

# LinearTurboFold: Linear-Time Global Prediction of Conserved Structures for RNA Homologs with Applications to SARS-CoV-2

Sizhen Li<sup>a</sup>, He Zhang<sup>b,a</sup>, Liang Zhang<sup>b,a</sup>, Kaibo Liu<sup>b</sup>, Boxiang Liu<sup>b</sup>, David H. Mathews<sup>d,\*</sup>, and Liang Huang<sup>a,b,\*</sup>

<sup>a</sup>School of Electrical Engineering & Computer Science, Oregon State University, Corvallis, OR; <sup>b</sup>Baidu Research, Sunnyvale, CA; <sup>d</sup>Department of Biochemistry & Biophysics, Center for RNA Biology, and Department of Biostatistics & Computational Biology, University of Rochester Medical Center, Rochester, NY

**The constant emergence of COVID-19 variants reduces the effectiveness of existing vaccines and test kits. Therefore, it is critical to identify conserved structures in SARS-CoV-2 genomes as potential targets for variant-proof diagnostics and therapeutics. However, the algorithms to predict these conserved structures, which simultaneously fold and align multiple RNA homologs, scale at best cubically with sequence length, and are thus infeasible for coronaviruses, which possess the longest genomes (~30,000 nt) among RNA viruses. As a result, existing efforts on modeling SARS-CoV-2 structures resort to single sequence folding as well as local folding methods with short window sizes, which inevitably neglect long-range interactions that are crucial in RNA functions. Here we present LinearTurboFold, an efficient algorithm for folding RNA homologs that scales linearly with sequence length, enabling unprecedented global structural analysis on SARS-CoV-2. Surprisingly, on a group of SARS-CoV-2 and SARS-related genomes, LinearTurboFold's purely *in silico* prediction not only is close to experimentally-guided models for local structures, but also goes far beyond them by capturing the end-to-end pairs between 5' and 3' UTRs (~29,800 nt apart) that match perfectly with a purely experimental work. Furthermore, LinearTurboFold identifies novel conserved structures and conserved accessible regions as potential targets for designing efficient and mutation-insensitive small-molecule drugs, antisense oligonucleotides, siRNAs, CRISPR-Cas13 guide RNAs and RT-PCR primers. LinearTurboFold is a general technique that can also be applied to other RNA viruses and full-length genome studies, and will be a useful tool in fighting the current and future pandemics. Availability and implementation: Our source code is available at <https://github.com/LinearFold/LinearTurboFold>.**

RNA secondary structure | homologous folding | conserved structures | structural alignment | SARS-CoV-2

*n*. Several software packages provide implementations of the Sankoff algorithm<sup>12,13,14,15,16,17</sup> that use simplifications to reduce runtime.\*

As an alternative, TurboFold II,<sup>18</sup> an extension of TurboFold,<sup>19</sup> provides a more computationally efficient method to align and fold sequences. Taking multiple unaligned sequences as input, TurboFold II iteratively refines alignments and structure predictions so that they conform more closely to each other and converge on conserved structures. TurboFold II is significantly more accurate than other methods<sup>12,14,20,21,22</sup> when tested on RNA families with known structures and alignments.

However, the cubic runtime and quadratic memory usage of TurboFold II prevent it from scaling to longer sequences such as full-length SARS-CoV-2 genomes, which contain ~30,000 nucleotides; in fact, no joint-align-and-fold methods can scale to these genomes, which are the longest among RNA viruses. As a (not very principled) workaround, most existing efforts for modeling SARS-CoV-2 structures<sup>29,24,25,27,28,26</sup> resort to local folding methods<sup>30,31</sup> with sliding windows plus a limited pairing distance, abandoning all long-range interactions, and only consider one SARS-CoV-2 genome (Fig. 1B–C), ignoring signals available in multiple homologous sequences. To address this challenge, we designed a linearized version of TurboFold II, *LinearTurboFold* (Fig. 1A), which is a global homologous folding algorithm that scales linearly with sequence length. This linear runtime makes it the first joint-fold-and-align algorithm to scale to full-length coronavirus genomes without any constraints on window size or pairing distance, taking about 13 hours to analyze a group of 25 SARS-CoV homologs. It also leads to significant improvement

\* Besides these joint-fold-and-align algorithms, there exist two alternative approaches to homologous folding: *align-then-fold* and *fold-then-align*; see Fig. S6 for details.

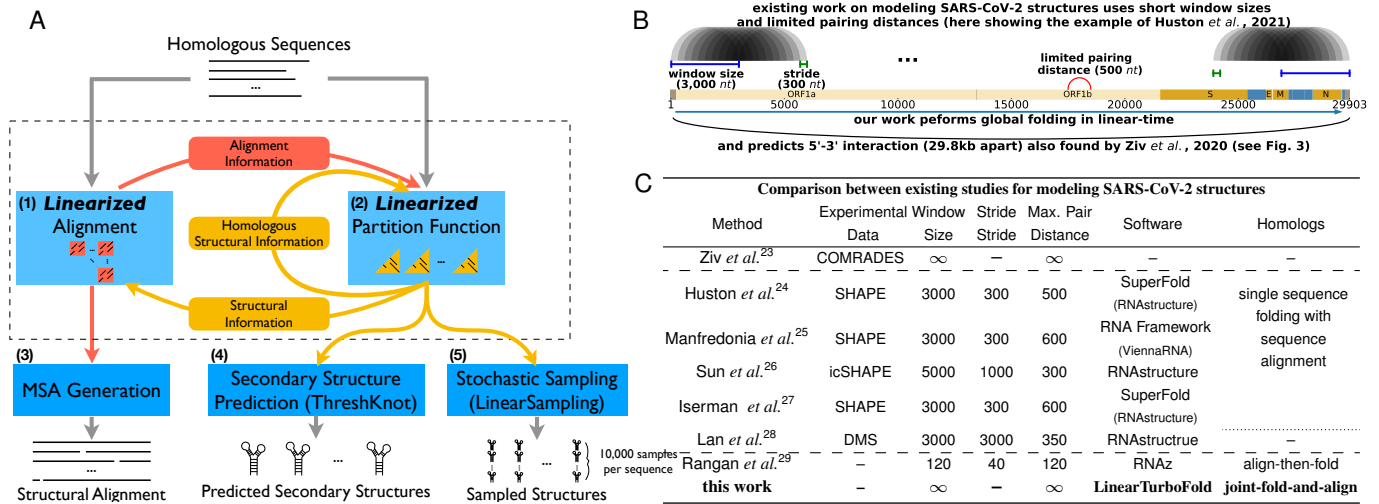
## Significance Statement

Conserved RNA structures are critical for designing diagnostic and therapeutic tools for many diseases including COVID-19. However, existing algorithms are much too slow to model the global structures of full-length RNA viral genomes. We present LinearTurboFold, a linear-time algorithm that is orders of magnitude faster, making it the first method to simultaneously fold and align whole genomes of SARS-CoV-2 variants, the longest known RNA virus (~30 kilobases). Our work enables unprecedented global structural analysis and captures long-range interactions that are out of reach for existing algorithms but crucial for RNA functions. LinearTurboFold is a general technique for full-length genome studies and can help fight the current and future pandemics.

\* Corresponding authors: David\_Mathews@urmc.rochester.edu, liang.huang.sh@gmail.com.

1 **R**ibonucleic acid (RNA) plays important roles in many cellular  
2 processes.<sup>1,2</sup> To maintain their functions, secondary structures of  
3 RNA homologs are conserved across evolution.<sup>3,4,5</sup> These conserved  
4 structures provide critical targets for diagnostics and treatments. Thus,  
5 there is a need for developing fast and accurate computational methods  
6 to identify structurally conserved regions.

7 Commonly, conserved structures involve compensatory base pair  
8 changes, where two positions in primary sequences mutate across  
9 evolution and still conserve a base pair, for instance, an AU or a  
10 CG pair replaces a GC pair in homologous sequences. These com-  
11 pensatory changes provide strong evidence for evolutionarily con-  
12 served structures.<sup>6,7,8,9,10</sup> Meanwhile, they make it harder to align  
13 sequences when structures are unknown. To solve this issue, Sankoff  
14 proposed a dynamic programming algorithm that simultaneously  
15 predicts structures and a structural alignment for two or more se-  
16 quences.<sup>11</sup> The major limitation of this approach is that the algorithm  
17 runs in  $O(n^{3k})$  against  $k$  sequences with the average sequence length



**Fig. 1.** A: The LinearTurboFold framework. Like TurboFold II, LinearTurboFold takes multiple unaligned homologous sequences as input and outputs a secondary structures for each sequence, and a multiple sequence alignment (MSA). But unlike TurboFold II, LinearTurboFold employs two linearizations to ensure linear runtime: a *linearized* alignment computation (module 1) to predict posterior co-occurrence probabilities (red squares) for all pairs of sequences (see [Methods §1-4](#)), and a *linearized* partition function computation (module 2) to estimate base-pairing probabilities (yellow triangles) for all the sequences (see [Methods §5-6](#)). These two modules take advantage of information from each other and iteratively refine predictions (Fig. S7). After several iterations, module 3 generates the final multiple sequence alignments (see [Methods §7](#)), and module 4 predicts secondary structures. Module 5 can stochastically sample structures. B-C: Prior studies (except for the purely experimental work by Ziv et al.) used local folding methods with limited window size and maximum pairing distance. B shows the local folding of the SARS-CoV-2 genome by Huston et al., which used a window of 3,000 nt that was advanced 300 nt. It also limited the distance between nucleotides that can base pair at 500. Some work also used homologous sequences to identify conserved structures, but they only predicted structures for one genome and utilized sequence alignments to identify mutations. By contrast, LinearTurboFold is a global folding method without any limitations on sequence length or pairing distance, and it jointly folds and aligns homologs to obtain conserved structures. Consequently, LinearTurboFold can capture long-range interactions even across the whole genome (the long arc in B and Fig. 3).

45 on secondary structure prediction accuracy as well as an alignment  
46 accuracy comparable to or higher than all benchmarks.

47 Over a group of 25 SARS-CoV-2 and SARS-related homologous  
48 genomes, LinearTurboFold predictions are close to the canonical struc-  
49 tures<sup>32</sup> and structures modeled with the aid of experimental data<sup>24,25,27</sup>  
50 for several well-studied regions. Thanks to global rather than local  
51 folding, LinearTurboFold discovers a long-range interaction involving  
52 5' and 3' UTRs (~29,800 nt apart), which is consistent with recent  
53 purely experimental work,<sup>28</sup> and yet is out of reach for local folding  
54 methods used by existing studies (Fig. 1B-C). In short, our *in silico*  
55 method of folding multiple homologs can achieve results similar to,  
56 and sometimes more accurate than, experimentally-guided models  
57 for one genome. Moreover, LinearTurboFold identifies conserved  
58 structures supported by compensatory mutations, which are poten-  
59 tial targets for small molecule drugs<sup>33</sup> and antisense oligonucleotides  
60 (ASOs).<sup>26</sup> We further identify regions that are (a) sequence-level  
61 conserved, (b) at least 15 nt long, and (c) accessible (i.e., likely to be  
62 completely unpaired) as potential targets for ASOs,<sup>34</sup> small interfering  
63 RNA (siRNA),<sup>35</sup> CRISPR-Cas13 guide RNA (gRNA)<sup>36</sup> and reverse  
64 transcription polymerase chain reaction (RT-PCR) primers.<sup>37</sup> Lin-  
65 earTurboFold is a general technique that can also be applied to other  
66 RNA viruses (e.g., influenza, Ebola, HIV, Zika, etc.) and full-length  
67 genome studies.

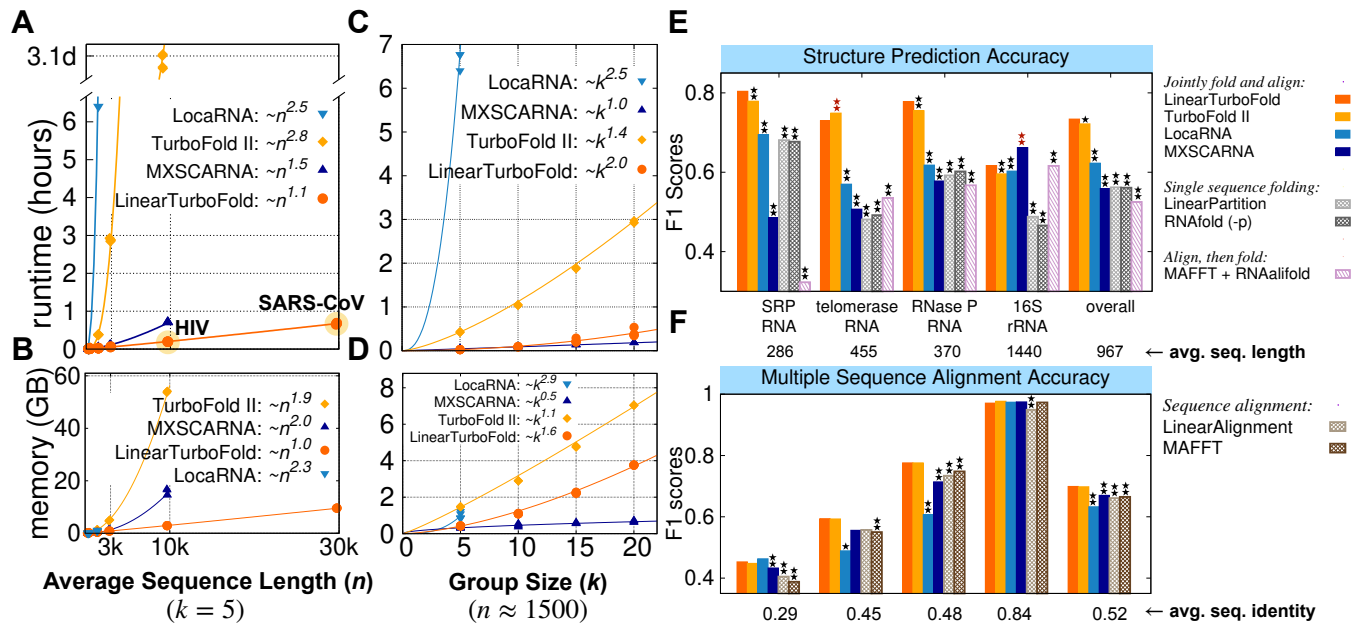
## 68 Results

69 The framework of LinearTurboFold has two major aspects (Fig. 1A):  
70 linearized structure-aware pairwise alignment estimation (module 1);  
71 and linearized homolog-aware structure prediction (module 2). Lin-  
72 earTurboFold iteratively refines alignments and structure predictions,  
73 specifically, updating pairwise alignment probabilities by incorporat-  
74 ing predicted base-pairing probabilities (from module 2) to form struc-

75 tural alignments, and modifying base-pairing probabilities for each  
76 sequence by integrating the structural information from homologous  
77 sequences via the estimated alignment probabilities (from module 1)  
78 to detect conserved structures. After several iterations, LinearTurbo-  
79 Fold generates the final multiple sequence alignment (MSA) based  
80 on the latest pairwise alignment probabilities (module 3) and predicts  
81 secondary structures using the latest pairing probabilities (module 4).

82 LinearTurboFold achieves linear time regarding sequence length  
83 with two major linearized modules: our recent work LinearParti-  
84 tion<sup>38</sup> (Fig. 1A module 2), which approximates the RNA partition  
85 function<sup>39</sup> and base pairing probabilities in linear time, and a novel  
86 algorithm LinearAlignment (module 1). LinearAlignment aligns two  
87 sequences by Hidden Markov Model (HMM) in linear time by apply-  
88 ing the same beam search heuristic<sup>40</sup> used by LinearPartition. Finally,  
89 LinearTurboFold assembles the secondary structure from the final  
90 base pairing probabilities using an accurate and linear-time method  
91 named ThreshKnot<sup>41</sup> (module 4). LinearTurboFold also integrates a  
92 linear-time stochastic sampling algorithm named LinearSampling<sup>42</sup>  
93 (module 5), which can independently sample structures according to  
94 the homolog-aware partition functions and then calculate the proba-  
95 bility of being unpaired for regions, which is an important property  
96 in, for example, siRNA sequence design.<sup>35</sup> Therefore, the overall  
97 end-to-end runtime of LinearTurboFold scales linearly with sequence  
98 length (see [Methods §1-7](#) for more details).

99 **Scalability and Accuracy.** To evaluate the efficiency of LinearTur-  
100 boFold against the sequence length, we collected a dataset consisting  
101 of seven families of RNAs with sequence length ranging from 210 nt  
102 to 30,000 nt, including five families from the RNAstralign dataset plus  
103 23S ribosomal RNA, HIV genomes and SARS-CoV genomes, and  
104 the calculation for each family uses five homologous sequences (see  
105 [Methods §8](#) for more details). Fig. 2A compares the running times of



**Fig. 2.** End-to-end Scalability and Accuracy Comparisons. **A–B:** End-to-end runtime and memory usage comparisons between benchmarks and LinearTurboFold against the sequence length. **C–D:** End-to-end runtime and memory usage comparisons against the group size. LinearTurboFold is the first joint-fold-and-align algorithm to scale to full-length coronavirus genomes ( $\sim 30,000$  nt) due to its linear runtime. **E–F:** The F1 accuracy scores of the structure prediction and multiple sequence alignment (see Tab. S1 for more details). LocARNA and MXSCARNA are Sankoff-style simultaneous folding and alignment algorithms for homologous sequences. As negative controls, LinearPartition and Vienna RNAfold-predicted structures for each sequence separately; LinearAlignment and MAFFT generated sequence-level alignments; RNAalifold folded pre-aligned sequences (e.g., from MAFFT) and predicted conserved structures. Statistical significances (two-tailed permutation test) between the benchmarks and LinearTurboFold are marked with one star (\*) on the top of the corresponding bars if  $p < 0.05$  or two stars (\*\*) if  $p < 0.01$ . The benchmarks whose accuracies are significantly lower than LinearTurboFold are annotated with black stars, while benchmarks higher than LinearTurboFold are marked with dark red stars. Overall, on structure prediction, LinearTurboFold achieves significantly higher accuracy than all evaluated benchmarks, and on multiple sequence alignment, it achieves accuracies comparable to TurboFold II and significantly higher than other methods (See Tab. S1 for detailed accuracies).

106 LinearTurboFold with TurboFold II and two Sankoff-style simultane- 134  
 107 ous folding and alignment algorithms, LocARNA and MXSCARNA. 135  
 108 Clearly, LinearTurboFold scales linearly with sequence length, and 136  
 109 is substantially faster than other algorithms, which scale superlinearly. 137  
 110 The linearization in LinearTurboFold brought orders of magnitude 138  
 111 speedup over the cubic-time TurboFold II, taking only 12 minutes on 139  
 112 the HIV family (average length 9,686 nt) while TurboFold II takes 3.1 140  
 113 days (372 $\times$  speedup). More importantly, LinearTurboFold takes only 141  
 114 40 minutes on five SARS-CoV sequences while all other benchmarks 142  
 115 fail to scale. Regarding the memory usage (Fig. 2B), LinearTurbo- 143  
 116 Fold costs linear memory space with sequence length, while other 144  
 117 benchmarks use quadratic or more memory. In Fig. 2C–D, we also 145  
 118 demonstrate that the runtime and memory usage against the number of 146  
 119 homologs ( $k = 5 \sim 20$ ), using sets of 16S rRNAs about 1,500 nt in 147  
 120 length. The apparent complexity against the group size of LinearTur- 148  
 121 boFold is higher than TurboFold II because the cubic-time partition 149  
 122 function calculation, which dominates the runtime of TurboFold II, 150  
 123 was linearized in LinearTurboFold by LinearPartition (Fig. S10C).

