LinearTurboFold: Fast Folding and Alignment for RNA Homologs with Applications to Coronavirus

Sizhen Li,[1] He Zhang,[2] Liang Zhang,[2,1] Kaibo Liu,[2,1]
Boxiang Liu,[2] David H. Mathews,[3,4,5*] Liang Huang[1,2*]

[1]School of Electrical Engineering & Computer Science,
Oregon State University, Corvallis, OR 97330, USA
[2]Baidu Research USA, Sunnyvale, CA 94089, USA
[3]Department of Biochemistry & Biophysics, [4]Center for RNA Biology,
[5]Department of Biostatistics & Computational Biology,
University of Rochester Medical Center, Rochester, NY 14642, USA

[*]To whom correspondence should be addressed;
E-mail: David_Mathews@urmc.rochester.edu, liang.huang.sh@gmail.com.

As the COVID-19 outbreak spreads, there is a growing need for an efficient tool to identify conserved RNA structures as critical targets for diagnostics and therapeutics. To address this need, we present LinearTurboFold, an algorithm that scales *linearly* with sequence length, to predict conserved structures for a set of unaligned RNA homologs. LinearTurboFold uses the same iterative refinement of structures and alignments as TurboFold, but is substantially faster than previous methods and can fold full-length coronavirus genomes without constraints on base-pairing distance. It also significantly improves structure prediction accuracy and achieves comparable alignment accuracy. On SARS-CoV-2 genomes, LinearTurboFold identifies not only conserved structures but also accessible and conserved regions as potential targets for designing efficient small-molecule drugs, antisense oligonucleotides, siRNAs, CRISPR-Cas13 gRNAs and RT-PCR primers.

Introduction

RNAs play important roles in many cellular processes ([1, 2]). To maintain their functions, secondary structures of RNA homologs are conserved across evolution ([3, 4, 5]). These conserved structures provide critical targets for diagnostics and treatments. Thus, there is a need for developing fast and accurate computational methods to identify structurally conserved regions.

Commonly, conserved structures involve compensatory base pair changes, where two positions in primary sequences mutate across evolution and still conserve a base pair, for instance, an AU or a CG pair replaces a GC pair in homologous sequences. These compensatory changes provide strong evidence for evolutionarily conserved structures ([6,7,8,9,10]). Meanwhile, they make it harder to align sequences when structures are unknown. To solve this issue, Sankoff proposed a dynamic algorithm that simultaneously predicts structures and a structural alignment for two or more sequences ([11]). The major limitation of this approach is that the algorithm runs in $O(n^{3k})$ against k sequences with the average sequence length n. Several software packages provide implementations of the Sankoff algorithm ([12, 13, 14, 15, 16, 17]) that use simplifications to reduce runtime.[1]

As an alternative, TurboFold II ([18]), an extension of TurboFold ([19]), 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 ([12, 14, 20, 21, 22]) 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 ([23, 24, 25, 26, 27, 28]) resort to local folding methods ([29, 30]) with sliding windows plus a limited pairing distance, abandoning all non-local interactions, and only consider one SARS-

[1] Besides these joint-fold-and-align algorithms, there exist two alternative approaches to homologous folding: *align-then-fold* and *fold-then-align*; see Fig. S1 for details.

CoV-2 genome (Fig. 1B–C), ignoring homology signals. To address this challenge, we design a linearized version of TurboFold II, *LinearTurboFold* (Fig. 1A), which is a global homologous folding algorithm that scale 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 on secondary structure prediction accuracy as well as an alignment accuracy comparable to or higher than all benchmarks.

