviviruses. *Viruses*, **11**, 298.
35. Halvorsen,M., Martin,J.S., Broadaway,S. and Laederach,A. (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.*, **6**, e1001074.
36. Zinshteyn,B., Chan,D., England,W., Feng,C., Green,R. and Spitale,R.C. (2019) Assaying RNA structure with LASER-Seq. *Nucleic Acids Res.*, **47**, 43–55.
37. Weng,X., Gong,J., Chen,Y., Wu,T., Wang,F., Yang,S., Yuan,Y., Luo,G., Chen,K., Hu,L. *et al.* (2020) Keth-seq for transcriptome-wide RNA structure mapping. *Nat. Chem. Biol.*, **16**, 489–492.
38. Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.*, **20**, 295–304.
39. Mailler,E., Paillart,J.C., Marquet,R., Smyth,R.P. and Vivet-Boudou,V. (2019) The evolution of RNA structural probing methods: from gels to next-generation sequencing. *Wiley Interdiscip. Rev. RNA*, **10**, e1518.
40. Aviran,S., Trapnell,C., Lucks,J.B., Mortimer,S.A., Luo,S., Schroth,G.P., Doudna,J.A., Arkin,A.P. and Pachter,L. (2011) Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl. Acad. Sci. USA*, **108**, 11069–11074.
41. Deigan,K.E., Li,T.W., Mathews,D.H. and Weeks,K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. USA*, **106**, 97–102.
42. Underwood,J.G., Uzilov,A. V., Katzman,S., Onodera,C.S., Mainzer,J.E., Mathews,D.H., Lowe,T.M., Salama,S.R. and Haussler,D. (2010) FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.
43. 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.
44. Spitale,R.C., Flynn,R.A., Zhang,Q.C., Crisalli,P., Lee,B., Jung,J.-W., Kuchelmeister,H.Y., Batista,P.J., Torre,E.A., Kool,E.T. *et al.* (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519**, 486–490.
45. Ding,Y., Tang,Y., Kwok,C.K., Zhang,Y., Bevilacqua,P.C. and Assmann,S.M. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**, 696–700.
46. Wan,Y., Qu,K., Zhang,Q.C., Flynn,R.A., Manor,O., Ouyang,Z., Zhang,J., Spitale,R.C., Snyder,M.P., Segal,E. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.
47. Choudhary,K., Deng,F. and Aviran,S. (2017) Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. *Quant. Biol.*, **5**, 3–24.
48. Saha,K., England,W., Fernandez,M.M., Biswas,T., Spitale,R.C. and Ghosh,G. (2020) Structural disruption of exonic stem-loops immediately upstream of the intron regulates mammalian splicing. *Nucleic Acids Res.*, **48**, 6294–6309.
49. Yang,M., Woolfenden,H.C., Zhang,Y., Fang,X., Liu,Q., Vigh,M.L., Cheema,J., Yang,X., Norris,M., Yu,S. *et al.* (2020) Intact RNA structurome reveals mRNA structure-mediated regulation of miRNA cleavage in vivo. *Nucleic Acids Res.*, **48**, 8767–8781.
50. Mauger,D.M., Joseph Cabral,B., 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. USA*, **116**, 24075–24083.
51. Lai,Y.-H., Choudhary,K., Cloutier,S.C., Xing,Z., Aviran,S. and Tran,E.J. (2019) Genome-Wide discovery of DEAD-Box RNA helicase targets reveals RNA structural remodeling in transcription termination. *Genetics*, **212**, 153–174.
52. Guenther,U.-P., Weinberg,D.E., Zubradt,M.M., Tedeschi,F.A., Stawicki,B.N., Zagore,L.L., Brar,G.A., Licatalosi,D.D., Bartel,D.P., Weissman,J.S. *et al.* (2018) The helicase ded1p controls use of near-cognate translation initiation codons in 5' UTRs. *Nature*, **559**, 130–134.