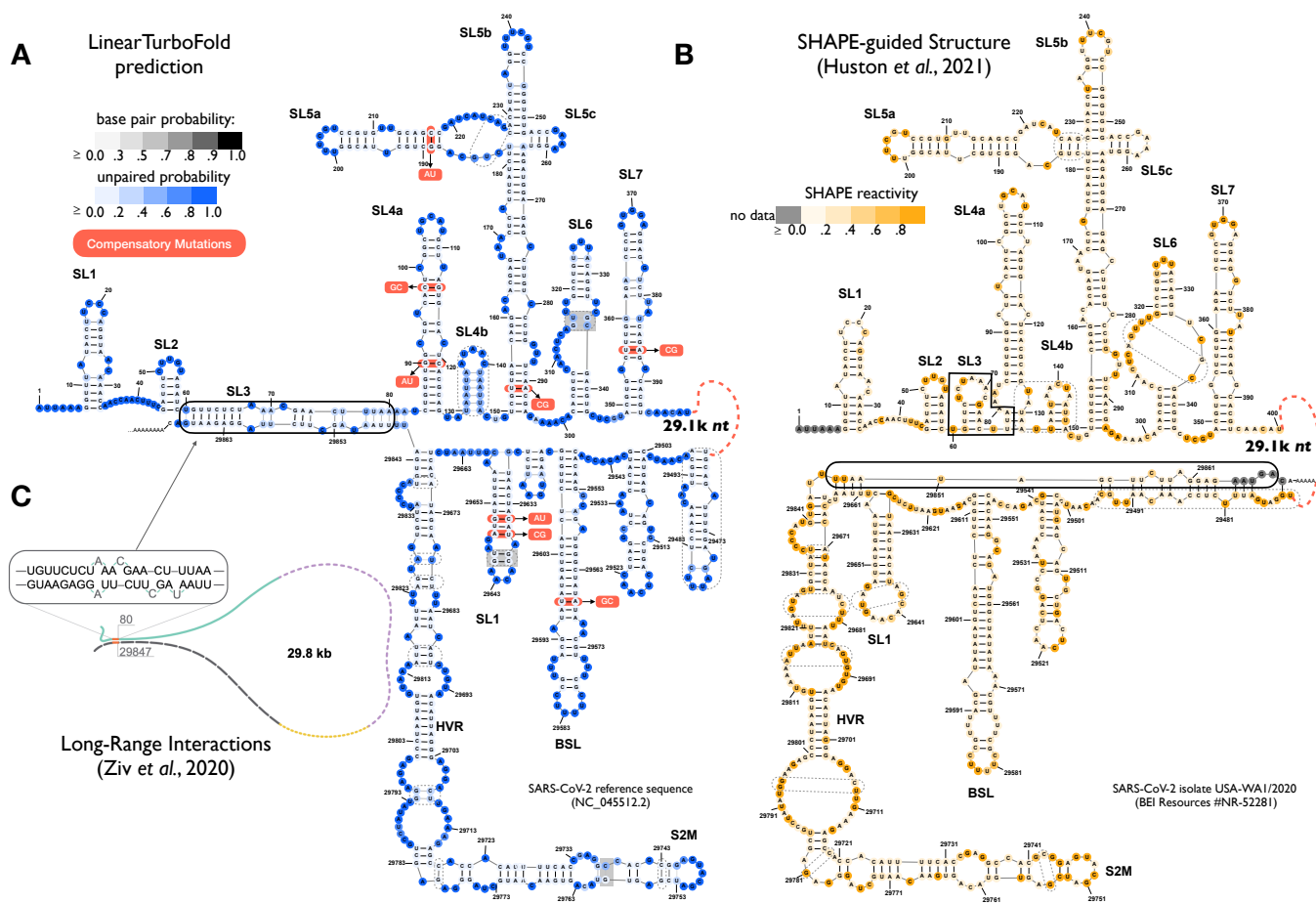
124 We next compare the accuracies of predicted secondary structures 151  
 125 and MSAs between LinearTurboFold and several benchmark meth- 152  
 126 ods (see Methods §9). Besides Sankoff-style LocARNA and MXS- 153  
 127 CARNA, we also consider three types of negative controls: (a) single 154  
 128 sequence folding (partition function-based): Vienna RNAfold<sup>31</sup> (-p 155  
 129 mode) and LinearPartition; (b) sequence-only alignment: MAFFT<sup>21</sup> 156  
 130 and LinearAlignment (a standalone version of the alignment method 157  
 131 developed for this work, but without structural information in Lin- 158  
 132 earTurboFold); and (c) an align-then-fold method that predicts con- 159  
 133 sensus structures from MSAs (Fig. S6): MAFFT + RNAalifold.<sup>20</sup> 160

For secondary structure prediction, LinearTurboFold, TurboFold 134  
 II and LocARNA achieve higher F1 scores than single sequence fold- 135  
 ing methods (Vienna RNAfold and LinearPartition) (Fig. 2E), which 136  
 demonstrates folding with homology information performs better than 137  
 folding sequences separately. Overall, LinearTurboFold performs sig- 138  
 nificantly better than all the other benchmarks on structure prediction. 139  
 For the accuracy of MSAs (Fig. 2F), the structural alignments from 140  
 LinearTurboFold obtain higher accuracies than sequence-only align- 141  
 ments (LinearAlignment and MAFFT) on all four families, especially 142  
 for families with low sequence identity. On average, LinearTurbo- 143  
 Fold performs comparably with TurboFold II and significantly better 144  
 than other benchmarks on alignments. We also note that the struc- 145  
 ture prediction accuracy of the align-then-fold approach (MAFFT + 146  
 RNAalifold) depends heavily on the alignment accuracy, and is the 147  
 worst when the sequence identity is low (e.g., SRP RNA) and the best 148  
 when the sequence identity is high (e.g., 16S rRNA) (Fig. 2E–F). 149

**Highly Conserved Structures in SARS-CoV-2 and SARS-related 150  
 Betacoronaviruses.** RNA sequences with conserved second- 151  
 ary structures play vital biological roles and provide potential 152  
 targets. The current COVID-19 outbreak raises an emergent require- 153  
 ment of identifying potential targets for diagnostics and therapeutics. 154  
 Given the strong scalability and high accuracy, we used LinearTur- 155  
 boFold on a group of full-length SARS-CoV-2 and SARS-related 156  
 (SARSr) genomes to obtain global structures and identify highly con- 157  
 served structural regions. 158

We used a greedy algorithm to select the 16 most diverse genomes 159  
 from all the valid SARS-CoV-2 genomes submitted to the Global 160





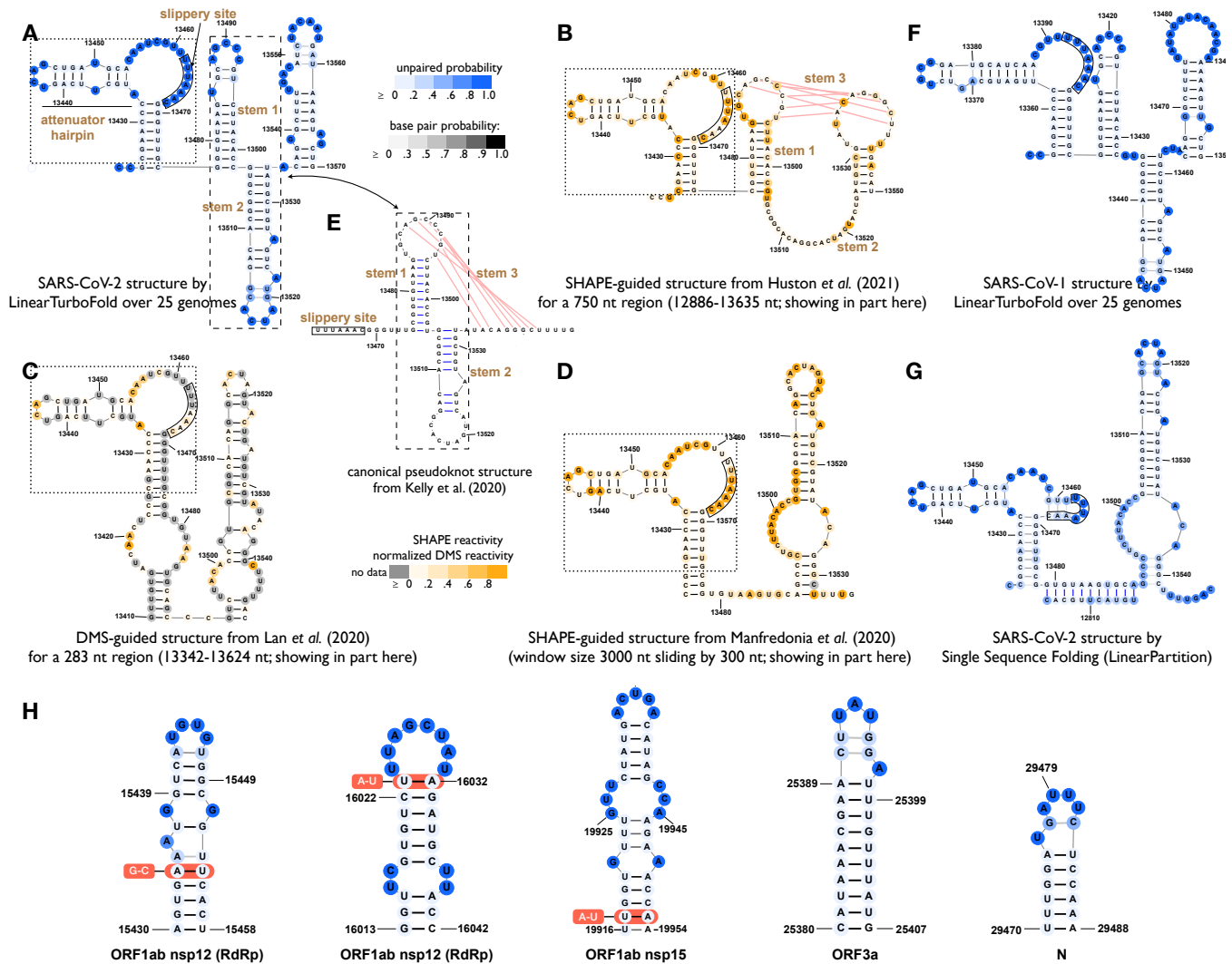
**Fig. 3.** Secondary structures predictions of SARS-CoV-2 extended 5' and 3' UTRs. **A:** LinearTurboFold prediction. The nucleotides and base pairs are colored by unpaired probabilities and base-pairing probabilities, respectively. The compensatory mutations extracted by LinearTurboFold are annotated with alternative pairs in red boxes (see Tab. S3 for more fully conserved pairs with co-variational changes). **B:** SHAPE-guided model by Huston *et al.*<sup>24</sup> (window size 3000 nt sliding by 300 nt with maximum pairing distance 500 nt). The nucleotides are colored by SHAPE reactivities. Dashed boxes enclose the different structures between **A** and **B**. Our model is close to Huston *et al.*'s, but the major difference is that LinearTurboFold predicts the end-to-end pairs involving 5' and 3' UTRs (solid box in **A**), which is *exactly* the same interaction detected by Ziv *et al.* using the COMRADES experimental technique<sup>23</sup> (**C**). Such long-range interactions cannot be captured by the local folding methods used by prior experimentally-guided models (Fig. 1B). The similarity between models **A** and **B** as well as the exact agreement between **A** and **C** show that our *in silico* method of folding multiple homologs can achieve results similar to, if not more accurate than, experimentally-guided single-genome prediction. As negative controls (Fig. S11), the align-then-fold (RNAfold) method cannot predict such long-range interactions. Although the single sequence folding algorithm (LinearPartition) predicts a long-range 5'-3' interaction, the positions are not the same as the LinearTurboFold model and Ziv *et al.*'s experimental result.

161 Initiative on Sharing Avian Influenza Data (GISAID)<sup>43</sup> up to De- 160  
162 cember 2020 (Methods §11). We further extended the group by 161  
163 adding 9 SARS-related homologous genomes (5 human SARS-CoV-1 162  
164 and 4 bat coronaviruses).<sup>44</sup> In total, we built a dataset of 25 full- 163  
165 length genomes consisting of 16 SARS-CoV-2 and 9 SARS-related 164  
166 sequences (Tab. S2). The average pairwise sequence identities of 165  
167 the 16 SARS-CoV-2 and the total 25 genomes are 99.9% and 89.6%, 166  
168 respectively. LinearTurboFold takes about 13 hours and 43 GB on the 167  
169 25 genomes. 168

170 To evaluate the reliability of LinearTurboFold predictions, we first 169  
171 compare them with the Huston *et al.*'s SHAPE-guided models<sup>24</sup> for 170  
172 regions with well-characterized structures across betacoronaviruses. 171  
173 For the extended 5' and 3' untranslated regions (UTRs), LinearTurbo- 172  
174 Fold's predictions are close to the SHAPE-guided structures (Fig. 3A- 173  
175 B), i.e., both identify the stem-loops (SLs) 1-2 and 4-7 in the extended 174  
176 5' UTR, and the bulged stem-loop (BSL), SL1, and a long bulge stem 175  
177 for the hypervariable region (HVR) including the stem-loop II-like 176  
178 motif (S2M) in the 3' UTR. Interestingly, in our model, the high 177  
179 unpaired probability of the stem in the SL4b indicates the possibility 178

of being single-stranded as an alternative structure, which is supported 180  
by experimental studies.<sup>26,25</sup> In addition, the compensatory mutations 181  
LinearTurboFold found in UTRs strongly support the evolutionary 182  
conservation of structures (Fig. 3A). 183

The most important difference between LinearTurboFold's prediction 184  
and Huston *et al.*'s experimentally-guided model is that Linear 185  
TurboFold discovers an end-to-end interaction (29.8 kilobases 186  
apart) between the 5' UTR (SL3, 60-82 nt) and the 3' UTR (final 187  
region, 29845-29868 nt), which fold locally by themselves in Hus- 188  
ton *et al.*'s model. Interestingly, this 5'-3' interaction matches *ex- 189*  
actly with the one discovered by the purely experimental work of 190  
Ziv *et al.*<sup>23</sup> using the COMRADES technique to capture long-range 191  
base-pairing interactions (Fig. 3C). These end-to-end interactions have 192  
been well established by theoretical and experimental studies<sup>45,46,47</sup> to 193  
be common in natural RNAs, but are far beyond the reaches of local 194  
folding methods used in existing studies on SARS-CoV-2 secondary 195  
structures.<sup>24,25,27,28</sup> By contrast, LinearTurboFold predicts secondary 196  
structures globally without any limit on window size or base-pairing 197  
distance, enabling it to discover long-distance interactions across the 198



**Fig. 4.** A–D: Secondary structure predictions of SARS-CoV-2 extended frameshifting stimulation element (FSE) region (13425–13545 nt). A: LinearTurboFold prediction. B–D: Experimentally-guided predictions from the literature,<sup>24,28,25</sup> which are sensitive to the context and region boundaries due to the use of local folding methods (Fig. S12). E: The canonical pseudoknot structure by the comparative analysis between SARS-CoV-1 and SARS-CoV-2 genomes.<sup>32</sup> For the 5' region of the FSE shown in dotted boxes (attenuator hairpin, internal loop with slippery site, and a stem), the LinearTurboFold prediction (A) is consistent with B–D; for the 3' region of the FSE shown in dashed boxes, our prediction (predicting stems 1–2 but missing 3) is closer to the canonical structure in E compared to B–D. F: LinearTurboFold prediction on SARS-CoV-1. G: Single sequence folding algorithm (LinearPartition) prediction on SARS-CoV-2, which is quite different from LinearTurboFold's. As another negative control, the align-then-fold method (RNAalifold) predicts a rather dissimilar structure (Fig. S12G). H: Five examples from 59 fully conserved structures among 25 genomes (see Tab. S4 for details), 26 of which are novel compared with prior work.<sup>29,24</sup>

199 whole genome. The similarity between our predictions and the experimental work shows that our *in silico* method of folding multiple  
 200 homologs can achieve results similar to, if not more accurate than,  
 201 those experimentally-guided single-genome prediction. We also observed  
 202 that LinearPartition, as a single sequence folding method, can also  
 203 predict a long-range interaction between 5' and 3' UTRs, but it  
 204 involves SL2 instead of SL3 of the 5' UTR (Fig. 3A), which indicates  
 205 that the homologous information assists to adjust the positions of  
 206 base pairs to be conserved in LinearTurboFold. Additionally, the  
 207 align-then-fold approach (MAFFT + RNAalifold) fails to predict such  
 208 long-range interactions (Fig. S11B).  
 209

210 The frameshifting stimulation element (FSE) is another well-  
 211 characterized region. For an extended FSE region, the LinearTurbo-  
 212 Fold prediction consists of two substructures (Fig. 4A): the 5' part  
 213 includes an attenuator hairpin and a stem, which are connected by a

long internal loop (16 nt) including the slippery site, and the 3' part  
 includes three stem loops. We observe that our predicted structure  
 of the 5' part is consistent with experimentally-guided models<sup>24,25,28</sup>  
 (Fig. 4B–D). In the attenuator hairpin, the small internal loop motif  
 (UU) was previously selected as a small molecule binder that stabilizes  
 the folded state of the attenuator hairpin and impairs frameshifting.<sup>33</sup>  
 For the long internal loop including the slippery site, we will show  
 in the next section that it is both highly accessible and conserved  
 (Fig. 5), which makes it a perfect candidate for drug design. For the  
 3' region of the FSE, LinearTurboFold successfully predicts stems  
 1–2 (but misses stem 3) of the canonical three-stem pseudoknot<sup>32</sup>  
 (Fig. 4E). Our prediction is closer to the canonical structure compared  
 to the experimentally-guided models<sup>24,25,28</sup> (Fig. 4B–D); one such  
 model (Fig. 4B) identified the pseudoknot (stem 3) but with an open  
 stem 2. Note that all these experimentally-guided models  
 214  
 215  
 216  
 217  
 218  
 219  
 220  
 221  
 222  
 223  
 224  
 225  
 226  
 227  
 228





271 accessible, and it takes advantage of not only SARS-CoV-2 variants  
272 but also homologous sequences, e.g., SARS-CoV-1 and bat coron-  
273 avirus genomes, to identify conserved regions from historical and  
274 evolutionary perspectives.