Over a group of 25 SARS-CoV-2 and SARS-related homologous genomes, LinearTurboFold predictions are close to the canonical structures (*31*) and structures modeled with the aid of experimental data (*24*, *25*, *26*) for several well-studied regions. Thanks to global rather than local folding, LinearTurboFold discovers a long-range interaction involving 5' and 3' UTRs ($\sim$29,800 *nt* apart), which is consistent with recent purely experimental work (*27*), and yet is out of reach for local folding methods used by existing studies (Fig. 1B–C). In short, our *in silico* method of folding multiple homologs can achieve results similar to, and sometimes more accurate than, experimentally-guided models for one genome. Moreover, LinearTurboFold identifies conserved structures supported by compensatory mutations, which are potential targets for small molecule drugs (*32*) and antisense oligonucleotides (ASOs) (*28*). We further identify regions that are (a) sequence-level conserved, (b) at least 15 *nt* long, and (c) accessible (i.e., likely to be completely unpaired) as potential targets for ASOs (*33*), small interfering RNA (siRNA) (*34*), CRISPR-Cas13 guide RNA (gRNA) (*35*) and reverse transcription polymerase chain reaction (RT-PCR) primers (*36*).

LinearTurboFold is a general technique that can also be applied to other RNA viruses (e.g., influenza, Ebola, HIV, Zika, etc.) for full-length genome studies.

Results

The framework of LinearTurboFold has two major aspects (Fig. 1A): linearized structure-aware pairwise alignment estimation (module **1**); and linearized homolog-aware structure prediction (module **2**). LinearTurboFold iteratively refines alignments and structure predictions, specifically, updating pairwise alignment probabilities by incorporating predicted base-pairing probabilities (from module **2**) to form structural alignments, and modifying base-pairing probabilities for each sequence by integrating the structural information from homologous sequences via the estimated alignment probabilities (from module **1**) to detect conserved structures. After several iterations, LinearTurboFold generates the final multiple sequence alignment (MSA) based on the latest pairwise alignment probabilities (module **3**) and predicts secondary structures using the latest pairing probabilities (module **4**).

LinearTurboFold achieves linear time regarding sequence length with two major linearized modules: our recent work LinearPartition (*37*) (Fig. 1A module **2**), which approximates the RNA partition function (*38*) and base pairing probabilities in linear time, and a novel algorithm LinearAlignment (module **1**). LinearAlignment aligns two sequences by Hidden Markov Model (HMM) in linear time by applying the same beam search heuristic (*39*) used by LinearPartition. Finally, LinearTurboFold assembles the secondary structure from the final base pairing probabilities using an accurate and linear-time method named ThreshKnot (*40*) (module **4**). LinearTurboFold also integrates a linear-time stochastic sampling algorithm named LinearSampling (*41*) (module **5**), which can independently sample structures according to the homolog-aware partition functions and then calculate the probability of being unpaired for regions, which is an important property in siRNA sequence design (*34*). So overall, the end-to-end runtime of LinearTurboFold scales linearly with sequence length (see **Methods** for more details).

Scalability and Accuracy

To evaluate the efficiency of LinearTurboFold against the sequence length, we collected a dataset consisting of seven families of RNAs with sequence length ranging from 210 *nt* to 30,000 *nt*, including five families from the RNAstralign dataset plus 23S ribosomal RNA, HIV and SARS-CoV genomes, and each family has five homologous sequences (see **Methods** for more details). Fig. 2A compares the running times of LinearTurboFold with TurboFold II and two Sankoff-style simultaneous folding and alignment algorithms, LocARNA and MXSCARNA. Clearly, LinearTurboFold scales linearly with sequence length n, and is substantially faster than other benchmarks which scale superlinearly. The linearization in LinearTurboFold brought orders of magnitude speedup over the cubic-time TurboFold II, taking only 12 minutes on the HIV family (average length 9,686 *nt*) while TurboFold II takes 3.1 days (372× speedup). More importantly, LinearTurboFold takes only 40 minutes on five SARS-CoV sequences while all other benchmarks fail to scale. Regarding the memory usage (Fig. 2B), LinearTurboFold costs linear memory space with sequence length, while other benchmarks use quadratic or more memory. In Fig. 2C–D, we also demonstrate that the runtime and memory usage against the number of homologs ($k = 5 \sim 20$), using homologs of 16S rRNAs about 1,500 *nt* in length. The apparent complexity against the group size of LinearTurboFold is higher than TurboFold II because the cubic-time partition function calculation, which dominates the runtime of TurboFold II, has been linearized in LinearTurboFold by LinearPartition (Fig. S5C).