53. Waldron,J.A., Tack,D.C., Ritchey,L.E., Gillen,S.L., Wilczynska,A., Turro,E., Bevilacqua,P.C., Assmann,S.M., Bushell,M. and Quesne,J. Le (2019) mRNA structural elements immediately upstream of the start codon dictate dependence upon eIF4A helicase activity. *Genome Biol.*, **20**, 300.
54. Lee,Y.J., Wang,Q. and Rio,D.C. (2018) Coordinate regulation of alternative pre-mRNA splicing events by the human RNA chaperone proteins hnRNPA1 and DDX5. *Genes Dev.*, **32**, 1060–1074.
55. Twittenhoff,C., Brandenburg,V.B., Righetti,F., Nuss,A.M., Mosig,A., Dersch,P. and Narberhaus,F. (2020) Lead-seq: Transcriptome-wide structure probing in vivo using lead(II) ions. *Nucleic Acids Res.*, **48**, E71–E71.
56. Nussinov,R. and Jacobson,A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA*, **77**, 6309–6313.
57. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
58. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinform.*, **5**, 140.
59. Choudhary,K., Lai,Y.-H., Tran,E.J. and Aviran,S. (2019) dStruct: identifying differentially reactive regions from RNA structurome profiling data. *Genome Biol.*, **20**, 40.
60. Marangio,P., Law,K.Y.T., Sanguinetti,G. and Granneman,S. (2021) Differential BUM-HMM: a robust statistical modelling approach for detecting RNA flexibility changes in high-throughput structure probing data. *Genome Biol.*, **22**, 165.
61. Ledda,M. and Aviran,S. (2018) PATTERNA: Transcriptome-wide search for functional RNA elements via structural data signatures. *Genome Biol.*, **19**, 28.
62. Radecki,P., Ledda,M. and Aviran,S. (2018) Automated recognition of RNA structure motifs by their SHAPE data signatures. *Genes (Basel)*, **9**, 300.
63. Andronescu,M., Bereg,V., Hoos,H.H. and Condon,A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinform.*, **9**, 340.
64. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z. *et al.* (2020) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, **49**, D192–D200.
65. Siegfried,N.A., Busan,S., Rice,G.M., Nelson,J.A.E. and Weeks,K.M. (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*, **11**, 959–965.
66. Rabiner,L.R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
67. Eddy,S.R. (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu. Rev. Biophys.*, **43**, 433–456.
68. Busan,S., Weidmann,C.A., Sengupta,A. and Weeks,K.M. (2019) Guidelines for SHAPE reagent choice and detection strategy for RNA structure probing studies. *Biochemistry*, **58**, 2655–2664.
69. Marinus,T., Fessler,A.B., Ogle,C.A. and Incarnato,D. (2021) A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic Acids Res.*, **49**, e34.
70. Sükösd,Z., Swenson,M.S., Kjems,J. and Heitsch,C.E. (2013) Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Res.*, **41**, 2807–2816.
71. Sexton,A.N., Wang,P.Y., Rutenberg-Schoenberg,M. and Simon,M.D. (2017) Interpreting reverse transcriptase termination and mutation events for greater insight into the chemical probing of RNA. *Biochemistry*, **56**, 4713–4721.
72. Mauger,D.M., Golden,M., Yamane,D., Williford,S., Lemon,S.M., Martin,D.P. and Weeks,K.M. (2015) Functionally conserved architecture of hepatitis c virus RNA genomes. *Proc. Natl. Acad. Sci. USA*, **112**, 3692–3697.
73. Simon,L.M., Morandi,E., Luganini,A., Gribaudo,G., Martinez-Sobrido,L., Turner,D.H., Oliviero,S. and Incarnato,D. (2019) In vivo analysis of influenza a mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Res.*, **47**, 7003–7017.
74. Ramanouskaya,T. V. and Grinev,V. V. (2017) The determinants of alternative RNA splicing in human cells. *Mol. Genet. Genomics*, **292**, 1175–1195.