275 To get unstructured regions, Rangan *et al.*<sup>29</sup> imposed a threshold  
276 on unpaired probability of each position, which is a crude approxima-  
277 tion because the probabilities are not independent of each other. By  
278 contrast, the widely-used stochastic sampling algorithm<sup>53,42</sup> builds  
279 a representative ensemble of structures by sampling independent  
280 secondary structures according to their probabilities in the Boltz-  
281 mann distribution. Thus the accessibility for a region can be approx-  
282 imated as the fraction of sampled structures in which the region is  
283 single-stranded. LinearTurboFold utilized LinearSampling<sup>42</sup> to gener-  
284 ate 10,000 independent structures for each genome according to  
285 the modified partition functions after the iterative refinement (Fig. 1A  
286 module 5), and calculated accessibilities for regions at least 15 nt  
287 long. We then define *accessible regions* that are with at least 0.5  
288 accessibility among all 16 SARS-CoV-2 genomes (Fig. 5A–B). We  
289 also measure the free energy to open a target region  $[i, j]$ ,<sup>54</sup> notated:  
290  $\Delta G_u[i, j] = -RT(\log Z_u[i, j] - \log Z) = -RT \log P_u[i, j]$  where  
291  $Z$  is the partition function which sums up the equilibrium constants  
292 of all possible secondary structures,  $Z_u[i, j]$  is the partition function  
293 over all structures in which the region  $[i, j]$  is fully unpaired,  $R$  is  
294 the universal gas constant and  $T$  is the thermodynamic temperature.  
295 Therefore  $P_u[i, j]$  is the unpaired probability of the target region and  
296 can be approximated via sampling by  $s_0/s$ , where  $s$  is the sample  
297 size and  $s_0$  is the number of samples in which the target region is  
298 single-stranded. The regions whose free energy changes are close to  
299 zero need less free energy to open, thus more accessible to bind with  
300 siRNAs, ASOs, CRISPR-Cas13 gRNAs and RT-PCR primers.

301 Next, to identify *conserved regions* that are highly conserved  
302 among both SARS-CoV-2 and SARS-related genomes, we require  
303 that these regions contain at most three mutated sites on the 9 SARS-  
304 related genomes compared to the SARS-CoV-2 reference sequence  
305 because historically conserved sites are also unlikely to change in the  
306 future,<sup>55</sup> and the average sequence identity with reference sequence  
307 over a large SARS-CoV-2 dataset is at least 0.999 (here we use a  
308 dataset of  $\sim 2M$  SARS-CoV-2 genomes submitted to GISAID up to  
309 June 30, 2021<sup>†</sup>; see **Methods §11**). Finally, we identified 33 *accessi-*  
310 *ble and conserved regions* (Fig. 5G and Tab. S6), which are not only  
311 structurally accessible among SARS-CoV-2 genomes but also highly  
312 conserved among SARS-CoV-2 and SARS-related genomes (Fig. 5C).  
313 Because the specificity is also a key factor influencing siRNA effi-  
314 ciency,<sup>56</sup> we used BLAST against the human transcript dataset for  
315 these regions (Tab. S6). Finally, we also listed the GC content of each  
316 region. Among these regions, region 16 corresponds to the internal  
317 loop containing the slippery site in the extended FSE region, and it  
318 is conserved at both structural and sequence levels (Fig. 5D and 5H).  
319 Besides SARS-CoV-2 genomes, the SARS-related genomes such as  
320 the SARS-CoV-1 reference sequence (NC\_004718.3) and a bat coron-  
321 avirus (BCoV, MG772934.1) also form similar structures around  
322 the slippery site (Fig. 5A). By removing the constraint of conserva-  
323 tion on SARS-related genomes, we identified 38 additional candidate  
324 regions (Tab. S7) that are accessible but only highly conserved on  
325 SARS-CoV-2 variants.

326 We also designed a negative control by analyzing the SARS-CoV-  
327 2 reference sequence alone using LinearSampling, which can also  
328 predict accessible regions. However, these regions are not structurally  
329 conserved among the other 15 SARS-CoV-2 genomes, resulting in  
330 vastly different accessibilities, except for one region in the M gene

(Tab. S8). The reason for this difference is that, even with a high se-  
331 quence identity (over 99.9%), single sequence folding algorithms still  
332 predict greatly dissimilar structures for the SARS-CoV-2 genomes  
333 (Fig. 5E–F). Both regions (in nsp11 and N genes) are fully conserved  
334 among the 16 SARS-CoV-2 genomes, yet they still fold into vastly dif-  
335 ferent structures due to mutations outside the regions; as a result, the  
336 accessibilities are either low (nsp11) or in a wide range (N) (Fig. 5D).  
337 Conversely, addressing this by folding each sequence with proclivity  
338 of base pairing inferred from all homologous sequences, LinearTur-  
339 boFold structure predictions are more consistent with each other and  
340 thus can detect conserved structures (Fig. 5A–B). 341

## 342 Discussion

343 The constant emergence of new SARS-CoV-2 variants is reducing the  
344 effectiveness of exiting vaccines and test kits. To cope with this issue,  
345 there is an urgent need to identify conserved structures as promis-  
346 ing targets for therapeutics and diagnostics that would work in spite  
347 of current and future mutations. Here we presented LinearTurbo-  
348 Fold, an end-to-end linear-time algorithm for structural alignment and  
349 conserved structure prediction of RNA homologs, which is the first  
350 joint-fold-and-align algorithm to scale to full-length SARS-CoV-2  
351 genomes without imposing any constraints on base-pairing distance.  
352 We also demonstrate that LinearTurboFold leads to significant im-  
353 provement on secondary structure prediction accuracy as well as an  
354 alignment accuracy comparable to or higher than all benchmarks. 354

355 Unlike existing work on SARS-CoV-2 using local folding and  
356 single-sequence folding workarounds, LinearTurboFold enables un-  
357 precedented global structural analysis on SARS-CoV-2 genomes; in  
358 particular, it can capture long-range interactions, especially the one  
359 between 5' and 3' UTRs across the whole genome, which matches  
360 perfectly with a recent purely experiment work. Over a group of  
361 SARS-CoV-2 and SARS-related homologs, LinearTurboFold identi-  
362 fies not only conserved structures supported by compensatory muta-  
363 tions and experimental studies, but also accessible and conserved  
364 regions as vital targets for designing efficient small-molecule drugs,  
365 siRNAs, ASOs, CRISPR-Cas13 gRNAs and RT-PCR primers. Lin-  
366 earTurboFold is widely applicable to the analysis of other RNA viruses  
367 (influenza, Ebola, HIV, Zika, etc.) and full-length genome analysis.

## 368 Methods

369 Detailed description of our algorithms, datasets, and evaluation met-  
370 rics are available in the online version of the paper.

<sup>1</sup> Sean R. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929, 2001.

<sup>2</sup> Jennifer A. Doudna and Thomas R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418(6894):222–228, 2002.

<sup>3</sup> Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

<sup>4</sup> Edwin A. Brown, Hangchun Zhang, Li-Hua Ping, and Stanley M. Lemon. Secondary structure of the 5' nontranslated regions of hepatitis C virus and pestivirus genomic RNAs. *Nucleic Acids Research*, 20(19):5041–5045, 1992.

<sup>5</sup> Justin Ritz, Joshua S. Martin, and Alain Laederach. Evolutionary evidence for alternative structure in RNA sequence co-variation. *PLoS Computational Biology*, 9(7):e1003152–e1003152, 2013.

<sup>6</sup> Elena Rivas, Jody Clements, and Sean R Eddy. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*, 36(10):3072–3076, 2020.

<sup>7</sup> Robert W Holley, Jean Appar, George A Everett, James T Madison, Mark Marquisee, Susan H Merrill, John Robert Penswick, and Ada Zamir. Structure of a ribonucleic acid. *Science*, pages 1462–1465, 1965.

<sup>8</sup> Harry F Noller, JoAnn Kop, Virginia Wheaton, Jürgen Brosius, Robin R Gutell, Alexei M Kopylov, Ferdinand Dohme, Winship Herr, David A Stahl, Ramesh Gupta, et al. Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Research*, 9(22):6167–6189, 1981.

<sup>9</sup> Norman R Pace, David K Smith, Gary J Olsen, and Bryan D James. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA—a review. *Gene*, 82(1):65–75, 1989.

<sup>†</sup>The average sequence identity is 0.9987 on that  $\sim 2M$  dataset (downloaded on July 25, 2021).

- <sup>10</sup> KP Williams and DP Bartel. Phylogenetic analysis of tmRNA secondary structure. *RNA*, 2(12):1306–1310, 1996.
- <sup>11</sup> David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.
- <sup>12</sup> Sebastian Will, Kristin Reiche, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, 3(4):e65, 2007.
- <sup>13</sup> Jakob H Havgaard, Elfar Torarinnsson, and Jan Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Computational Biology*, 3(10):1896–1908, 2007.
- <sup>14</sup> Yasuo Tabei, Hisanori Kiryu, Taishin Kin, and Kiyoshi Asai. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, 9(1):33, 2008.
- <sup>15</sup> Zhenjiang Xu and David H Mathews. Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics*, 27(5):626–632, 2011.
- <sup>16</sup> David H Mathews and Douglas H Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 317(2):191–203, 2002.
- <sup>17</sup> Kengo Sato, Yuki Kato, Tatsuya Akutsu, Kiyoshi Asai, and Yasubumi Sakakibara. DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. *Bioinformatics*, 28(24):3218–3224, 2012.
- <sup>18</sup> Zhen Tan, Yinghan Fu, Gaurav Sharma, and David H. Mathews. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Research*, 45(20):11570–11581, 09 2017.
- <sup>19</sup> Arif O Harmanci, Gaurav Sharma, and David H Mathews. TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics*, 12(1):108, 2011.
- <sup>20</sup> Stephan H Bernhart, Ivo L Hofacker, Sebastian Will, Andreas R Gruber, and Peter F Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9(1):1–13, 2008.
- <sup>21</sup> Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.
- <sup>22</sup> Chuong B Do, Mahathi SP Mahabhashyam, Michael Brudno, and Serafim Batzoglou. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340, 2005.
- <sup>23</sup> Omer Ziv, Jonathan Price, Lyudmila Shalamova, Tsveta Kamenova, Ian Goodfellow, Friedemann Weber, and Eric A Miska. The short- and long-range RNA-RNA interactome of SARS-CoV-2. *Molecular cell*, 80(6):1067–1077, 2020.
- <sup>24</sup> Nicholas C Huston, Han Wan, Madison S Strine, Rafael de Cesaris Araujo Tavares, Craig B Wilen, and Anna Marie Pyle. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Molecular cell*, 81(3):584–598, 2021.
- <sup>25</sup> Ilaria Manfredonia, Chandran Nithin, Almudena Ponce-Salvaterra, Pritha Ghosh, Tomasz K Wirecki, Tycho Marinus, Natacha S Ogando, Eric J Snijder, Martijn J van Hemert, Janusz M Bujnicki, et al. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Research*, 48(22):12436–12452, 2020.
- <sup>26</sup> Lei Sun, Pan Li, Xiaohui Ju, Jian Rao, Wenzhe Huang, Lili Ren, Shaojun Zhang, Tuanlin Xiong, Kui Xu, Xiaolin Zhou, et al. In vivo structural characterization of the SARS-CoV-2 rna genome identifies host proteins vulnerable to repurposed drugs. *Cell*, 184(7):1865–1883, 2021.
- <sup>27</sup> Christiane Iserman, Christine A Roden, Mark A Boerneke, Rachel SG Sealfon, Grace A McLaughlin, Irwin Jungreis, Ethan J Fritch, Yixuan J Hou, Joanne Ekena, Chase A Weidmann, et al. Genomic RNA elements drive phase separation of the SARS-CoV-2 nucleocapsid. *Molecular cell*, 80(6):1078–1091, 2020.
- <sup>28</sup> Tammy CT Lan, Matthew F Allan, Lauren Malsick, Stuti Khandwala, Sherry SY Nyee, Mark Bathe, Anthony Griffiths, and Silvi Rouskin. Structure of the full SARS-CoV-2 RNA genome in infected cells. *BioRxiv*, 2020.
- <sup>29</sup> Ramya Rangan, Ivan N Zheludev, Rachel J Hagey, Edward A Pham, Hannah K Waymett-Steele, Jeffrey S Glenn, and Rhiju Das. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA*, 26(8):937–959, 2020.
- <sup>30</sup> Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):1–9, 2010.
- <sup>31</sup> Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdisen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):1, 2011.
- <sup>32</sup> Jamie A Kelly, Alexandra N Olson, Krishna Neupane, Sneha Munshi, Josue San Emeterio, Lois Pollack, Michael T Woodside, and Jonathan D Dinman. Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *Journal of Biological Chemistry*, 295(31):10741–10748, 2020.
- <sup>33</sup> Hafeez S Haniif, Yuquan Tong, Xiaohui Liu, Jonathan L Chen, Blessy M Suresh, Ryan J Andrews, Jake M Peterson, Collin A O’Leary, Raphael I Benhamou, Walter N Moss, et al. Targeting the SARS-CoV-2 RNA genome with small molecule binders and ribonuclease targeting chimera (RIBOTAC) degraders. *ACS Central Science*, 6(10):1713–1721, 2020.
- <sup>34</sup> Zhi John Lu and David H Mathews. Fundamental differences in the equilibrium considerations for siRNA and antisense oligodeoxynucleotide design. *Nucleic Acids Research*, 36(11):3738–3745, 2008.
- <sup>35</sup> Steffen Schubert, Arnold Grünweller, Volker A Erdmann, and Jens Kurreck. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. *Journal of Molecular Biology*, 348(4):883–893, 2005.
- <sup>36</sup> Omar O Abudayyeh, Jonathan S Gootenberg, Patrick Essletzbichler, Shuo Han, Julia Joung, Joseph J Belanto, Vanessa Verdine, David BT Cox, Max J Kellner, Aviv Regev, et al. RNA targeting with CRISPR-Cas13. *Nature*, 550(7675):280–284, 2017.
- <sup>37</sup> Stephen A Bustin and Tania Nolan. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *Journal of Biomolecular Techniques: JBT*, 15(3):155, 2004.
- <sup>38</sup> He Zhang, Liang Zhang, David H Mathews, and Liang Huang. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics*, 36(Supplement\_1):i258–i267, 2020.
- <sup>39</sup> John S. McCaskill. The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, 29:11105–1119, 1990.
- <sup>40</sup> Liang Huang and Kenji Sagae. Dynamic programming for linear-time incremental parsing. In *Proceedings of ACL 2010*, page 1077–1086, Uppsala, Sweden, 2010. ACL.
- <sup>41</sup> Liang Zhang, He Zhang, David H. Mathews, and Liang Huang. ThreshKnot: Thresholded prob-knot for improved RNA secondary structure prediction. *BioRxiv*, 2019.
- <sup>42</sup> He Zhang, Liang Zhang, Sizhen Li, David Mathews, and Liang Huang. LinearSampling: Linear-time stochastic sampling of RNA secondary structure with applications to SARS-CoV-2. *BioRxiv*, 2020.
- <sup>43</sup> Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID’s innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017.
- <sup>44</sup> Carmine Ceraolo and Federico M Giorgi. Genomic variance of the 2019-nCoV coronavirus. *Journal of Medical Virology*, 92(5):522–528, 2020.
- <sup>45</sup> Matthew G Seetin and David H Mathews. RNA structure prediction: an overview of methods. In *Bacterial Regulatory RNA*, pages 99–122. Springer, 2012.
- <sup>46</sup> Thomas JX Li and Christian M Reidys. The rainbow spectrum of RNA secondary structures. *Bulletin of Mathematical Biology*, 80(6):1514–1538, 2018.
- <sup>47</sup> Wan-Jung C Lai, Mohammad Kayedkhordeh, Erica V Cornell, Elie Farah, Stanislav Bellaousov, Robert Rietmeijer, Enea Salsi, David H Mathews, and Dmitri N Ermolenko. mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances. *Nature Communications*, 9(1):1–11, 2018.
- <sup>48</sup> Ramya Rangan, Andrew M Watkins, Jose Chacon, Rachael Kretsch, Wipapat Kladwang, Ivan N Zheludev, Jill Townley, Mats Rynge, Gregory Thain, and Rhiju Das. De novo 3D models of SARS-CoV-2 RNA elements from consensus experimental secondary structures. *Nucleic Acids Research*, 49(6):3092–3108, 2021.
- <sup>49</sup> Valeria Lulla, Michal P Wandel, Katarzyna J Bandyra, Rachel Uliferts, Mary Wu, Tom Dendooven, Xiaofei Yang, Nicole Doyle, Stephanie Oerum, Rupert Beale, et al. The stem loop 2 motif is a site of vulnerability for SARS-CoV-2. *BioRxiv*, pages 2020–09, 2021.
- <sup>50</sup> Zhi J. Lu and David H. Mathews. Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Research*, 36:640–647, 2008.
- <sup>51</sup> Stephen A Bustin, Vladimir Benes, Jeremy A Garson, Jan Hellemans, Jim Huggett, Mikael Kubista, Reinhold Mueller, Tania Nolan, Michael W Pfaffl, Gregory L Shipley, et al. The MIQE guidelines: Minimum information for publication of quantitative real-time pcr experiments. *Clinical Chemistry*, 55:611–622, 2009.
- <sup>52</sup> Myungsun Park, Joungha Won, Byung Yoon Choi, and C Justin Lee. Optimization of primer sets and detection protocols for SARS-CoV-2 of coronavirus disease 2019 (COVID-19) using PCR and real-time PCR. *Experimental & Molecular Medicine*, 52(6):963–977, 2020.
- <sup>53</sup> Ye Ding and Charles E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.
- <sup>54</sup> Ulrike Mückstein, Hakim Tafer, Jörg Hacker Müller, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.
- <sup>55</sup> Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.
- <sup>56</sup> Elham Fakhri, F Zare, and Ladan Teimoori-Toolabi. Precise and efficient siRNA design: a key point in competent gene silencing. *Cancer Gene Therapy*, 23(4):73–82, 2016.
- <sup>57</sup> Arif Ozgun Harmanci, Gaurav Sharma, and David H Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, 8(1):130, 2007.
- <sup>58</sup> Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.
- <sup>59</sup> Ivo L Hofacker, Stephan HF Bernhart, and Peter F Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227, 2004.
- <sup>60</sup> Stanislav Bellaousov and David H Mathews. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, 16(10):1870–1880, 2010.
- <sup>61</sup> Jamie J Cannone, Sankar Subramanian, Murray N Schnare, James R Collett, Lisa M D’Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi V Madabusi, Kirsten M Müller, et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3(1):2, 2002.
- <sup>62</sup> Yasuo Tabei, Koji Tsuda, Taishin Kin, and Kiyoshi Asai. SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinformatics*, 22(14):1723–1729, 2006.
- <sup>63</sup> Nima Aghaepour and Holger H Hoos. Ensemble-based prediction of RNA secondary structures. *BMC Bioinformatics*, 14(1):139, 2013.
- <sup>64</sup> Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269, 2020.
- <sup>65</sup> Paul P Gardner and Robert Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5(1):1–18, 2004.
- <sup>66</sup> Matthias Hochsmann, Thomas Toller, Robert Giegerich, and Stefan Kurtz. Local similarity in RNA secondary structures. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference*. CSB2003, pages 159–168. IEEE, 2003.
- <sup>67</sup> David H Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, 2004.
- <sup>68</sup> Adeel Afzal. Molecular diagnostic technologies for COVID-19: Limitations and challenges. *Journal of Advanced Research*, 2020.