We next compare the accuracies of predicted secondary structures and MSAs between LinearTurboFold and several benchmark methods. Besides Sankoff-style LocARNA and MXSCARNA, we also consider three types of negative controls: (a) single sequence folding (partition function-based): Vienna RNAfold (*30*) (-p mode) and LinearPartition; (b) sequence-only alignment: MAFFT (*21*) and LinearAlignment (a standalone version without structural information); and (c) an align-then-fold method that predicts consensus structures from MSAs (Fig. S1): MAFFT + RNAalifold (*20*).

For secondary structure prediction, LinearTurboFold, TurboFold II and LocARNA achieve higher F1 scores than single sequence folding methods (Vienna RNAfold and LinearPartition) (Fig. 2E), which demonstrates folding with homology information performs better than folding sequences separately. Overall, LinearTurboFold performs significantly better than all the other benchmarks on structure prediction. For the accuracy of MSAs (Fig. 2F), the structural alignments from LinearTurboFold obtain higher accuracies than sequence-only alignments (LinearAlignment and MAFFT) on all four families, especially for families with low sequence identity. On average, LinearTurboFold performs comparably with TurboFold II and significantly better than other benchmarks on alignments. We also note that the structure prediction accuracy of the align-then-fold approach (MAFFT + RNAalifold) depends heavily on the alignment accuracy, and is the worst when the sequence identity is low (e.g., SRP RNA) and the best when the sequence identity is high (e.g., 16S rRNA) (Fig. 2E–F).

Highly Conserved Structures in SARS-CoV-2 and SARS-related Betacoronaviruses

RNA sequences with conserved secondary structures play vital biological roles and provide potential targets. The current COVID-19 outbreak raises an emergent requirement of identifying potential targets for diagnostics and therapeutics. Given the strong scalability and high accuracy, we used LinearTurboFold on a group of full-length SARS-CoV-2 and SARS-related (SARSr) genomes to obtain global structures and identify highly conserved structural regions.

We used a greedy algorithm to select the 16 most diverse genomes from all the valid SARS-CoV-2 genomes submitted to the Global Initiative on Sharing Avian Influenza Data (GISAID) (*42*) up to December 2020 (**Methods**). We further extended the group by adding 9 SARS-related homologous genomes (5 human SARS-CoV-1 and 4 bat coronaviruses). In total, we built a dataset of 25 full-length genomes consisting of 16 SARS-CoV-2 and 9 SARS-related sequences (Tab. S2). The average pairwise sequence identities of the 16 SARS-CoV-2 and the total 25 genomes are 99.9% and 89.6%, respectively. LinearTurboFold takes about 13 hours and 43 GB on the 25 genomes.

To evaluate the reliability of LinearTurboFold predictions, we first compare them with the Huston *et al.*'s SHAPE-guided models (*24*) for regions with well-characterized structures across betacoronaviruses. For the extended 5' and 3' untranslated regions (UTRs), LinearTurboFold's predictions are close to the SHAPE-guided structures (Fig. 3A–B), i.e., both identify the stem-loops (SLs) 1–2 and 4–7 in the extended 5' UTR, and the bulged stem-loop (BSL), SL1, and a long bulge stem for the hypervariable region (HVR) including the stem-loop II-like motif (S2M) in the 3' UTR. Interestingly, in our model, the high unpaired probability of the stem in the SL4b indicates the possibility of being single-stranded as an alternative structure, which is supported by experimental studies (*28*, *25*). In addition, the compensatory mutations LinearTurboFold found in UTRs strongly support the evolutionary conservation of structures (Fig. 3A).