75. Hiller,M., Zhang,Z., Backofen,R. and Stamm,S. (2007) Pre-mRNA secondary structures influence exon recognition. *PLoS Genet.*, **3**, 2147–2155.

76. Melnick,M., Gonzales,P., Cabral,J., Allen,M.A., Dowell,R.D. and Link,C.D. (2019) Heat shock in c. elegans induces downstream of gene transcription and accumulation of double-stranded RNA. *PLoS One*, **14**, e0206715.
77. Gawroński,P., Pałac,A. and Scharff,L.B. (2020) Secondary structure of chloroplast mRNAs in vivo and in vitro. *Plants*, **9**, 323.
78. Deng,F., Ledda,M., Vaziri,S. and Aviran,S. (2016) Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA*, **22**, 1109–1119.
79. Huston,N.C., Wan,H., Strine,M.S., de Cesaris Araujo Tavares,R., Wilen,C.B. and Pyle,A.M. (2021) Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol. Cell*, **81**, 584–598.
80. Ziv,O., Price,J., Shalamova,L., Kamenova,T., Goodfellow,I., Weber,F. and Miska,E.A. (2020) The Short- and Long-Range RNA-RNA interactome of SARS-CoV-2. *Mol. Cell*, **80**, 1067–1077.
81. Manfredonia,I. and Incarnato,D. (2020) Structure and regulation of coronavirus genomes: state-of-the-art and novel insights from SARS-CoV-2 studies. *Biochem. Soc. Trans.*, **0**, 1–12.
82. Wacker,A., Weigand,J.E., Akabayov,S.R., Altincekic,N., Bains,J.K., Banijamali,E., Binas,O., Castillo-Martinez,J., Cetiner,E., Ceylan,B. *et al.* (2020) Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic Acids Res.*, **48**, 12415–12435.
83. Yang,D., Liu,P., Wudeck,E. V., Giedroc,D.P. and Leibowitz,J.L. (2015) Shape analysis of the rna secondary structure of the mouse hepatitis virus 5' untranslated region and n-terminal nsp1 coding sequences. *Virology*, **475**, 15–27.
84. 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.
85. Beaudoin,J.D., Novoa,E.M., Vejnar,C.E., Yartseva,V., Takacs,C.M., Kellis,M. and Giraldez,A.J. (2018) Analyses of mRNA structure dynamics identify embryonic gene regulatory programs. *Nat. Struct. Mol. Biol.*, **25**, 677–686.
86. Cambray,G., Guimaraes,J.C. and Arkin,A.P. (2018) Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli. *Nat. Biotechnol.*, **36**, 1005.
87. Hofacker,I.L., Priwitzer,B. and Stadler,P.F. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.
88. Radecki,P., Uppuluri,R., Deshpande,K. and Aviran,S. (2021) Accurate detection of RNA stem-loops in structurome data reveals widespread association with protein binding sites. *RNA Biol.*, https://doi.org/10.1080/15476286.2021.1971382.
89. Aviv,T., Lin,Z., Ben-Ari,G., Smibert,C.A. and Sicheri,F. (2006) Sequence-specific recognition of RNA hairpins by the SAM domain of vts1p. *Nat. Struct. Mol. Biol.*, **13**, 168–176.
90. Jolma,A., Zhang,J., Mondragón,E., Morgunova,E., Kivioja,T., Laverty,K.U., Yin,Y., Zhu,F., Bourenkov,G., Morris,Q. *et al.* (2020) Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res.*, **30**, 962–973.
91. Groot,N.S. de, Armaos,A., Graña-Montes,R., Alriquet,M., Calloni,G., Vabulas,R.M., Tartaglia,G.G., Sanchez de Groot,N., Armaos,A., Graña-Montes,R. *et al.* (2019) RNA structure drives interaction with proteins. *Nat. Commun.*, **10**, 3246.
92. Jones,S. (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.