## 368 Methods

369 **§1 Pairwise Hidden Markov Model.** We use a pairwise Hidden Markov  
370 Model (pair-HMM) to align two sequences.<sup>57,58</sup> The model includes three  
371 actions ( $h$ ): aligning two nucleotides from two sequences (ALN), inserting a  
372 nucleotide in the first sequence without a corresponding nucleotide in the other  
373 sequence (INS1), and a nucleotide insertion in the second sequence without a  
374 corresponding nucleotide in the first sequence (INS2). We then define  $\mathcal{A}(\mathbf{x}, \mathbf{y})$   
375 as a set of all the possible alignments for the two sequences, and one alignment  
376  $a \in \mathcal{A}(\mathbf{x}, \mathbf{y})$  as a sequence of steps  $(h, i, j)$  with  $m + 2$  steps, where  $(h, i, j)$   
377 means an alignment step at the position pair  $(i, j)$  by the action  $h$ . Thus, for  
378 the  $l$ th step  $a_l = (h_l, i_l, j_l) \in a$ , the values of  $i_l$  and  $j_l$  depend on the action  
379  $h_l$  and the positions  $i_{l-1}$  and  $j_{l-1}$  of  $a_{l-1}$ :

$$380 \quad a_l = \begin{cases} (\text{ALN}, & i_{l-1} + 1, & j_{l-1} + 1), & h_l = \text{ALN} \\ (\text{INS1}, & i_{l-1} + 1, & j_{l-1}), & h_l = \text{INS1} \\ (\text{INS2}, & i_{l-1}, & j_{l-1} + 1), & h_l = \text{INS2} \end{cases}$$

381 with  $(\text{ALN}, 0, 0)$  as the first step, and  $(\text{ALN}, |\mathbf{x}| + 1, |\mathbf{y}| + 1)$  as the last  
382 one. For two sequences  $\{\text{ACAAGU}, \text{AACUG}\}$ , one possible alignment  
383  $\{-\text{ACAAGU}, \text{AAC}-\text{UG}\}$  can be specified as  $\{(\text{ALN}, 0, 0) \rightarrow (\text{INS2}, 0, 1) \rightarrow$   
384  $(\text{ALN}, 1, 2) \rightarrow (\text{ALN}, 2, 3) \rightarrow (\text{INS1}, 3, 3) \rightarrow (\text{INS1}, 4, 3) \rightarrow (\text{ALN}, 5, 4) \rightarrow$   
385  $(\text{ALN}, 6, 5) \rightarrow (\text{ALN}, 7, 6)\}$ , where a gap symbol  $(-)$  represents a nucleotide  
386 insertion in the other sequence at the corresponding position (Fig. S8). The  
387 action  $h_l$  in each step  $(h_l, i_l, j_l)$  corresponds to a line segment starting from  
388 the previous node  $(i_{l-1}, j_{l-1})$  and stopping at the node  $(i_l, j_l)$ . Thus the line  
389 segment is horizontal, vertical or diagonal towards the top-right corner when  
390  $h_l$  is INS1, INS2 or ALN, respectively (Fig. S8).

391 We initialize the first step with the state ALN of probability 1, thus  
392  $p_\pi(\text{ALN}) = 1$ .  $p_i(h_2 | h_1)$  is the transition probability from the state  
393  $h_1$  to  $h_2$ , and  $p_e((c_1, c_2) | h_1)$  is the probability of the state  $h_1$  emitting a  
394 character pair  $(c_1, c_2)$  with values from  $\{A, G, C, U, -\}$ . Both the emission  
395 and transition probabilities were taken from TurboFold II. The function  $e()$   
396 yields a character pair based on  $a_l$  and the nucleotides of two sequences:

$$397 \quad e(\mathbf{x}, \mathbf{y}, a_l) = \begin{cases} (x_{i_l}, y_{j_l}), & h_l = \text{ALN} \\ (x_{i_l}, -), & h_l = \text{INS1} \\ (-, y_{j_l}), & h_l = \text{INS2} \end{cases}$$

398 where  $x_i$  and  $y_j$  are the  $i$ th and  $j$ th nucleotides of sequences  $\mathbf{x}$  and  $\mathbf{y}$ , re-  
399 spectively. Note that the first step  $a_0 = (\text{ALN}, 0, 0)$  and the last  $a_{m+1} =$   
400  $(\text{ALN}, |\mathbf{x}| + 1, |\mathbf{y}| + 1)$  do not have emissions.

401 We denote forward probability  $\alpha_{i,j}^h$ , encompassing the probability of the  
402 partial alignments of  $\mathbf{x}$  and  $\mathbf{y}$  up to positions  $i$  and  $j$ , and all the alignments  
403 that go through the step  $(h, i, j)$ :

$$404 \quad \alpha_{i,j}^h = \sum_{\substack{a \in \mathcal{A}(\mathbf{x}, \mathbf{y}) \\ \exists k, a_k = (h, i, j)}} p(\mathbf{x}, \mathbf{y}, a[:k]) \\ = p_\pi(h_0) \cdot \prod_{l=1}^k p_i(h_l | h_{l-1}) p_e(e(\mathbf{x}, \mathbf{y}, a_l) | h_l)$$

405 where  $a[:k]$  indicates the partial alignments from the starting node up to the  
406  $k$ th step and  $a_k = (h, i, j)$ . For instance,  $\alpha_{3,3}^{\text{ALN}}$ ,  $\alpha_{3,3}^{\text{INS1}}$  and  $\alpha_{3,3}^{\text{INS2}}$  corresponds  
407 to the region circled by the blue dashed lines (Fig. S8B, C and D). Similarly,  
408 the backward probability  $\beta_{i,j}^h$  assembles the probability of partial alignments  
409  $a[k+1:]$  from the  $(k+1)$ th step up to the end one:

$$410 \quad \beta_{i,j}^h = \sum_{\substack{a \in \mathcal{A}(\mathbf{x}, \mathbf{y}) \\ \exists k, a_k = (h, i, j)}} p(\mathbf{x}, \mathbf{y}, a[k+1:]) \\ = \left\{ \prod_{l=k+1}^m p_i(h_l | h_{l-1}) p_e(e(\mathbf{x}, \mathbf{y}, a_l) | h_l) \right\} \cdot p_i(h_{m+1} | h_m)$$

411 For example,  $\beta_{3,3}^{\text{ALN}}$ ,  $\beta_{3,3}^{\text{INS1}}$  and  $\beta_{3,3}^{\text{INS2}}$  are the regions circled by the yellow  
412 dashed line (Fig. S8B, C and D). Thus, the probability of observing two  
413 sequences  $p(\mathbf{x}, \mathbf{y})$  is  $\alpha_{|\mathbf{x}|+1, |\mathbf{y}|+1}^{\text{ALN}}$  or  $\beta_{0,0}^{\text{ALN}}$ .

414 **§2 Posterior Co-incidence Probability Computation.** Nucleotide positions  
415  $i$  and  $j$  in two sequences  $\mathbf{x}$  and  $\mathbf{y}$  are said to be *co-incident* (notated as  $i \sim j$ )  
416 in an alignment  $a$  if the alignment path goes through the node  $(i, j)$ .<sup>57</sup> Since

the node  $(i, j)$  is reachable by three actions  $\mathcal{H} = \{\text{ALN}, \text{INS1}, \text{INS2}\}$ , the  
co-incidence probability for a position pair  $(i, j)$  given two sequences is:

$$p(i \sim j | \mathbf{x}, \mathbf{y}) = \frac{1}{p(\mathbf{x}, \mathbf{y})} \sum_{\substack{a \in \mathcal{A}(\mathbf{x}, \mathbf{y}) \\ \exists h, (h, i, j) \in a}} p(\mathbf{x}, \mathbf{y}, a) \quad [1] \quad 419$$

where  $p(\mathbf{x}, \mathbf{y}, a)$  is the probability of two sequences with the alignment  $a$ , and  
 $p(\mathbf{x}, \mathbf{y})$  is the probability of observing two sequences, which is the sum of  
probability of all the possible alignments:

$$p(\mathbf{x}, \mathbf{y}) = \sum_{a \in \mathcal{A}(\mathbf{x}, \mathbf{y})} p(\mathbf{x}, \mathbf{y}, a) \quad 423$$

The co-incidence probability for positions  $i$  and  $j$  (Equation 1) can be  
computed by:

$$p(i \sim j | \mathbf{x}, \mathbf{y}) = \frac{\sum_h \alpha_{i,j}^h \cdot \beta_{i,j}^h}{\alpha_{|\mathbf{x}|+1, |\mathbf{y}|+1}^{\text{ALN}}} \quad 426$$

**§3 LinearAlignment.** Unlike a previous method<sup>57</sup> that fills out all the nodes  
in the alignment matrix by columns (Fig. S8), LinearAlignment scans the  
matrix based on the *step count*  $s$ , which is the sum value of  $i$  and  $j$  ( $s = i + j$ )  
for the partial alignments of  $\mathbf{x}_{[1,i]}$  and  $\mathbf{y}_{[1,j]}$ . As shown in the pseudocode  
(Fig. S9), the forward phase starts from the node  $(0, 0)$  in the state ALN of  
probability 1, then iterates the step count  $s$  from 0 to  $|\mathbf{x}| + |\mathbf{y}| - 1$ . For each  
step count  $s$  with a specific state  $h$  from  $\mathcal{H}$ , we first collect all the nodes  $(i, j)$   
with the step count  $s$  with  $\alpha_{i,j}^h$  existing, which means the position pair  $(i, j)$   
has been visited via the state  $h$  before. Then each node makes transitions to  
next nodes by these states, and updates the corresponding forward probabilities  
 $\alpha_{i+1,j}^{\text{INS1}}$ ,  $\alpha_{i,j+1}^{\text{INS2}}$  and  $\alpha_{i+1,j+1}^{\text{ALN}}$ , respectively.

The current alignment algorithm is still an exhaustive-search algorithm  
and costs quadratic time and space for all the  $|\mathbf{x}| \times |\mathbf{y}|$  nodes. To reduce  
the runtime, LinearAlignment uses the beam search heuristic algorithm<sup>40</sup>  
and keeps a limited number of promising nodes at each step. For each step count  
 $s$  with a state  $h$ , LinearAlignment applies the beam search method first over  
 $B(s, h)$ , which is the collection of all the nodes  $(i, j)$  with step count  $s$  and  
the presence of  $\alpha_{i,j}^h$  (Fig. S9 line 6). This algorithm only saves the top  $b_1$   
nodes with the highest forward scores in  $B(s, h)$ , and these are subsequently  
allowed to make transitions to the next states. Here  $b_1$  is a user-specified beam  
size and the default value is 100. In total,  $O(b_1 n)$  nodes survive because the  
length of  $s$  is  $|\mathbf{x}| + |\mathbf{y}|$  and each step count keeps  $b_1$  nodes. For simplicity,  
we show the topological order and the beam search method with alignment  
examples (Fig. S8A), while the forward-backward algorithm adopts the same  
idea by summing the probabilities of all the possible alignments.

After the forward phase, the backward phase (Fig. S9) performs in linear  
time to calculate the co-incidence probabilities automatically because only a  
linear number of nodes in  $B(s, h)$  are stored. Thus by pruning low-scoring  
candidates at each step in the forward algorithm, we reduce the runtime from  
 $O(n^2)$  to  $O(b_1 n)$  for aligning two sequences. For  $k$  input homologous  
sequences, LinearTurboFold computes posterior co-incidence probabilities for  
each pair of sequences by LinearAlignment, which costs  $O(k^2 b_1 n)$  runtime  
in total.

**§4 Match Scores Computation and Modified LinearAlignment.** To encour-  
age the pairwise alignment conforming with estimated secondary structures,  
LinearTurboFold predicts structural alignments by incorporating the secondary  
structural conformation. PMcomp<sup>59</sup> first proposed the match score to measure  
the structural similarity for position pairs between a pair of sequences, and  
TurboFold II adapts it as a prior. Based on the base pair probabilities  $P_{\mathbf{x}}(i, j)$   
estimated from the partition function for a sequence  $\mathbf{x}$ , a position  $i$  could  
be paired with bases upstream, downstream or unpaired, with correspond-  
ing probability  $P_{\mathbf{x},>}(i) = \sum_{j<i} P_{\mathbf{x}}(i, j)$ ,  $P_{\mathbf{x},<}(i) = \sum_{j>i} P_{\mathbf{x}}(i, j)$   
and  $P_{\mathbf{x},o}(i) = 1 - P_{\mathbf{x},>}(i) - P_{\mathbf{x},<}(i)$ , respectively. The match score  
 $m_{\mathbf{x},\mathbf{y}}(i, j)$  for two positions  $i$  and  $j$  from two sequences  $\mathbf{x}$  and  $\mathbf{y}$  is based on  
the probabilities of these three structural propensities from the last iteration  
( $t - 1$ ):

$$m_{\mathbf{x},\mathbf{y}}^{(t)}(i, j) = \alpha_1 \left[ \sqrt{P_{\mathbf{x},>}^{(t-1)}(i) \cdot P_{\mathbf{y},>}^{(t-1)}(j)} \sqrt{P_{\mathbf{x},<}^{(t-1)}(i) \cdot P_{\mathbf{y},<}^{(t-1)}(j)} \right] \\ + \alpha_2 \sqrt{P_{\mathbf{x},o}^{(t-1)}(i) \cdot P_{\mathbf{y},o}^{(t-1)}(j)} + \alpha_3 \quad 473$$

474 where  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$  are weight parameters trained in TurboFold II. The  
475 forward-backward phrases integrate the match score as a prior when aligning  
476 two nucleotides (Fig. S9 line 10, and Fig. S9 line 12).

477 TurboFold II separately pre-computes match scores for all the  $O(n^2)$   
478 position pairs for pairs of sequences before the HMM alignment calculation.  
479 However, only a linear number of pairs  $O(b_1 n)$  survive after applying the  
480 beam pruning in LinearAlignment. To reduce redundant time and space usage,  
481 LinearTurboFold calculates the corresponding match scores for co-incident  
482 pairs when they are first visited in LinearAlignment. Overall, for  $k$  homologous  
483 sequences, LinearTurboFold reduces the runtime of the whole module of  
484 pairwise posterior co-incident probability computation from  $O(k^2 n^2)$   
485 to  $O(k^2 b_1 n)$  by applying the beam search heuristic to the pairwise HMM  
486 alignment, and only calculating the match scores for position pairs that are  
487 needed.

488 **§5 Extrinsic Information Calculation.** To update partition functions for  
489 each sequence with the structural information from homologs, TurboFold<sup>19</sup>  
490 introduces *extrinsic information* to model the the proclivity for base pairing  
491 induced from the other sequences in the input set  $\mathcal{S}$ . The extrinsic information  
492  $e_{\mathbf{x}}(i, j)$  for a base pair  $(i, j)$  in the sequence  $\mathbf{x}$  maps the estimated base  
493 pairing probabilities of other sequences to the target sequence via the co-  
494 incident nucleotides between each pair of sequences:

$$\sum_{\mathbf{y} \in \{\mathcal{S} \setminus \mathbf{x}\}} (1 - s_{\mathbf{x}, \mathbf{y}}) \sum_{k, l} p_{\mathbf{y}}^{(t-1)}(k, l) \cdot p_{\mathbf{x}, \mathbf{y}}^{(t)}(i \sim k) \cdot p_{\mathbf{x}, \mathbf{y}}^{(t)}(j \sim l)$$

495 where  $p_{\mathbf{y}}^{(t-1)}(k, l)$  is the base pair probability for a base pair  $(k, l)$  in the  
496 sequence  $\mathbf{y}$  from  $(t-1)$ th iteration.  $p_{\mathbf{x}, \mathbf{y}}^{(t)}(i \sim k)$  and  $p_{\mathbf{x}, \mathbf{y}}^{(t)}(j \sim l)$  are the  
497 posterior co-incident probabilities for position pairs  $(i, k)$  and  $(j, l)$ , respec-  
498 tively, from  $(t)$ th iteration. The extrinsic information  $e_{\mathbf{x}}^{(t)}(i, j)$  first sums all  
499 the base pair probabilities of alignable pairs from another one sequence with  
500 the co-incident probabilities and then iterates over all the other sequences.  
501  $s_{\mathbf{x}, \mathbf{y}}$  is the sequence identity for sequences  $\mathbf{x}$  and  $\mathbf{y}$ . The sequences with a  
502 low identity contribute more to the extrinsic information than sequences of  
503 higher identity. The sequence identity is defined as the fraction of nucleotides  
504 that are aligned and identical in the alignment.

505 **§6 LinearPartition for Base Pairing Probabilities Estimation with Extrinsic**  
506 **Information.** The classical partition function algorithm scales cubically  
507 with sequence length. The slowness limits its extension to longer sequences.  
508 To address this bottleneck, our recent LinearPartition<sup>38</sup> algorithm approxi-  
509 mates the partition function and base pairing probability matrix computation in  
510 linear time. LinearPartition is significantly faster, and correlates better with the  
511 ground truth structures than the traditional cubic partition function calculation.  
512 Thus LinearTurboFold uses LinearPartition to predict base pair probabilities  
513 instead of the traditional  $O(n^3)$ -time partition function.

514 TurboFold introduces the extrinsic information  $e_{\mathbf{x}}^{(t)}(i, j)$  in the partition  
515 function as a pseudo-free energy term for each base pair  $(i, j)$ . Similarly, in  
516 LinearPartition, for each span  $[i, j]$ , which is the subsequence  $x_i \dots x_j$ , and  
517 its associated partition function  $Q(i, j)$ , the partition function is modified as  
518  $\tilde{Q}(i, j) = Q(i, j) e_{\mathbf{x}}^{(t)}(i, j)^\lambda$  if  $(x_i, x_j)$  is an allowed pair, where  $\lambda$  denotes  
519 the contribution of the extrinsic information relative to the intrinsic informa-  
520 tion. Specifically, at each step  $j$ , among all possible spans  $[i, j]$  where  $x_i$   
521 and  $x_j$  are paired, we replace the original partition function  $Q(i, j)$  with  
522  $Q(i, j) e_{\mathbf{x}}^{(t)}(i, j)^\lambda$  by multiplying the extrinsic information. Then LinearTur-  
523 boFold applies the beam pruning heuristic over the modified partition function  
524  $\tilde{Q}(i, j)$  instead of the original.

525 Similarly, TurboFold II obtains the extrinsic information for all the  $O(n^2)$   
526 base pairs before the partition function calculation of each sequence, while  
527 only a linear number of base pairs survives in LinearPartition. Thus, Lin-  
528 earTurboFold only requires the extrinsic information for those promising base  
529 pairs that are visited in LinearPartition. Overall, for  $k$  homologous sequences,  
530 LinearTurboFold reduces the runtime of base pair probabilities estimation for  
531 each sequence from  $O(kn^3 + k^2 n^2)$  to  $O(kb_1^2 n + k^2 b_2 n)$  by applying the  
532 beam search heuristic to the partition function calculation, and only calculating  
533 extrinsic information for the saved base pairs.