The most important difference between LinearTurboFold's prediction and Huston *et al.*'s experimentally-guided model is that LinearTurboFold discovers an end-to-end interaction (29.8 kilobases apart) between the 5' UTR (SL3, 60-82 *nt*) and the 3' UTR (final region, 29845-29868 *nt*), which fold locally by themselves in Huston *et al.*'s model. Interestingly, this 5'-3' interaction matches *exactly* with the one discovered by the purely experimental work of Ziv *et al.* (*43*) using the COMRADES technique to capture long-range base-pairing interactions (Fig. 3C). These end-to-end interactions have been well established by theoretical and experimental studies (*44*, *45*, *46*) to be common in natural RNAs, but are far beyond the reaches of local folding methods used in existing studies on SARS-CoV-2 secondary structures (*24*,*25*,*26*,*27*). By contrast, LinearTurboFold predicts secondary structures globally without any limit on window size or base-pairing distance, enabling it to discover long-distance interactions across the whole genome. The similarity between our predictions and the experimental work shows that our *in silico* method of folding multiple homologs can achieve results similar to, if not more accurate than, those experimentally-guided single-genome prediction. We also observed that LinearPartition, as a single sequence folding method, can also predict a long-range interaction between 5' and 3' UTRs, but it involves SL2 instead of SL3 of the 5' UTR (Fig. 3A), which indicates that the homologous information assists to adjust the positions of base pairs to be conserved in LinearTurboFold. Additionally, the align-then-fold approach (MAFFT + RNAalifold) fails to predict such long-range interactions (Fig. S6B).

The frameshifiting stimulation element (FSE) is another well-characterized region. For an extended FSE region, the LinearTurboFold prediction consists of two substructures (Fig. 4A): the 5' part 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 (*24*, *25*, *27*) (Fig. 4B–D). In the attenuator hairpin, the small internal loop motif (UU) was previously selected as a small molecule binder which stabilizes the folded state of the attenuator hairpin and impairs frameshifting (*32*). 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 (*31*) (Fig. 4E). Our prediction is closer to the canonical structure compared to the experimentally-guided models (*24*, *25*, *27*) (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 for the FSE region were estimated for specific local regions. As a result, the models are sensitive to the context and region boundaries (*27*, *24*, *47*) (see Fig. S7D–F for alternative structures of Fig. 4B–D with different regions). LinearTurboFold, by contrast, does not suffer from this problem by virtue of global folding without local windows. Besides SARS-CoV-2, we notice that the estimated structure of the SARS-CoV-1 reference sequence (Fig. 4F) from LinearTurboFold is similar to SARS-CoV-2 (Fig. 4A), which is consistent with the observation that the structure of the FSE region is highly conserved among betacoronaviruses (*31*). Finally, as negative controls, both the single sequence folding algorithm (LinearPartition in Fig. 4G) and the align-then-fold method (RNAalifold in Fig. S7G) predict quite different structures compared with the LinearTurboFold prediction (Fig. 4A) (39%/61% of pairs from the LinearTurboFold model are not found by LinearPartition/RNAalifold, respectively).

In addition to the well-studied UTRs and FSE regions, LinearTurboFold discovers 50 conserved structures with identical structures among 25 genomes, and 26 regions are novel compared to previous studies (*23*, *24*) (Fig. 4H and Tab. S4), which might be potential targets for small-molecule drugs (*32*) and antisense oligonucleotides (*28*, *48*). LinearTurboFold also recovers fully conserved base pairs with compensatory mutations (Tab. S3), which imply highly conserved structural regions whose functions might not have been explored.