534 **§7 MSA Generation and Secondary Structure Prediction.** After several  
535 iterations, TurboFold II builds the multiple sequence alignment using a prob-  
536 abilistic consistency transformation, generating a guide tree and performing  
537 progressive alignment over the pairwise posterior co-incident probabilities.<sup>22</sup>

The whole procedure is accelerated in virtue of the sparse matrix by discarding  
538 alignment pairs of probability smaller than a threshold (0.01 by default). Since  
539 LinearAlignment uses the beam search method and only saves a linear number  
540 of co-incident pairs, the MSA generation in LinearTurboFold costs linear  
541 runtime against the sequence length straightforwardly.

542 Estimated base pair probabilities are fed into downstream methods to predict  
543 secondary structures. To maintain the end-to-end linear-time property,  
544 LinearTurboFold uses ThreshKnot,<sup>41</sup> which is a thresholded version of Prob-  
545 Knot<sup>60</sup> and only considers base pairs of probability exceeding a threshold  $\theta$   
546 ( $\theta = 0.3$  by default). We evaluate the performance of ThreshKnot and MEA  
547 with different hyperparameters ( $\theta$  and  $\gamma$ ). On a sampled RNAStrAlign training  
548 set, ThreshKnot is closer to the upper right-hand than MEA, which indicates  
549 that ThreshKnot always has a higher Sensitivity than MEA at a given PPV  
550 (Fig. S10B).

551 **§8 Efficiency and Scalability Datasets.** Four datasets are built and used for  
552 measuring efficiency and scalability. To evaluate the efficiency and scalability  
553 of LinearTurboFold with sequence length, we collected groups of homologous  
554 RNA sequences with sequence length ranging from 200 *nt* to 29,903 *nt* with  
555 a fixed group size 5. Sequences are sampled from RNAStrAlign dataset,<sup>18</sup>  
556 the Comparative RNA Web (CRW) Site,<sup>61</sup> the Los Alamos HIV database  
557 (<http://www.hiv.lanl.gov/>) and the SARS-related betacoronaviruses (SARS-  
558 related).<sup>44</sup> RNAStrAlign, aggregated and released with TurboFold II, is an  
559 RNA alignment and structure database. Sequences in RNAStrAlign are cate-  
560 gorized into families, i.e. sets of homologs, and some of families are further  
561 split into subfamilies. Each subfamily or family includes a multiple sequence  
562 alignment and ground truth structures for all the sequences. 20 groups of  
563 five homologs were randomly chosen from the small subunit ribosomal RNA  
564 (Alphaproteobacteria subfamily), SRP RNA (Protozoan subfamily), RNase P  
565 RNA (bacterial type A subfamily) and telomerase RNA families. For longer  
566 sequences, we sampled five groups of 23S rRNA (of sequence length ranging  
567 from 2,700 *nt* to 2,926 *nt*) from the CRW Site, HIV-1 genetic sequences (of  
568 sequence length ranging from 9,597 *nt* to 9,738 *nt*) from the Los Alamos  
569 HIV database, and SARS-related sequences (of sequence length ranging from  
570 29,484 *nt* to 29,903 *nt*). All the sequences in one group belong to the same  
571 subfamily or subtype. We sampled five groups for each family and obtained  
572 35 groups in total. Due to the runtime and memory limitations, we did not run  
573 TurboFold II on SARS-CoV-2 groups (Fig. 2, A and D).

574 To assess the runtime and memory usage of LinearTurboFold with group  
575 size, we fixed the sequence length around 1,500 *nt*, and sampled 5 groups  
576 of sequences from the small subunit ribosomal RNA (Alphaproteobacteria  
577 subfamily) with group size 5, 10, 15 and 20, respectively (Fig. 2, B and F). We  
578 used a Linux machine (CentOS 7.7.1908) with 2.30 GHz Intel Xeon E5-2695  
579 v3 CPU and 755 GB memory, and gcc 4.8.5 for benchmarks.

580 We built a test set from the RNAStrAlign dataset to measure and compare  
581 the performance between LinearTurboFold and other methods. 60 groups  
582 of input sequences consisting of five homologous sequences were randomly  
583 selected from the small subunit ribosomal RNA (rRNA) (Alphaproteobacteria  
584 subfamily), SRP RNA (Protozoan subfamily), RNase P RNA (bacterial type  
585 A subfamily) and telomerase RNA families from RNAStrAlign dataset. We  
586 removed sequences shorter than 1,200 *nt* for the small subunit rRNA to filter  
587 out subdomains, and removed sequences that are shorter than 200 *nt* for SRP  
588 RNA following the TurboFold II paper to filter out less reliable sequences. We  
589 resampled the test set five times and show the average PPV, Sensitivity and F1  
590 scores over the five samples (Fig. 2, C and F).

591 An RNAStrAlign training set was built to compare accuracies between  
592 MEA and ThreshKnot. 40 groups of 3, 5 and 7 homologs were randomly  
593 sampled from 5S ribosomal RNA (Eubacteria subfamily), group I intron (IC1  
594 subfamily), tmRNA, and tRNA families from RNAStrAlign dataset. We chose  
595  $\theta = 0.1, 0.2, 0.3, 0.4$  and  $0.5$  for ThreshKnot, and  $\gamma = 1, 1.5, 2, 2.5, 3, 3.5, 4,$   
596  $8$  and  $16$  for MEA. We reported the average secondary structure prediction  
597 accuracies (PPV and Sensitivity) across all training families (Fig. S10B).

598 **§9 Benchmarks.** The Sankoff algorithm<sup>11</sup> uses dynamic programming to  
599 simultaneously fold and align two or more sequences, and it requires  $O(n^{3k})$   
600 time and  $O(n^{2k})$  space for  $k$  input sequences with the average length  $n$ . Both  
601 LocARNA<sup>12</sup> and MXSCARNA<sup>14</sup> are Sankoff-style algorithms.

602 LocARNA (local alignment of RNA) costs  $O(n^2(n^2 + k^2))$  time and  
603  $O(n^2 + k^2)$  space by restricting the alignable regions. MXSCARNA progres-  
604 sively aligns multiple sequences as an extension of the pairwise alignment  
605 algorithm SCARNA<sup>62</sup> with improved score functions. SCARNA first aligns  
606 stem fragment candidates, then removes the inconsistent matching in the post-  
607 processing to generate the sequence alignment. MXSCARNA reduces runtime  
608 to  $O(k^3 n^2)$  and space to  $O(k^2 n^2)$  with a limited searching space of folding  
609

610 and alignment. Both MXSCARNA and LocARNA uses pre-computed base  
611 pair probabilities for each sequence as structural input. All the benchmarks  
612 use the default options and hyper-parameters running on the RNAStrAlign test  
613 set. TurboFold II iterates three times, then predicts secondary structures by  
614 MEA ( $\gamma=1$ ). LinearTurboFold also runs three iterations with default beam  
615 sizes ( $b_1 = b_2 = 100$ ) in LinearAlignment and LinearPartition, then predicts  
616 structures with ThreshKnot ( $\theta = 0.3$ ).

617 **§10 Significance Test.** We use a paired, two-tailed permutation test<sup>63</sup> to mea-  
618 sure the significant difference. Following the common practice, the repetition  
619 number is 10,000, and the significance threshold  $\alpha$  is 0.05.

620 **§11 SARS-CoV-2 Datasets.** We used two large SARS-CoV-2 datasets. The  
621 first dataset is used to draw a representative sample of most diverse SARS-  
622 CoV-2 genomes. We downloaded all the genomes submitted to GISAID<sup>43</sup>  
623 by December 29, 2020 (downloaded on December 29, 2020), and filtered out  
624 low-quality genomes (with more than 5% unknown characters and degenerate  
625 bases, shorter than 29,500 *nt*, or with framing error in the coding region), and  
626 we also discard genomes with more than 600 mutations compared with the  
627 SARS-CoV-2 reference sequence (NC\_0405512.2).<sup>64</sup> After preprocessing, this  
628 dataset includes about 258,000 genomes. To identify a representative group  
629 of samples with more variable mutations, we designed a greedy algorithm to  
630 select 16 most diverse genomes found at least twice in the 258,000  
631 genomes. The general idea of the greedy algorithm is to choose genomes  
632 one by one with the most new mutations compared with the selected samples,  
633 which consists of only the reference sequence at the beginning.

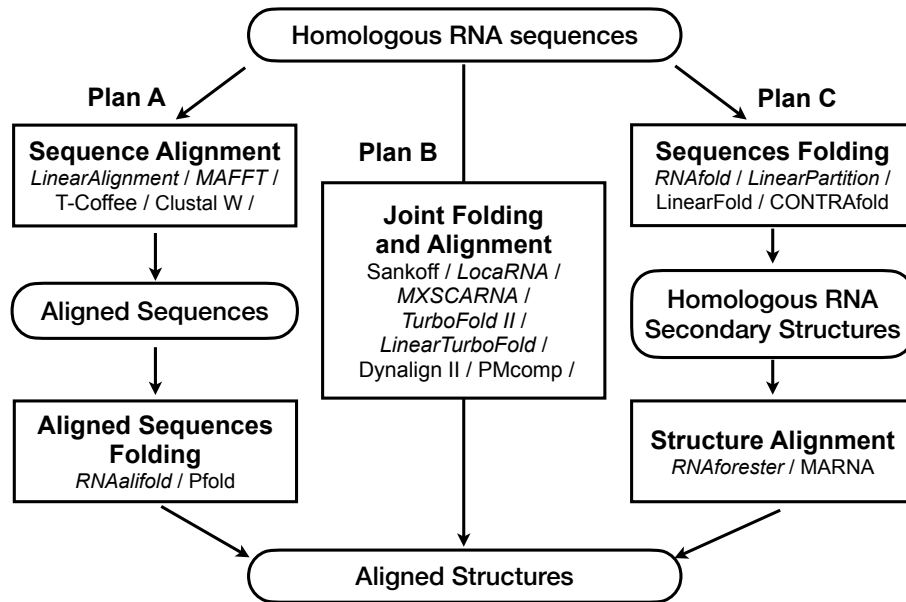
634 The second, larger, dataset is to evaluate the conservation of regions with  
635 respect to more up-to-date variants. We downloaded all the genomes submitted  
636 to GISAID by June 30, 2021 (downloaded on July 25, 2021), and did the same  
637 preprocessing as the first dataset. This resulted in a dataset of  $\sim 2M$  genomes,  
638 which was used to evaluate conservation in Figure 5 and Tables S5, S6, S7.



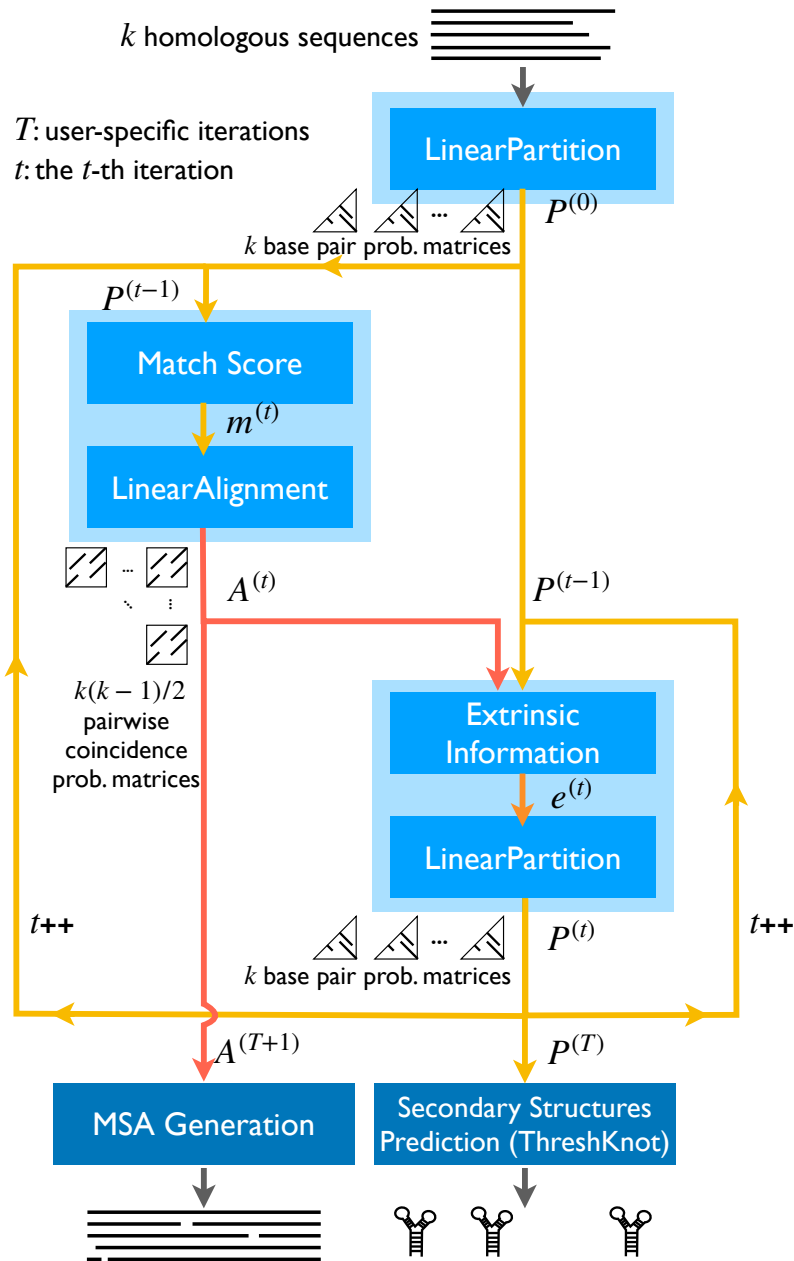
## Supporting Information

### LinearTurboFold: Linear-Time Global Prediction of Conserved Structures for RNA Homologs with Applications to SARS-CoV-2

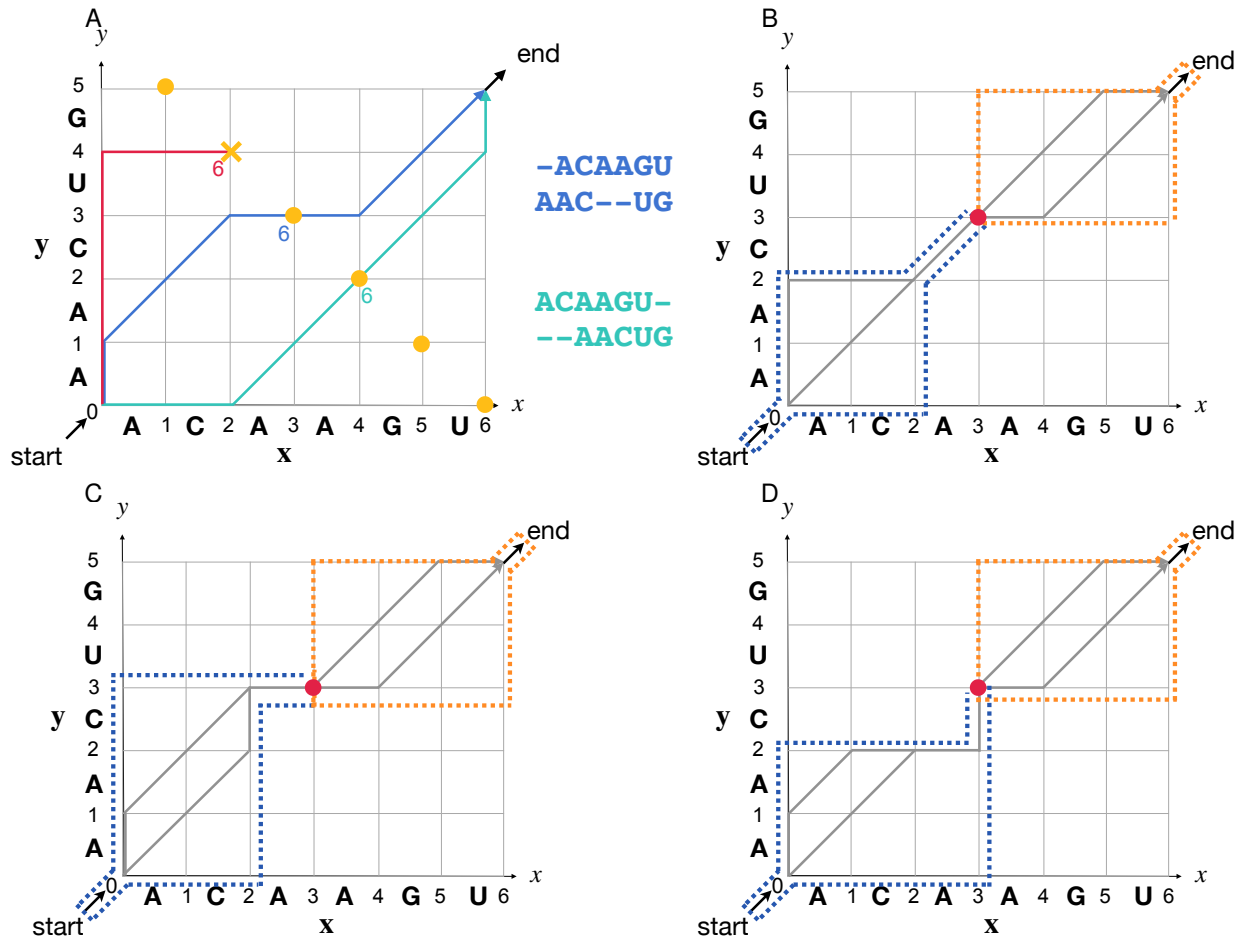
Sizhen Li, He Zhang, Liang Zhang, Kaibo Liu, Boxiang Liu, David H. Mathews, and Liang Huang



**Fig. S6.** Approaches for analyzing homologous sequence can be categorized into three plans<sup>65</sup> (related to Fig. 1). Plan A involves two steps: first aligning sequences and then folding aligned multiple sequences. This line works well for homologs with a high sequence identity. Plan B employs joint folding and alignment for multiple sequences, and it requires more time and space. Plan C folds sequences separately first and then aligns structures. Italic methods in each plan are evaluated on RNAStrAlign dataset (Tab. S1).



**Fig. S7.** The flowchart of LinearTurboFold with more detailed information (related to Fig. 1). At iteration 0, LinearPartition calculates the partition function and estimates the base pair probabilities for each sequence. From iteration 1 to  $T$ , the two major modules LinearAlignment and LinearPartition are conducted and updated in order with the match score and extrinsic information, respectively. The match score and extrinsic information are required and calculated for promising position pairs and base pairs during the LinearAlignment and LinearPartition computations, respectively. After  $T$  iterations, the match score and LinearAlignment computations are performed one more time over the latest the base pair probabilities. A multiple sequence is generated based on the pairwise co-incidence probabilities from the  $(T+1)$ -th iteration, and secondary structures are predicted according to the base pair probabilities for each sequence from the  $T$ -th iteration.



**Fig. S8.** Illustrations of LinearAlignment. **A:** An example of aligning two sequences and the beam search method based on the step count. The  $x$ -axis and  $y$ -axis of the matrix represent two sequences  $x$  and  $y$ . Yellow notes have the same step count 6. At step count 6, the red path is discarded because of its lower probability compared to others. There are two complete alignment paths (in green and blue) and the observed alignments are on the right side of the matrix with corresponding colors. **B:** The area enclosed by the blue dashed line corresponds to  $\alpha_{3,3}^{ALN}$ , which includes all the partial alignments arriving at the node  $(3, 3)$  by the state  $h$ . And the region circled by the orange dashed line maintains all the partial alignments starting from the step  $(ALN, 3, 3)$  ( $\beta_{3,3}^{ALN}$ ). **C and D:** The regions circled by the blue dashed lines are  $\alpha_{3,3}^{INS1}$  and  $\alpha_{3,3}^{INS2}$ , and regions circled by the orange dashed lines are  $\beta_{3,3}^{INS1}$  and  $\beta_{3,3}^{INS2}$ , respectively.



```

1: function FORWARD( $\mathbf{x}, \mathbf{y}, b_1$ )
2:  $\alpha_{0,0}^{ALN} \leftarrow 1$ 
3: for  $s = 0 \dots |\mathbf{x}| + |\mathbf{y}| - 1$  do
4:   for each  $h$  in  $\mathcal{H}$  do
5:      $B(s, h) \leftarrow$  all the nodes  $(i, j)$  such that  $\alpha_{i,j}^h$  exists and  $i + j = s$ 
6:     BEAMPRUNE( $B(s, h), b_1$ )
7:     for each node  $(i, j)$  in  $B(s, h)$  do
8:        $\alpha_{i+1,j}^{INS1} += \alpha_{i,j}^h \cdot p_t(INS1 | h) \cdot p_e((x_{i+1}, -) | INS1)$ 
9:        $\alpha_{i,j+1}^{INS2} += \alpha_{i,j}^h \cdot p_t(INS2 | h) \cdot p_e((- , y_{j+1}) | INS2)$ 
10:       $\alpha_{i+1,j+1}^{ALN} += \alpha_{i,j}^h \cdot p_t(ALN | h) \cdot p_e((x_{i+1}, y_{j+1}) | ALN)$ 
11: return  $\alpha$ 

```

▷ the forward phase  
 ▷ initial probability distribution  
 ▷ topological order  
 ▷ keep top  $b_1$  nodes in  $B(s, h)$  by  $\alpha_{i,j}^h$   
 ▷ transitions to next states

1. Pseudocode of the LinearAlignment algorithm forward phase

```

1: function BACKWARD( $\mathbf{x}, \mathbf{y}, \alpha, B$ )
2:  $\beta \leftarrow$  hash()
3:  $\beta_{|\mathbf{x}|+1, |\mathbf{y}|+1}^{ALN} \leftarrow 1$ 
4:  $p_{\mathbf{x}, \mathbf{y}} \leftarrow \alpha_{|\mathbf{x}|+1, |\mathbf{y}|+1}^{ALN}$ 
5:  $p_{i,j} \leftarrow 0$ 
6: for  $s = |\mathbf{x}| + |\mathbf{y}| \dots 0$  do
7:   for each  $h$  in  $\mathcal{H}$  do
8:     for each node  $(i, j)$  in  $B(s, h)$  do
9:       if  $i = |\mathbf{x}|$  and  $j = |\mathbf{y}|$  then
10:         $\beta_{i,j}^h = p_t(ALN | h) \cdot \beta_{|\mathbf{x}|+1, |\mathbf{y}|+1}^{ALN}$ 
11:       else
12:         $\beta_{i,j}^h = p_t(ALN | h) \cdot p_e((x_{i+1}, y_{j+1}) | ALN) \cdot \beta_{i+1, j+1}^{ALN}$ 
13:         $+ p_t(INS1 | h) \cdot p_e((x_{i+1}, -) | INS1) \cdot \beta_{i+1, j}^{INS1}$ 
14:         $+ p_t(INS2 | h) \cdot p_e((- , y_{j+1}) | INS2) \cdot \beta_{i, j+1}^{INS2}$ 
15:         $p_{i,j} += \frac{\alpha_{i,j}^h \cdot \beta_{i,j}^h}{p_{\mathbf{x}, \mathbf{y}}}$ 

```

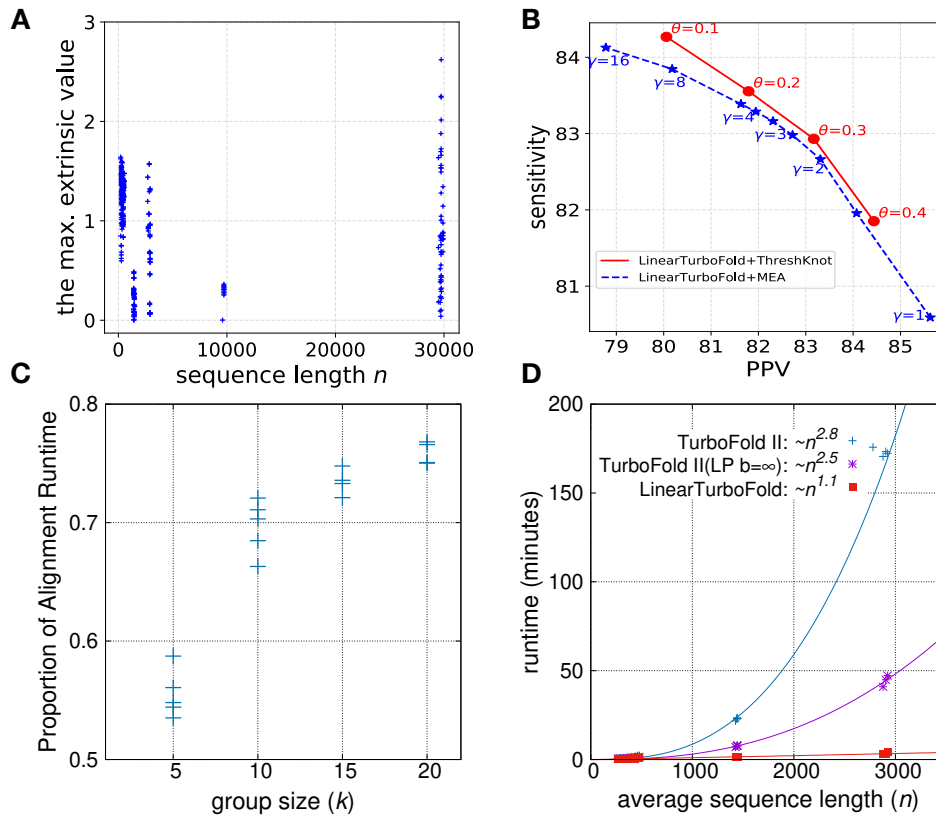
▷ the backward phase  
 ▷ initialization  
 ▷ probability of observing two sequences  
 ▷ co-incident probability initialization  
 ▷  $B(s, h)$  saves  $b_1$  entries during the forward phase  
 ▷ boundary conditions  
 ▷ update co-incident probabilities

2. Pseudocode of the LinearAlignment algorithm backward phase (co-incident probability computation)

**Fig. S9.** The pseudocode of the LinearAlignment algorithm forward and backward phases (co-incident probability computation). The pseudocode ignores boundary conditions for simplicity.

**Table S1.** Structure prediction and multiple sequence alignment accuracies (related to Fig. 2).

	<b>Structure Prediction Accuracy</b>						
	LinearTurboFold	TurboFold II	LocARNA	MXSCARNA	LinearPartition	Vienna RNAfold	RNAalifold
	PPV						
SRP	0.801	<b>0.819</b>	0.698	0.485	0.662	0.673	0.629
telomerase	0.650	<b>0.685</b>	0.516	0.465	0.409	0.430	0.602
RNase P RNA	0.734	<b>0.752</b>	0.606	0.573	0.543	0.571	0.698
16S rRNA	0.615	0.608	0.586	<b>0.662</b>	0.467	0.464	0.628
overall	0.700	<b>0.716</b>	0.602	0.546	0.520	0.534	0.639
	Sensitivity						
SRP	<b>0.806</b>	0.743	0.693	0.488	0.700	0.682	0.218
telomerase	<b>0.832</b>	0.826	0.637	0.558	0.584	0.576	0.482
RNase P RNA	<b>0.828</b>	0.758	0.630	0.584	0.650	0.636	0.478
16S rRNA	0.620	0.584	0.622	<b>0.663</b>	0.511	0.469	0.605
overall	<b>0.772</b>	0.728	0.645	0.573	0.611	0.591	0.446
	F1 scores						
SRP	<b>0.804</b>	0.779	0.695	0.486	0.681	0.677	0.323
telomerase	0.730	<b>0.749</b>	0.570	0.507	0.481	0.492	0.535
RNase P RNA	<b>0.778</b>	0.755	0.618	0.578	0.592	0.602	0.567
16S rRNA	0.617	0.596	0.603	<b>0.662</b>	0.488	0.466	0.616
overall	<b>0.734</b>	0.722	0.623	0.559	0.562	0.561	0.525
	<b>Multiple Sequence Alignment Accuracy</b>						
	LinearTurboFold	TurboFold II	LocARNA	MXSCARNA	LinearAlignment	MAFFT	RNAforester <sup>66</sup>
	PPV						
SRP	<b>0.463</b>	0.458	0.305	0.387	0.414	0.393	0.263
telomerase	<b>0.617</b>	0.615	0.311	0.554	0.575	0.572	0.239
RNase P RNA	<b>0.788</b>	0.787	0.615	0.692	0.744	0.759	0.258
16S rRNA	0.971	<b>0.977</b>	0.647	0.971	0.947	0.974	0.239
overall	<b>0.710</b>	0.709	0.470	0.651	0.670	0.675	0.250
	Sensitivity						
SRP	<b>0.443</b>	0.438	0.452	0.384	0.396	0.382	0.271
telomerase	<b>0.573</b>	0.572	0.470	0.523	0.540	0.529	0.262
RNase P RNA	<b>0.765</b>	<b>0.765</b>	0.596	0.684	0.724	0.738	0.286
16S rRNA	0.971	<b>0.977</b>	0.974	0.971	0.951	0.973	0.298
overall	<b>0.688</b>	<b>0.688</b>	0.623	0.641	0.653	0.656	0.280
	F1 scores						
SRP	<b>0.453</b>	0.448	0.364	0.385	0.405	0.388	0.267
telomerase	<b>0.594</b>	0.593	0.375	0.538	0.557	0.550	0.250
RNase P RNA	<b>0.776</b>	<b>0.776</b>	0.605	0.688	0.734	0.748	0.271
16S rRNA	0.971	<b>0.977</b>	0.778	0.971	0.949	0.973	0.265
overall	<b>0.699</b>	0.698	0.535	0.646	0.661	0.665	0.264

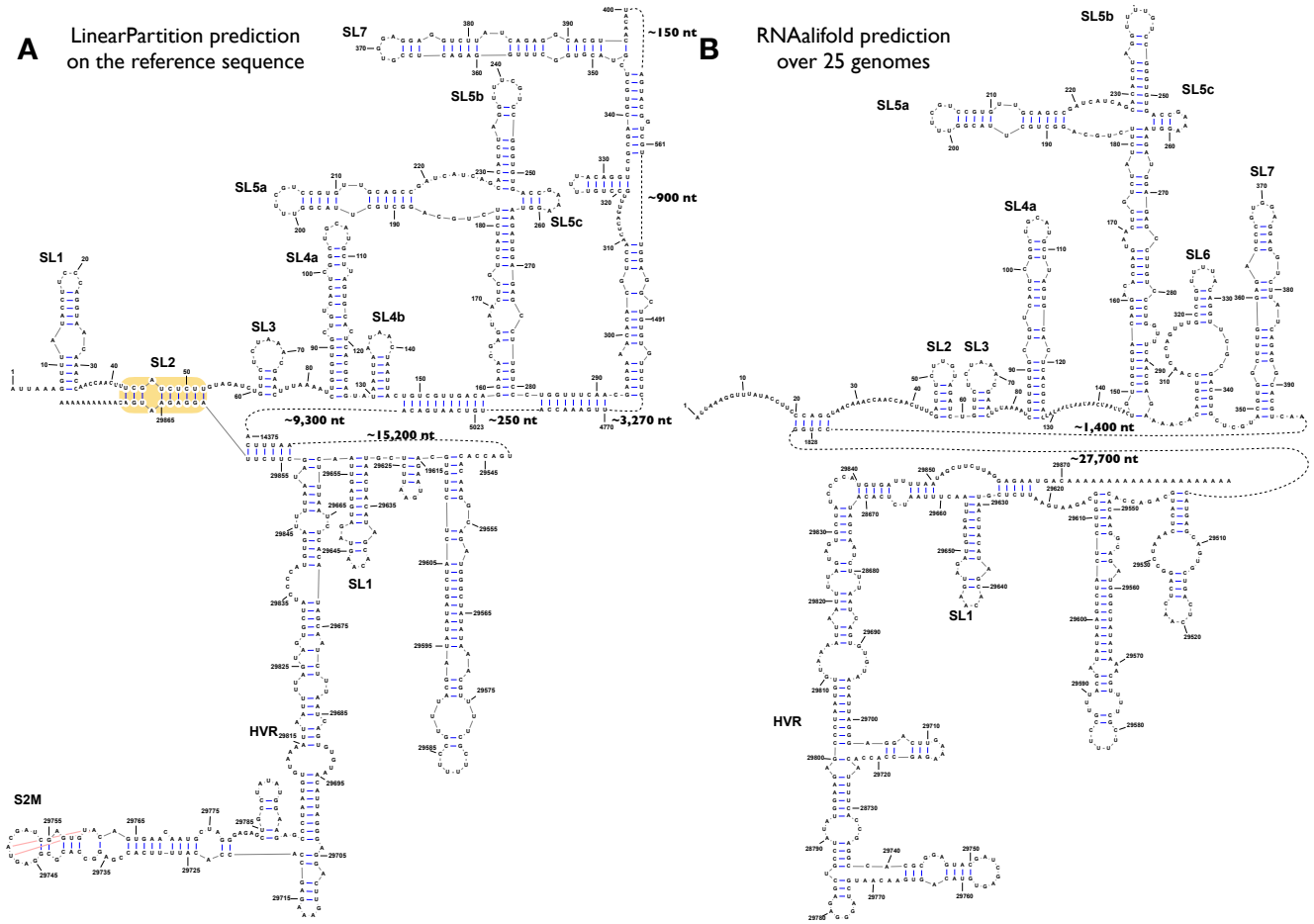


**Fig. S10.** A: The maximum values of the extrinsic information as a function of sequence length. The maximal value for each sequence is recorded when running LinearTurboFold on the collected dataset of sequence length ranging from 200 nt to 30,000 nt. B: Accuracy comparison between ThreshKnot and MEA on the training set with different hyper-parameters. C: The proportion of alignment runtime in the total runtime as the group size grows from 5 to 20. D: LinearPartition uses thermodynamic parameters from Vienna RNAfold,<sup>31</sup> which is a subset of the RNAstructure<sup>67</sup> partition function terms. By only replacing the TurboFold II partition function with LinearPartition with an infinite beam size (i.e., no approximation), the runtime decreases. This indicates part of speedup of LinearTurboFold profits from a simplified energy model.

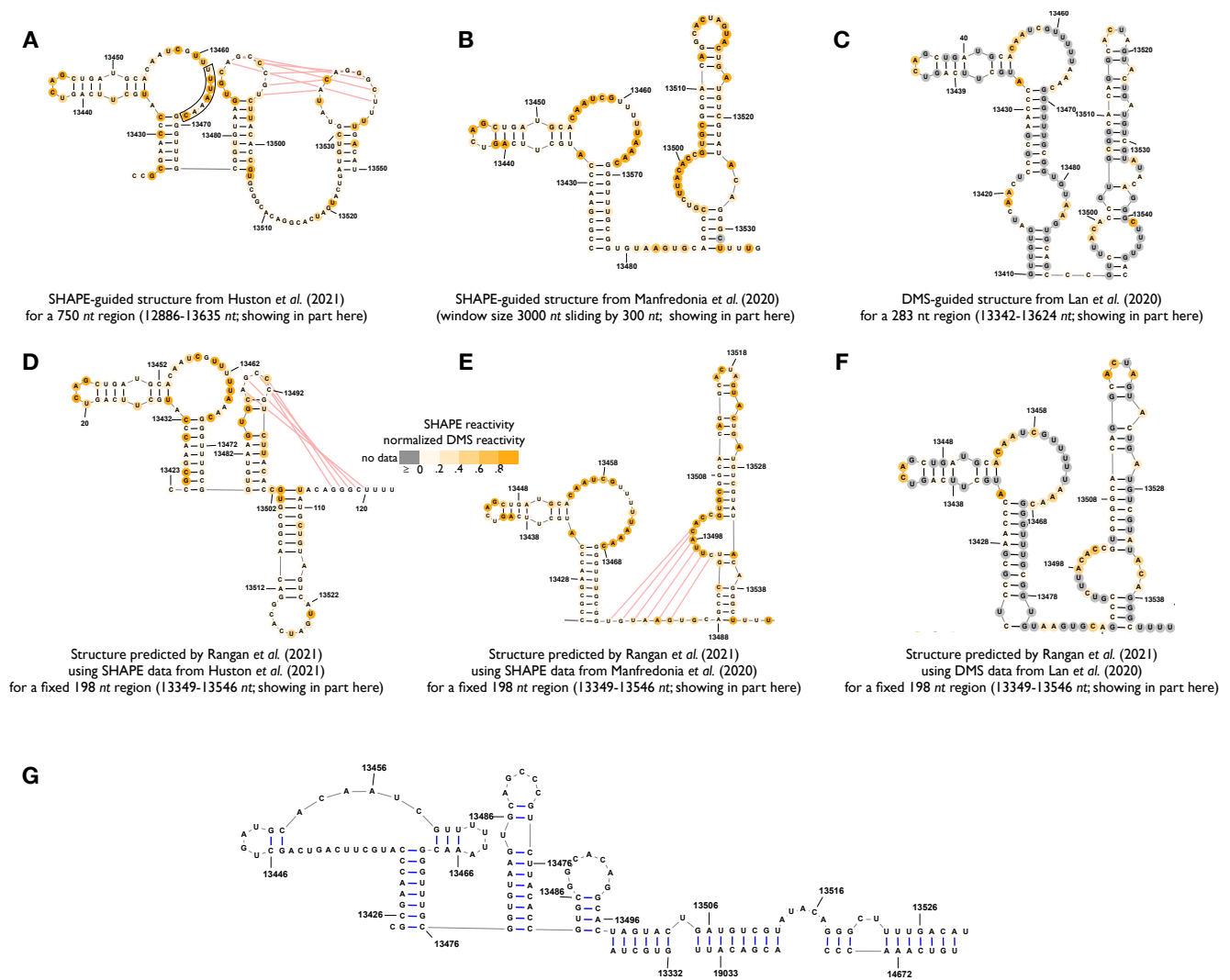


**Table S2.** Detailed information of the sampled 16 SARS-CoV-2 genomes and 9 SARS-related genomes (related to Fig. 3, 4 and 5). This dataset includes the reference sequences of SARS-CoV-2 and SARS-CoV-1 (NC\_0405512.2, NC\_004718.3). Most of the SARS-CoV-2 genomes include the D614G mutation, which has been a dominate mutation in the SARS-CoV-2 spike protein. B.1.1.7 lineage is a more infectious and lethal variant of SARS-CoV-2 first detected in the United Kingdom around November 2020. We utilized MAFFT<sup>21</sup> to generate the multiple sequence alignment and calculated the sequence identity with the reference sequence.