Highly Accessible and Conserved Regions in SARS-CoV-2 and SARS-related Betacoronaviruses

Studies show that the siRNA silencing efficiency, ASOs inhibitory efficacy, CRISPR-Cas13 knockdown efficiency and RT-PCR testing efficiency all correlate with the target *accessibility* (*34*,*35*,*36*,*49*), which is the probability of a target site being fully unpaired. To get unstructured regions, Rangan *et al.* (*23*) imposed a threshold on unpaired probabilities of each position, which is not a truly correct method because the unpaired probabilities are dependent. By contrast, the widely-used stochastic sampling algorithm (*50*, *41*) builds a representative ensemble of structures by sampling independent secondary structures according to their probabilities in the Boltzmann distribution. Thus the accessibility for a region can be approximated as the fraction of sampled structures in which the region is single-stranded. LinearTurboFold utilized LinearSampling (*41*) to generate 10,000 independent structures for each genome according to the modified partition functions after the iterative refinement (Fig. 1A module **5**), and calculated accessibilities for regions at least 15 *nt* long. We then identify *accessible regions* with at least 0.5 accessibility among all 16 SARS-CoV-2 genomes (Fig. 5A–B).

In addition to accessibility, sequence conservation is another critical aspect for efficient therapeutic and diagnostic target sites. We further identify *accessible and conserved regions* that are not only structurally accessible among SARS-CoV-2 genomes, but also fully conserved among SARS-CoV-2 genomes with at most one mutation at each position across SARS-related genomes (Fig. 5C). These regions are less likely to accumulate mutations in the future. Finally, we identified 35 accessible and conserved regions (Fig. 5G and Tab. S5). Because the nucleotide content and specificity are also key factors influencing siRNA efficient (*51*), we searched BLAST against the human mRNA dataset for these regions and calculated the GC content (Tab. S5). Among these regions, region 16 corresponds to the internal loop containing the slippery site in the extended FSE region, and it is conserved at both structural and sequence levels (Fig. 5D and 5H). Region 29 in the ORF3a gene is fully conserved among all the 25 genomes with average accessibility 0.936 (Fig. 5D). Besides SARS-CoV-2 genomes, the SARS-related genomes such as the SARS-CoV-1 reference sequence (NC_004718.3) and a bat coronavirus (BCoV, MG772934.1) also form similar structures around the slippery site (Fig. 5A). To investigate if the the mutations are sensitive to the sampled SARS-CoV-2 genomes, we further checked the conservation of these regions among a dataset including 257,672 valid genomes submitted to GISAID up to December 2020, and most of these regions are still highly conserved[2] (Tab. S5). The mutations of new lineages of SARS-CoV-2 in South African, Brazil[3] and India[4] are outside of these predicted regions, which implies that the sequence conservation constraint imposed on SARS-related genomes is helpful in selecting evolutionarily conserved regions.

We also designed a negative control by analyzing the SARS-CoV-2 reference sequence alone,

[2] the fraction of valid genomes in which the whole region is identical.

[3] https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html

[4] https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html

which can also obtain some accessible regions. However, these regions are not structurally conserved among the other 15 SARS-CoV-2 genomes, resulting in vastly different accessibilities, except for one region in the M gene (Tab. S6). The reason behind this is that even with a high sequence identity (over 99.9%), single sequence folding algorithms still predict greatly dissimilar structures for the SARS-CoV-2 genomes (Fig. 5E–F). Both regions (in nsp11 and N genes) are fully conserved among the 16 SARS-CoV-2 genomes, yet they still fold into vastly different structures due to mutations outside the regions; as a result, the accessibilities are either low (nsp11) or in a wide range (N) (Fig. 5D). Conversely, addressing this by folding each sequence with proclivity of base pairing inferred from all homologous sequences, LinearTurboFold structure predictions are more consistent with each other and thus can detect conserved structures (Fig. 5A–B).

Summary

We have presented LinearTurboFold, an end-to-end linear-time algorithm for structural alignment and conserved structure prediction of RNA homologs, which is the first joint-fold-and-align algorithm to scale to full-length SARS-CoV-2 genomes without imposing any constraints on base-pairing distance. We also demonstrate that LinearTurboFold leads to significant improvement on secondary structure prediction accuracy as well as an alignment accuracy comparable to or higher than all benchmarks.

Unlike existing work using local folding workarounds, LinearTurboFold enables unprecedented global structural analysis on the SARS-CoV-2 genomes; in particular, it can capture long-range interactions, especially the one between 5' and 3' UTRs across the whole genome, which matches perfectly with a recent purely experiment work. Over a group of 25 SARS-CoV-2 and SARS-related homologs, LinearTurboFold identifies not only conserved structures supported by compensatory mutations and experimental studies, but also accessible and conserved regions as vital targets for designing efficient small-molecule drugs, siRNAs, ASOs, CRISPR-Cas13 gRNAs and RT-PCR primers. LinearTurboFold is widely applicable to the analysis of other RNA viruses (influenza, Ebola, HIV, Zika, etc.) and full-length genome analysis.

References

1. S. R. Eddy., *Nature Reviews Genetics* **2**, 919 (2001).
2. J. A. Doudna, T. R. Cech, *Nature* **418**, 222 (2002).
3. E. P. Nawrocki, S. R. Eddy, *Bioinformatics* **29**, 2933 (2013).
4. E. A. Brown, H. Zhang, L.-H. Ping, S. M. Lemon, *Nucleic Acids Research* **20**, 5041 (1992).
5. J. Ritz, J. S. Martin, A. Laederach, *PLoS Computational Biology* **9**, e1003152 (2013).
6. E. Rivas, J. Clements, S. R. Eddy, *Bioinformatics* **36**, 3072 (2020).
7. R. W. Holley, *et al.*, *Science* pp. 1462–1465 (1965).
8. H. F. Noller, *et al.*, *Nucleic Acids Research* **9**, 6167 (1981).
9. N. R. Pace, D. K. Smith, G. J. Olsen, B. D. James, *Gene* **82**, 65 (1989).
10. K. Williams, D. Bartel, *RNA* **2**, 1306 (1996).
11. D. Sankoff, *SIAM Journal on Applied Mathematics* **45**, 810– (1985).
12. S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, R. Backofen, *PLoS Computational Biology* **3**, e65 (2007).
13. J. H. Havgaard, E. Torarinsson, J. Gorodkin, *PLoS Computational Biology* **3**, 1896–1908 (2007).
14. Y. Tabei, H. Kiryu, T. Kin, K. Asai, *BMC Bioinformatics* **9**, 33 (2008).
15. Z. Xu, D. H. Mathews, *Bioinformatics* **27**, 626 (2011).
16. D. H. Mathews, D. H. Turner, *Journal of Molecular Biology* **317**, 191 (2002).
17. K. Sato, Y. Kato, T. Akutsu, K. Asai, Y. Sakakibara, *Bioinformatics* **28**, 3218 (2012).
18. Z. Tan, Y. Fu, G. Sharma, D. H. Mathews, *Nucleic Acids Research* **45**, 11570 (2017).