Accession ID	Species	Type	Submitted Date	Location	Length	Frequency	Mutations	Sequence identity	Note
NC_045512.2	human	SARS-CoV-2	2020-01-17	Wuhan, Asia	29903	2	-	-	-
EPI_ISL_454994	human	SARS-CoV-2	2020-03-02	Wuhan, Asia	29864	3	36	0.999	-
EPI_ISL_572982	human	SARS-CoV-2	2020-08-28	England, Europe	29882	2	28	0.999	D614G
EPI_ISL_573173	human	SARS-CoV-2	2020-09-08	England, Europe	29851	2	22	0.999	D614G
EPI_ISL_573220	human	SARS-CoV-2	2020-09-09	England, Europe	29784	2	19	0.999	D614G
EPI_ISL_576666	human	SARS-CoV-2	2020-09-18	England, Europe	29891	2	22	0.999	D614G
EPI_ISL_639684	human	SARS-CoV-2	2020-10-03	Latvia, Europe	29840	3	23	0.999	D614G
EPI_ISL_648168	human	SARS-CoV-2	2020-10-13	Sweden, Europe	29858	3	22	0.999	D614G
EPI_ISL_706936	human	SARS-CoV-2	2020-10-13	England, Europe	29828	2	23	0.999	D614G
EPI_ISL_638950	human	SARS-CoV-2	2020-10-14	Scotland, Europe	29891	2	29	0.999	D614G
EPI_ISL_654499	human	SARS-CoV-2	2020-10-20	Sweden, Europe	29876	3	20	0.999	D614G
EPI_ISL_666966	human	SARS-CoV-2	2020-10-30	USA, NorthAmerica	29879	2	23	0.999	D614G
EPI_ISL_704698	human	SARS-CoV-2	2020-11-01	England, Europe	29834	5	50	0.999	D614G B.1.1.7
EPI_ISL_723671	human	SARS-CoV-2	2020-11-08	England, Europe	29876	2	32	0.999	D614G
EPI_ISL_602304	human	SARS-CoV-2	2020-11-12	England, Europe	29838	2	27	0.999	D614G
EPI_ISL_710589	human	SARS-CoV-2	2020-11-19	Sweden, Europe	29815	2	28	0.999	D614G
NC_004718.3	human	SARS-CoV-1	2003-04-13	Vancouver, Canada	29751	-	6277	0.789	-
AY297028	human	SARS-CoV-1	2003-05-19	Beijing, Asia	29715	-	6306	0.788	-
AY515512.1	human	SARS-CoV-1	2005-01-01	Hong Kong, Asia	29731	-	6298	0.788	-
DQ182595.1	human	SARS-CoV-1	2005-08-26	Zhejiang, Asia	29706	-	6298	0.788	-
GU553363.1	human	SARS-CoV-1	2010-01-15	USA, NorthAmerica	29644	-	6351	0.786	-
EPI_ISL_402131	bat	SARS-CoV-2	2013-07-24	Yunnan, Asia	29855	-	1176	0.961	-
DQ022305.2	bat	SARS-CoV-1	2005-04-29	Hong Kong, Asia	29728	-	6337	0.787	-
DQ648857.1	bat	SARS-CoV-1	2006-05-23	Hong Kong, Asia	29741	-	6285	0.789	-
MG772934.1	bat	SARS-CoV-1	2008-01-05	Jiangsu, Asia	29732	-	3740	0.874	-



**Fig. S11.** Secondary structure prediction of SARS-CoV-2 for extended 5' and 3' UTRs (related to Fig. 3). **A:** LinearPartition prediction of the SARS-CoV-2 reference sequence (NC\_0405512.2) alone (single sequence folding). LinearPartition also predicts a long-range interaction between 5' and 3' UTRs. However, it involves the SL2 of the 5' UTR not SL3, which disagrees with LinearTurboFold prediction and Ziv *et al.* (Fig 3). **B:** RNAalifold (MFE) prediction over 25 genomes. RNAalifold did not find any 5'-3' pairs.



**Fig. S12.** Secondary structure predictions of SARS-CoV-2 for extended frameshifting stimulation element (13425-13545 nt) (related to Fig. 4). A–F: Experimentally-guided structures with different probing data for different regions. The structures in each column were estimated with the same experimental data but different regions. The structures in the second row were predicted by Rangan *et al.* for a fixed region of 198 nt<sup>48</sup>. G: RNAalifold (MFE) prediction over 25 genomes.



**Table S3.** Fully conserved base pairs across 25 complete SARS-CoV-2 and SARS-related genomes with compensatory mutations (related to Fig. 3 and 4). The positions and nucleotide type of base pairs correspond to the reference sequence of SARS-CoV-2 (NC\_0405512.2). The mutations are from the other 24 genomes.

5' End	3' End	Base Pair	Average Probability	Mutations	5' End	3' End	Base Pair	Average Probability	Mutations
90	121	GC	0.971	AU/GU	12931	12964	UA	0.851	AU/GC
97	115	AU	0.969	GU/GC	13069	13108	UA	1.000	CG
153	291	UA	0.972	CG	13078	13099	UA	1.000	UG/CG
159	282	GC	0.961	AU/GU	13216	13222	UA	0.958	UG/CG
189	217	GC	0.972	AU	13599	13628	UA	1.000	CG
358	385	UA	0.960	UG/CG	13638	13695	UA	0.986	AU
367	373	CG	0.961	UA	13641	13692	UA	0.977	CG
407	478	GC	0.936	AU	14091	14107	UA	0.989	CG
442	448	CG	0.960	UA/UG	14161	14194	UA	1.000	UG/CG
484	555	UA	0.877	AU	14205	14211	AU	0.933	CG
570	616	AU	0.946	UA/CG	14224	14251	AU	0.996	GU/GC
652	724	AU	0.947	GC	14355	14361	AU	0.996	GC
677	703	GC	0.933	AU	14487	14532	AU	0.973	GU/UA/UG/CG
880	889	AU	0.962	CG	14595	14604	UA	0.999	UG/CG
970	981	GC	0.963	AU	15435	15453	AU	0.778	GC
1231	1251	GC	0.968	AU	15582	15607	AU	0.993	GC
1949	1956	UA	0.929	UG/CG	16023	16032	UA	0.998	AU
2278	2303	UA	0.970	CG	16080	16110	CG	0.971	UA
2855	2875	CG	0.962	UA/UG	16089	16101	GC	1.000	AU
2896	2923	UA	0.973	AU/GU	16125	16155	AU	0.999	UA
2959	2986	UA	0.973	UG/CG	16230	16236	CG	0.999	UA
3712	3721	AU	0.977	GC	16677	16716	GC	1.000	AU
3913	3928	UA	0.979	AU/UG	17241	17256	UA	0.980	CG
3915	3926	AU	0.965	GC	17244	17253	AU	0.980	GC
4096	4108	UA	0.926	UG/CG	17304	17331	CG	0.981	UA
4189	4225	CG	0.980	GC/UG	18006	18054	UA	0.980	AU/GU
4603	4624	UA	0.980	UG/CG	18439	18468	UA	0.980	AU
4978	4987	UA	0.982	CG	18549	18561	AU	0.982	CG
5164	5203	GC	0.975	AU/GU	18717	18774	UA	0.983	UG/CG
5347	5374	UG	0.984	AU/GC/UA	19074	19098	UA	0.882	AU
5356	5371	UA	0.953	AU/GC	19386	19419	CG	0.986	UA/UG
5417	5428	UA	0.982	AU	19395	19410	UA	0.981	CG
5476	5521	AU	0.940	GU/GC	19707	19732	CG	0.981	UA
5482	5515	CG	0.984	UA/UG	19708	19731	AU	0.985	GU/GC
5739	5770	GC	0.984	AU	19917	19953	UA	0.980	AU
6034	6055	AU	0.983	GC	19929	19941	UA	0.984	AU
6037	6052	CG	0.982	UA	20172	20187	UA	0.977	CG
6154	6202	AU	0.987	GC	20217	20265	UA	0.939	CG
6328	6343	AU	0.988	UA	20223	20260	AU	0.981	GC
6364	6388	GC	0.989	AU	20523	20541	UA	0.988	CG
6367	6385	GC	0.989	AU	20841	20901	AU	0.988	GU/GC
6458	6490	AU	0.988	GC	20985	20997	AU	0.757	UA
6460	6488	UA	0.988	CG	21163	21201	AU	0.980	GU/GC
6903	6922	CG	0.895	UA	21300	21321	AU	0.988	GU/GC
6977	7006	GC	0.970	AU/GU	21411	21423	CG	0.989	UA
7103	7135	AU	0.891	GU/GC	21513	21523	CG	0.988	UA
7480	7531	UA	0.942	UG/CG	22837	22903	AU	0.988	GU/GC
7558	7597	AU	0.956	GC	23531	23548	AU	0.717	GC
7864	7876	AU	0.972	GU/GC	23621	23647	GC	0.991	AU
8146	8219	CG	0.993	UA	23797	23806	UA	0.931	UG/CG
8147	8218	AU	0.992	GU/GC	23980	24088	AU	0.978	GU/GC

**Table S3 continued from previous page**

5' End	3' End	Base Pair	Average Probability	Mutations	5' End	3' End	Base Pair	Average Probability	Mutations
8153	8212	UA	0.987	CG	23983	24085	AU	0.974	UA/CG
8317	8332	AU	0.995	GU/GC	24121	24152	AU	0.994	GC
8437	8458	AU	0.915	GU/GC	24553	24586	CG	0.996	UA
8698	8738	UA	0.824	CG	24757	24766	GC	0.974	GU/CG
8860	8881	CG	0.996	AU/UA	25336	25370	AU	0.906	GU/GC
9046	9079	UA	0.995	UG/CG	25991	26004	GC	0.996	AU/GU
9055	9070	AU	0.969	GC	26145	26190	UA	0.997	CG
9427	9433	UA	0.991	CG	26262	26305	GC	0.911	AU
9472	9511	AU	0.996	UA	26630	26658	AU	0.903	GC
9689	9703	AU	0.932	GC	26676	26706	AU	0.986	GC
9842	9874	UA	0.990	UG/CG	26939	26975	AU	0.996	CG
10213	10248	UA	0.997	CG	27412	27456	UA	0.998	UG/CG
10651	10669	UA	0.998	CG	27415	27453	GC	0.994	AU
10864	10906	AU	0.926	GC	27603	27613	CG	0.996	UA
10873	10898	AU	0.946	GC	27699	27744	UA	0.935	CG
10984	11026	AU	0.976	GC	27717	27725	GC	0.990	AU/GU
11782	11803	AU	0.981	GU/GC	28642	28664	UA	0.997	UG/CG
11788	11797	CG	1.000	UA	28910	28930	AU	0.997	GU/GC
11971	12013	AU	0.966	GU/GC	29567	29597	AU	0.999	GC
11989	11995	UA	0.763	UG/CG	29635	29651	CG	0.999	AU
12538	12577	UA	0.940	UG/CG	29637	29649	UA	0.991	CG

**Table S4.** Fully conserved structures among 25 genomes (related to Fig. 3 and 4). Regions with compensatory mutations are annotated with alternative base pairs. Novel regions compared with Rangan *et al.*<sup>29,24</sup> are annotated with stars.

Region	Sequence & Structure	Compensatory Mutations
45-59	GAUCUCUUGUAGAUC ((((.....))))	
84-127	CUGUGUGGCUGUCACUCGGCUGCAUGCUUAGUGCACUCACGCAG (((((((.....)))))))))	GC ->AU AU ->GC
626-643*	GUUCUUCUUCGUAAGAAC ((((.....))))	
1324-1341	UGCCACUACUUGUGGUUA ((((.....))))	
1685-1698	GCCAUUAUUUUGGC (((.....)))	
1785-1800*	GUAUUUUAAAGUUAC ((((.....))))	
2818-2837*	AGUACUUAUGAGAAGUGCU ((((.....))))	
4973-4993	GUGUUACAACAGUAGACAAC (((((((.....)))))))))	UA ->CG
5021-5033	AUGUCA AUGACAU (((.....)))	
6450-6498	UUGAGUGUAUGUGAAAACUACCGAAGUUGUAGGAGACUUUAUACUUA (((((((.....)))))))))	AU ->GC UA ->CG
8078-8084	CCA AUGG ((.....))	
10494-10509*	GUGUUGGUUUUAACAU ((((.....))))	
11131-11140	UGCUUUUGCA (((.....)))	
12203-12218*	UUGAAGAAGUCUUUGA ((((.....))))	
12257-12268	CAACGUAAGUUG (((.....)))	
12386-12412	GAUAAUGAUGCACUCAACAACAUUAUC (((((((.....)))))))))	
12672-12685	CUGUCAAAUUAACAG (((.....)))	
12904-12926*	UAGGUUUGUACAGACACACCUA (((((((.....)))))))))	
12970-12988*	CAACCUAAUJAGGUAUG ((((.....))))	
13409-13422*	AGUUGUGAUCAACU ((((.....))))	
14729-14769	AGGAAGGAAGUUCUGUUGAAUAAAACACUUCUUCUUUGCU (((.....)))	
14773-14790	GAUGGUAUUGCUGCUAUC ((((.....))))	
14794-14818	GAUUUAGACUACUUCGUUAUAUC (((((((.....)))))))))	
15254-15269*	UUUAUAGUGAUGUAGA ((((.....))))	
15430-15458*	AGUGAAUGGUC AUGUGGCGGUUCACU (((((((.....)))))))))	AU ->GC
15502-15509*	GCUUAUGC (((.....)))	
15618-15628	ACUUUAUGAGU (((.....)))	
15775-15801*	AUAAAGAACUUUAAGUCAGUUCUUUAU (((((((.....)))))))))	
16013-16042*	GGUUCGUGUCUUUAGCUAUUGAUGC UUACC (((.....)))	UA ->AU
16180-16194	AGGUAUUGGGAACCU (((.....)))	
16955-16973*	UAGUGCCACAAGAGCACUA ((((.....))))	







**Table S5.** Accessibility and conservation of target regions for public RT-PCR forward/reverse primers and probes.<sup>68</sup> The accessibility is computed by LinearTurboFold, and it is underlined if larger than zero. The conservation on 9 SARS-related genomes is the number of mutated sites. The conservation on the ~2M SARS-CoV-2 dataset is the percentage of exact matches, which is underlined or bold if less than 0.97 or 0.5, respectively. (The average sequence identity of SARS-CoV-2 genomes is 0.9987, and the average length of primers and probes is 23 nt. Therefore, the probability of randomly sampling a region of length 23 nt without mutations is  $0.9987^{23} \approx 0.97$ ).

Gene	Institute	Forward Primer / Probe / Reverse Primer			Conservation	
		Start	Length	Accessibility	SARS-related # mut. site	SARS-CoV-2 (2M) exact match
ORF1ab nsp9	Institut Pasteur (1)	12689 / 12717 / 12779	18 / 21 / 18	0.0000 / 0.0000 / <u>0.0160</u>	4 / 3 / 5	0.9989 / 0.9967 / 0.9829
ORF1ab nsp10	China CDC (1)	13341 / 13377 / 13441	21 / 30 / 19	0.0000 / 0.0000 / 0.0000	4 / 3 / 11	0.9937 / 0.9875 / 0.9868
ORF1ab nsp12 (RdRp)	Institut Pasteur (2)	14079 / 14105 / 14166	19 / 19 / 20	0.0000 / <u>0.0001</u> / 0.0000	4 / 8 / 9	0.9978 / 0.9332 / 0.9941
	Charite Germany (1)	15430 / 15469 / 15504	22 / 25 / 26	0.0000 / 0.0000 / 0.0000	0 / 6 / 1	<u>0.9167</u> / 0.9938 / 0.9982
ORF1ab nsp14	HKU (1)	18777 / 18849 / 18888	20 / 24 / 21	0.0000 / 0.0000 / 0.0000	1 / 1 / 3	0.9958 / 0.9969 / 0.9933
E	Charite Germany (2)	26268 / 26332 / 26359	26 / 26 / 22	0.0000 / 0.0000 / 0.0000	0 / 0 / 0	0.9958 / 0.9969 / 0.9933
	CDC (1)	28286 / 28309 / 28334	20 / 24 / 24	0.0000 / 0.0000 / 0.0000	12 / 2 / 8	0.9913 / 0.9762 / 0.9934
	NIH Thailand	28319 / 28341 / 28357	20 / 16 / 19	0.0000 / <u>0.0026</u> / 0.0000	3 / 6 / 6	0.9908 / 0.9953 / 0.9927
	CDC (2)	28680 / 28704 / 28731	22 / 24 / 21	0.0000 / 0.0000 / <u>0.0010</u>	4 / 4 / 2	0.9862 / 0.9796 / 0.9895
N	Charite Germany (3)	28705 / 28753 / 28813	19 / 25 / 20	0.0000 / <u>0.0003</u> / 0.0000	2 / 0 / 6	0.9914 / 0.9920 / 0.9858
	China CDC (2)	28880 / 28934 / 28957	22 / 20 / 22	<u>0.0710</u> / 0.0000 / 0.0000	5 / 7 / 4	<b>0.2734</b> / 0.9911 / <b>0.4844</b>
	NIID Japan	29124 / 29222 / 29262	20 / 20 / 20	0.0000 / 0.0000 / 0.0000	6 / 1 / 7	0.9953 / 0.9785 / 0.9853
	HKU (2)	29144 / 29179 / 29235	22 / 20 / 19	0.0000 / 0.0000 / 0.0000	2 / 2 / 1	0.9904 / 0.9945 / 0.9895
	CDC (3)	29163 / 29188 / 29212	20 / 23 / 18	0.0000 / 0.0000 / 0.0000	2 / 6 / 2	0.9892 / 0.9797 / 0.9901

**Table S6.** Accessible and conserved regions (related to Fig. 5) with two kinds of constraints on conservation: 1) at most three mutations on SARS-related genomes; 2) the average sequence identity on the SARS-CoV-2 dataset at least 0.999. The start positions and sequences correspond to the reference sequence of SARS-CoV-2 (NC\_0405512.2). The accessibilities are calculated from folding with homologs (LinearTurboFold) and single sequence folding (LinearPartition), respectively. We searched for these regions among human representative transcript set (RefSeq Select RNA sequences, refseq\_select) using BLAST, and several regions have the exact matches with human transcripts (underlined). Using single sequence folding can only get one accessible region (bold). The conservation of these regions on 9 SARS-related genomes is the number of mutated sites. The table also shows two types of conservations on a large SARS-CoV-2 dataset containing ~2M genomes submitted to GISAID up to June 30, 2021: the average sequence identity with reference sequence, and the percentage of exact matches of the whole region, respectively.