19. A. O. Harmanci, G. Sharma, D. H. Mathews, *BMC Bioinformatics* **12**, 108 (2011).

20. S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, P. F. Stadler, *BMC Bioinformatics* **9**, 1 (2008).

21. K. Katoh, D. M. Standley, *Molecular Biology and Evolution* **30**, 772 (2013).

22. C. B. Do, M. S. Mahabhashyam, M. Brudno, S. Batzoglou, *Genome Research* **15**, 330 (2005).

23. R. Rangan, *et al.*, *RNA* **26**, 937 (2020).

24. N. C. Huston, *et al.*, *Molecular cell* **81**, 584 (2021).

25. I. Manfredonia, *et al.*, *Nucleic Acids Research* **48**, 12436 (2020).

26. C. Iserman, *et al.*, *Molecular cell* **80**, 1078 (2020).

27. T. C. Lan, *et al.*, *BioRxiv* (2020).

28. L. Sun, *et al.*, *Cell* **184**, 1865 (2021).

29. J. S. Reuter, D. H. Mathews, *BMC Bioinformatics* **11**, 1 (2010).

30. R. Lorenz, *et al.*, *Algorithms for Molecular Biology* **6**, 1 (2011).

31. J. A. Kelly, *et al.*, *Journal of Biological Chemistry* **295**, 10741 (2020).

32. H. S. Haniff, *et al.*, *ACS Central Science* **6**, 1713 (2020).

33. Z. J. Lu, D. H. Mathews, *Nucleic Acids Research* **36**, 3738 (2008).

34. S. Schubert, A. Grünweller, V. A. Erdmann, J. Kurreck, *Journal of Molecular Biology* **348**, 883 (2005).

35. O. O. Abudayyeh, *et al.*, *Nature* **550**, 280 (2017).

36. I. Peters, C. Helps, E. Hall, M. Day, *Journal of Immunological Methods* **286**, 203 (2004).

37. H. Zhang, L. Zhang, D. H. Mathews, L. Huang, *Bioinformatics* **36**, i258 (2020).

38. J. S. McCaskill, *Biopolymers* **29**, 11105 (1990).

39. L. Huang, K. Sagae, *Proceedings of ACL 2010* (ACL, Uppsala, Sweden, 2010), p. 1077–1086.

40. L. Zhang, H. Zhang, D. H. Mathews, L. Huang, *BioRxiv* (2019).

41. H. Zhang, L. Zhang, S. Li, D. Mathews, L. Huang, *BioRxiv* (2020).

42. S. Elbe, G. Buckland-Merrett, *Global Challenges* **1**, 33 (2017).

43. O. Ziv, *et al.*, *Molecular cell* **80**, 1067 (2020).

44. M. G. Seetin, D. H. Mathews, *Bacterial Regulatory RNA* (Springer, 2012), pp. 99–122.

45. T. J. Li, C. M. Reidys, *Bulletin of Mathematical Biology* **80**, 1514 (2018).

46. W.-J. C. Lai, *et al.*, *Nature Communications* **9**, 1 (2018).

47. R. Rangan, *et al.*, *Nucleic Acids Research* **49**, 3092 (2021).

48. V. Lulla, *et al.*, *BioRxiv* pp. 2020–09 (2021).

49. Z. J. Lu, D. H. Mathews, *Nucleic Acids Research* **36**, 640 (2008).

50. Y. Ding, C. E. Lawrence, *Nucleic Acids Research* **31**, 7280 (2003).

51. E. Fakhr, F. Zare, L. Teimoori-Toolabi, *Cancer gene therapy* **23**, 73 (2016).

52. A. O. Harmanci, G. Sharma, D. H. Mathews, *BMC Bioinformatics* **8**, 130 (2007).