Region	Start	Length	Sequence	Gene	Accessibility				Conservation		BLAST Match	GC (%)
					LinearTurboFold (Homologous Folding)			Single Seq.	SARSr (9)	SARS-CoV-2 (2M)		
					Average	Range	$\Delta G(\text{kcal/mol})$	Folding Range	# Mut. Sites	Identity / Exact		
Region 1	739	18	AGAAAACUGGAACACUAA	ORF1ab nsp1	0.71 ± 0.04	0.62 – 0.76	0.22 ± 0.04	0.00 – 0.00	2/18	0.9998 / 0.9970	14/18	33
Region 2	995	17	CGUUCUGAAAAGAGCUA		0.99 ± 0.00	0.99 – 0.99	0.01 ± 0.00	0.00 – 0.00	3/17	0.9999 / 0.9985	14/17	41
Region 3	998	17	UCUGAAAAGAGCUAUGA	ORF1ab nsp2	1.00 ± 0.00	1.00 – 1.00	0.00 ± 0.00	0.00 – 0.00	3/17	0.9999 / 0.9984	15/17	35
Region 4	1001	15	GAAAAGAGCUAUGAA		0.74 ± 0.08	0.51 – 0.80	0.19 ± 0.07	0.00 – 0.00	3/15	0.9999 / 0.9985	15/15	33
Region 5	6765	16	AUUUAUAGCCUUUUUU		0.96 ± 0.01	0.95 – 0.97	0.02 ± 0.00	0.30 – 0.40	3/16	1.0000 / 0.9993	14/16	19
Region 6	6767	15	UAUAUGCCUUUUUUC	ORF1ab nsp3 (PLpro)	0.96 ± 0.01	0.95 – 0.97	0.02 ± 0.00	0.37 – 0.42	3/15	0.9999 / 0.9981	15/15	27
Region 7	7691	22	CAGUUUAAAAGACCAUUAAAUC		0.77 ± 0.03	0.69 – 0.83	0.16 ± 0.03	0.00 – 0.00	3/22	1.0000 / 0.9991	20/22	27
Region 8	9527	18	UCAUUCACUGUACUCUGU		0.66 ± 0.03	0.60 – 0.70	0.26 ± 0.03	0.00 – 0.00	3/18	0.9997 / 0.9945	15/18	39
Region 9	9530	17	UUCACUGUACUCUGUUU	ORF1ab nsp4	0.64 ± 0.03	0.57 – 0.68	0.28 ± 0.03	0.00 – 0.00	3/17	0.9998 / 0.9965	15/17	35
Region 10	9905	15	UACAAGUAAUUUAGU		0.75 ± 0.10	0.51 – 0.82	0.18 ± 0.10	0.00 – 0.00	2/15	0.9999 / 0.9980	15/15	20
Region 11	10010	17	CUUUACCAACCACCACA		0.75 ± 0.06	0.54 – 0.79	0.18 ± 0.06	0.00 – 0.01	3/17	0.9999 / 0.9989	15/17	47
Region 12	11536	25	UAUUGUUUUUAUGUGUUGAGUUAU		0.67 ± 0.06	0.59 – 0.78	0.25 ± 0.05	0.00 – 0.01	3/25	0.9998 / 0.9961	17/25	24
Region 13	11540	22	GUUUUUUAUGUGUUGAGUUAUU	ORF1ab nsp6	0.79 ± 0.08	0.69 – 0.92	0.15 ± 0.06	0.00 – 0.02	3/22	0.9998 / 0.9965	17/22	27
Region 14	11543	20	UUUAUGUGUUGAGUUAUUG		0.78 ± 0.08	0.69 – 0.92	0.15 ± 0.06	0.00 – 0.00	3/20	0.9998 / 0.9970	17/20	30
Region 15	11547	19	UGUGUGUUGAGUUAUGCCC		0.79 ± 0.08	0.69 – 0.92	0.15 ± 0.06	0.00 – 0.00	3/19	0.9998 / 0.9954	15/19	47
Region 16	13454	15	CAAUCGUUUUUAAAC	ORF1ab nsp11	0.96 ± 0.04	0.88 – 0.98	0.02 ± 0.02	0.07 – 0.11	3/15	0.9998 / 0.9972	14/15	27
Region 17	15141	22	CAAUAGACAGUUUCAUAAAAA	ORF1ab nsp12	0.61 ± 0.05	0.50 – 0.67	0.30 ± 0.06	0.00 – 0.00	3/22	0.9999 / 0.9986	18/22	27
Region 18	15890	15	CAAUGCUAGUUAAAC	(RdRp)	0.63 ± 0.06	0.50 – 0.68	0.29 ± 0.06	0.00 – 0.23	1/15	0.9999 / 0.9991	13/15	33
Region 19	15997	16	ACACUUUAUUGAUAAC		0.72 ± 0.03	0.65 – 0.76	0.20 ± 0.03	0.00 – 0.40	2/16	1.0000 / 0.9997	13/16	31
Region 20	17194	22	AAGGCAUUAAAAUUUGCCUA		1.00 ± 0.00	0.99 – 1.00	0.00 ± 0.00	0.00 – 0.00	3/22	0.9999 / 0.9989	16/22	27
Region 21	18032	17	CUUUACAAGCUGAAAAU	ORF1ab nsp13	0.67 ± 0.05	0.57 – 0.73	0.25 ± 0.05	0.00 – 0.00	2/17	0.9999 / 0.9978	14/17	29
Region 22	18035	15	UACAAGCUGAAAAUG	(helicase)	0.91 ± 0.10	0.54 – 0.95	0.07 ± 0.08	0.00 – 0.02	1/15	1.0000 / 0.9993	13/15	33
Region 23	18036	17	ACAAGCUGAAAAUGUAA		0.93 ± 0.09	0.57 – 0.97	0.05 ± 0.08	0.00 – 0.02	2/17	0.9998 / 0.9992	15/17	29
Region 24	20134	20	GUAAAAACACAGUUCAAUUA	ORF1ab nsp15	0.62 ± 0.05	0.52 – 0.68	0.29 ± 0.05	0.00 – 0.07	3/20	0.9998 / 0.9959	14/20	25
Region 25	20135	21	UAAAAACACAGUUCAAUUUU		0.63 ± 0.04	0.54 – 0.68	0.29 ± 0.04	0.00 – 0.07	3/21	0.9998 / 0.9967	14/21	19
Region 26	25546	17	CUUCUUGCUGUUUUUCA	ORF3a	0.94 ± 0.02	0.89 – 0.95	0.04 ± 0.01	0.00 – 0.02	1/17	0.9998 / 0.9904	16/17	35
<b>Region 27</b>	<b>27132</b>	<b>15</b>	<b>UAUAAUUAAACACA</b>	<b>M</b>	<b>0.70 ± 0.01</b>	<b>0.68 – 0.72</b>	<b>0.22 ± 0.01</b>	<b>0.65 – 0.79</b>	3/15	0.9998 / 0.9963	15/15	13
Region 28	27525	16	ACCAUUUCAUCCUCUA	ORF7a	0.99 ± 0.00	0.99 – 1.00	0.00 ± 0.00	0.00 – 0.05	3/16	0.9993 / 0.9927	14/16	38
Region 29	28402	16	AGGUUUACCCAAUAAU		1.00 ± 0.00	0.99 – 1.00	0.00 ± 0.00	0.00 – 0.81	1/16	0.9999 / 0.9985	14/16	31
Region 30	28690	15	GAUACACCCAAAAGA		0.78 ± 0.01	0.77 – 0.78	0.16 ± 0.00	0.01 – 0.52	3/33	0.9992 / 0.9879	14/15	33
Region 31	28691	20	AAUACACAAAAGAUACAU	N	0.77 ± 0.01	0.76 – 0.78	0.16 ± 0.00	0.00 – 0.52	3/20	0.9994 / 0.9883	15/20	30
Region 32	28694	18	ACACCAAAGAUACACAUU		0.77 ± 0.01	0.76 – 0.78	0.16 ± 0.00	0.01 – 0.52	3/18	0.9994 / 0.9886	16/18	33
Region 33	29075	15	UACAAGUAAACACAA		1.00 ± 0.00	1.00 – 1.00	0.00 ± 0.00	0.01 – 0.83	3/15	0.9996 / 0.9942	12/15	27

**Table S7.** Accessible and conserved regions with a loose constraint on conservation: the average sequence identity on the ~2M SARS-CoV-2 dataset is at least 0.999. The table keeps the same format as Tab. S6 and only displays new regions not included in that table.

Region	Start	Length	Sequence	Gene	Accessibility				Conservation		BLAST Match	GC (%)
					LinearTurboFold (Homologous Folding)			Single Seq.	SARSr (9)	SARS-CoV-2 (2M)		
					Average	Range	$\Delta G$ (kcal/mol)	Folding Range	# Mut. Sites	Identity / Exact		
Region 1	1094	19	UUAAAUUCCAUAUAACAAGA		0.86 ± 0.03	0.80 – 0.89	0.10 ± 0.02	0.01 – 0.10	8/19	0.9998 / 0.9963	17/19	21
Region 2	1301	19	ACUGAGAAUUUGACUAAAAG		0.75 ± 0.05	0.64 – 0.79	0.18 ± 0.04	0.00 – 0.00	8/19	0.9997 / 0.9950	14/19	32
Region 3	1359	18	UUGUUAAAAUUUUUUGUC	ORF1ab nsp2	0.75 ± 0.02	0.72 – 0.81	0.17 ± 0.02	0.09 – 0.22	7/18	0.9999 / 0.9989	16/18	17
Region 4	1420	18	CGAAUACCAUAAUGAAUC		0.94 ± 0.00	0.94 – 0.95	0.04 ± 0.00	0.01 – 0.03	7/18	0.9997 / 0.9941	13/18	33
Region 5	2550	19	UUUUACAACCAUUGAACA		0.93 ± 0.03	0.89 – 0.96	0.04 ± 0.02	0.24 – 0.31	12/19	0.9998 / 0.9971	13/19	26
Region 6	3648	15	UUCACUUCUUAAGA		0.93 ± 0.02	0.91 – 0.96	0.04 ± 0.01	0.00 – 0.03	8/15	0.9999 / 0.9980	14/15	27
Region 7	3733	19	UGACCCUUAACAUUCUUUA		0.91 ± 0.01	0.89 – 0.91	0.06 ± 0.00	0.00 – 0.01	8/19	0.9996 / 0.9928	13/19	32
Region 8	4405	17	ACAUGCAGAAAGAAACAC	ORF1ab nsp3	0.55 ± 0.02	0.51 – 0.59	0.36 ± 0.02	0.00 – 0.00	6/17	0.9999 / 0.9987	15/17	41
Region 9	4406	21	CAUUGCAGAAAGAAACGCAAA	(PLpro)	0.75 ± 0.03	0.71 – 0.80	0.18 ± 0.02	0.00 – 0.00	7/21	0.9999 / 0.9975	17/21	43
Region 10	4864	26	AAGUGUAUUUACACUAGUAUCCUA		0.85 ± 0.07	0.60 – 0.88	0.11 ± 0.06	0.00 – 0.00	14/26	0.9999 / 0.9975	16/26	27
Region 11	5773	23	UAAACAUUAACUUCUAAAGAAA		0.64 ± 0.04	0.57 – 0.70	0.28 ± 0.04	0.00 – 0.21	9/23	0.9998 / 0.9961	16/23	17
Region 12	6129	16	UUAAAGUUACAUUUUU		0.97 ± 0.04	0.79 – 0.98	0.02 ± 0.03	0.01 – 0.11	6/16	1.0000 / 0.9996	16/16	13
Region 13	6499	32	ACCAGCAAUAUAGUUUAAAAUUACAGAAG		0.61 ± 0.02	0.55 – 0.63	0.31 ± 0.03	0.00 – 0.00	15/32	0.9997 / 0.9941	19/32	25
Region 14	6622	19	GAAAACCCUUGCUACUCAU		0.95 ± 0.00	0.94 – 0.96	0.03 ± 0.00	0.00 – 0.01	8/19	0.9993 / 0.9869	15/19	42
Region 15	6697	15	UUUUCUUAACAAGU		0.96 ± 0.00	0.95 – 0.97	0.03 ± 0.00	0.00 – 0.00	11/15	0.9995 / 0.9920	14/15	20
Region 16	7010	15	GCUUUAGGUGUUUUA		0.76 ± 0.03	0.69 – 0.79	0.17 ± 0.02	0.00 – 0.01	7/15	0.9999 / 0.9982	14/15	33
Region 17	7073	20	UAUUUGAACUCUACUAAUGU		0.85 ± 0.02	0.83 – 0.88	0.10 ± 0.01	0.00 – 0.24	7/20	0.9998 / 0.9967	15/20	25
Region 18	7725	19	CUUCUUACAUJGUUGAUAG		0.76 ± 0.06	0.69 – 0.86	0.17 ± 0.05	0.00 – 0.00	8/19	0.9996 / 0.9930	15/19	37
Region 19	9336	15	UAAUUUACUUAACUA	ORF1ab nsp4	1.00 ± 0.00	1.00 – 1.00	0.00 ± 0.00	0.00 – 0.95	8/15	0.9998 / 0.9976	13/15	13
Region 20	9555	15	UUUACUUAUCUUAAC		0.89 ± 0.01	0.88 – 0.91	0.07 ± 0.01	0.00 – 0.62	9/15	0.9995 / 0.9927	13/15	27
Region 21	11629	18	UUUUUGUACUUGUUAACUU	ORF1ab nsp6	0.77 ± 0.07	0.69 – 0.88	0.16 ± 0.05	0.00 – 0.27	8/18	0.9998 / 0.9981	14/18	22
Region 22	12825	17	AUUUACAGGAUUUGAAA		0.75 ± 0.06	0.62 – 0.82	0.18 ± 0.05	0.00 – 0.01	9/17	0.9999 / 0.9991	16/17	24
Region 23	14170	16	AUAUUAAACCUUGACCA	ORF1ab nsp12 (RdRp)	0.90 ± 0.01	0.88 – 0.92	0.06 ± 0.01	0.00 – 0.27	7/16	0.9997 / 0.9948	14/16	31
Region 24	16339	18	AUAUCAACAUCACAUAAA	ORF1ab nsp13	0.96 ± 0.06	0.83 – 1.00	0.02 ± 0.04	0.00 – 0.13	5/18	0.9999 / 0.9988	15/18	22
Region 25	17651	16	UUAAAUGUUUUUAUA	(helicase)	0.96 ± 0.00	0.95 – 0.97	0.02 ± 0.00	0.00 – 0.17	5/16	0.9999 / 0.9988	15/16	6
Region 26	20652	16	AUUACAUCUAGUCA	ORF1ab nsp15	0.73 ± 0.02	0.70 – 0.77	0.19 ± 0.02	0.00 – 0.38	6/16	0.9995 / 0.9918	13/16	25
Region 27	20844	24	CUAUAUUAUGAGUUUAUACAUUU	ORF1ab nsp16	1.00 ± 0.00	1.00 – 1.00	0.00 ± 0.00	0.00 – 0.02	7/24	0.9999 / 0.9982	14/24	21
Region 28	21622	16	CAGAACUCAUUUACCC		0.81 ± 0.03	0.78 – 0.89	0.13 ± 0.02	0.00 – 0.36	15/16	0.9992 / 0.9872	13/16	44
Region 29	21922	16	UAUAACGCUCUUAUAU		0.96 ± 0.01	0.93 – 0.97	0.03 ± 0.01	0.00 – 0.00	7/16	0.9999 / 0.9988	14/16	25
Region 30	21950	15	GUCUGUGAAUUUCA		0.73 ± 0.05	0.67 – 0.82	0.19 ± 0.04	0.00 – 0.11	6/15	0.9999 / 0.9983	14/15	33
Region 31	22876	24	UAACAUCUUGAUUCUAAGGUUGG		0.94 ± 0.04	0.86 – 0.97	0.04 ± 0.02	0.00 – 0.00	23/24	0.9993 / 0.9825	16/24	33
Region 32	23031	16	UUCUUUAACAUAUA		0.98 ± 0.00	0.98 – 0.98	0.01 ± 0.00	0.00 – 0.04	14/16	0.9995 / 0.9920	13/16	25
Region 33	24058	15	CUUCAUCAACAUA	S	0.75 ± 0.13	0.51 – 0.87	0.19 ± 0.12	0.00 – 0.00	5/15	0.9999 / 0.9988	14/15	27
Region 34	24166	19	AAUGAUJGCUCAAUACACU		0.73 ± 0.06	0.61 – 0.78	0.20 ± 0.05	0.00 – 0.16	8/19	0.9999 / 0.9977	17/19	32
Region 35	24170	16	AUUGCUCAAUACACUU		0.56 ± 0.03	0.50 – 0.60	0.36 ± 0.04	0.00 – 0.08	8/16	0.9999 / 0.9978	13/16	31
Region 36	24368	17	GACUCACUUUCUCCAC		0.68 ± 0.07	0.53 – 0.74	0.24 ± 0.07	0.00 – 0.01	8/17	0.9992 / 0.9877	16/17	47
Region 37	25015	18	GUUAGAUAAAUAUUUAA		0.78 ± 0.04	0.70 – 0.83	0.15 ± 0.03	0.00 – 0.03	6/18	0.9999 / 0.9987	14/18	11
Region 38	27312	16	UAAAAUUUAUCUJAG	ORF6	0.76 ± 0.01	0.74 – 0.77	0.17 ± 0.01	0.00 – 0.19	7/16	0.9999 / 0.9977	14/16	13

**Table S8.** Accessible regions by single sequence folding (applying LinearSampling on the SARS-CoV-2 reference sequence alone). The accessibility of the corresponding regions in other 15 SARS-CoV-2 genomes are calculated for each sequence separately. Except for the region in the M gene (in bold), all accessible regions on the reference sequence are not accessible on the other sequences, and always result in a wide range of accessibilities. By contrast, LinearTurboFold is able to find regions that are accessible across all 16 SARS-CoV-2 genomes thanks to fact that consensus folding is determined across the homologous sequences (Tab. S6).

Start	Length	Sequence	Gene	Accessibility		
				Reference Sequence	SARS-CoV-2 sequences (15) Average	Range
9555	15	UUUACUCAUUCUJAC	ORF1ab	0.61	0.52 ± 0.19	0.00 – 0.62
20147	17	UCAAUUAUUAUAAGAAA	ORF1ab	0.56	0.07 ± 0.16	0.00 – 0.55
23705	16	CCCACAAUUUUACUA	S	0.71	0.55 ± 0.31	0.01 – 0.90
23985	15	AUCCAUCAAAACCAA	S	0.62	0.59 ± 0.15	0.05 – 0.72
25700	20	CCCCUUUUCUCUAUCUUUAU	S	0.97	0.17 ± 0.26	0.00 – 0.98
<b>27129</b>	<b>18</b>	<b>AACUAUAAAUAAACACA</b>	<b>M</b>	<b>0.77</b>	<b>0.76 ± 0.04</b>	<b>0.62 – 0.78</b>
28433	15	ACCGCUCUCACUCAA	N	0.55	0.27 ± 0.27	0.00 – 0.70
28691	17	AAUACACCAAAGAUCA	N	0.54	0.45 ± 0.22	0.01 – 0.67
29074	16	AUACAAUGUAACACAA	N	0.83	0.56 ± 0.40	0.01 – 0.83