53. R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge University Press, 1998).

54. I. L. Hofacker, S. H. Bernhart, P. F. Stadler, *Bioinformatics* **20**, 2222 (2004).

55. S. Bellaousov, D. H. Mathews, *RNA* **16**, 1870 (2010).

56. J. J. Cannone, *et al.*, *BMC Bioinformatics* **3**, 2 (2002).

57. C. Ceraolo, F. M. Giorgi, *Journal of Medical Virology* **92**, 522 (2020).

58. Y. Tabei, K. Tsuda, T. Kin, K. Asai, *Bioinformatics* **22**, 1723 (2006).

59. N. Aghaeepour, H. H. Hoos, *BMC Bioinformatics* **14**, 139 (2013).

60. F. Wu, *et al.*, *Nature* **579**, 265 (2020).

61. R. Madhugiri, M. Fricke, M. Marz, J. Ziebuhr, *Advances in Virus Research* (Elsevier, 2016), vol. 96, pp. 127–163.

62. E. Van Den Born, C. C. Posthuma, A. P. Gultyaev, E. J. Snijder, *Journal of Virology* **79**, 6312 (2005).

63. E. P. Plant, J. D. Dinman, *Frontiers in Bioscience: A Journal and Virtual Library* **13**, 4873 (2008).

64. S. J. Goebel, B. Hsue, T. F. Dombrowski, P. S. Masters, *Journal of Virology* **78**, 669 (2004).

65. S. J. Goebel, T. B. Miller, C. J. Bennett, K. A. Bernard, P. S. Masters, *Journal of Virology* **81**, 1274 (2007).

66. P. Liu, D. Yang, K. Carter, F. Masud, J. L. Leibowitz, *Virology* **443**, 40 (2013).

67. M. P. Robertson, *et al.*, *PLoS Biology* **3**, e5 (2004).

68. P. P. Gardner, R. Giegerich, *BMC Bioinformatics* **5**, 1 (2004).

69. M. Hochsmann, T. Toller, R. Giegerich, S. Kurtz, *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003* (IEEE, 2003), pp. 159–168.

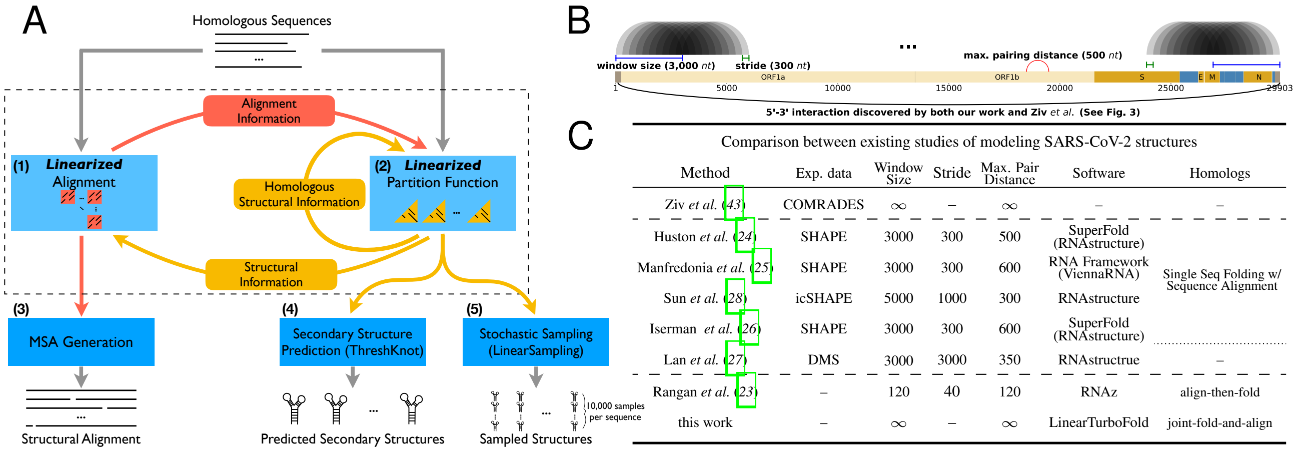
70. D. H. Mathews, *RNA* **10**, 1178 (2004).

Acknowledgments

Authors contributions: L.H. and D.H.M. conceived the idea and directed the project. S.L., H.Z., L.H., and D.H.M. designed the algorithm; S.L. implemented it. D.H.M. guided the evaluation that S.L. and L.Z. carried out. S.L. and H.Z. wrote the manuscript; L.H., and D.H.M. revised it. L.K. made the webserver. B.L. guided the SARS-CoV-2 experiment.

Competing interests: The authors declare no conflict of interest.

Figures


A
Homologous Sequences
...
Alignment Information
(1) Linearized Alignment
Homologous Structural Information
(2) Linearized Partition Function
...
Structural Information
(3)
MSA Generation
(4)
Secondary Structure Prediction (ThreshKnot)
(5)
Stochastic Sampling (LinearSampling)
...
Structural Alignment
...
Predicted Secondary Structures
...
10,000 samples per sequence
Sampled Structures
B
...
window size (3,000 nt)
stride (300 nt)
max. pairing distance (500 nt)
ORF1a
ORF1b
S
E
M
N
1
5000
10000
15000
20000
25000
29903
5'-3' interaction discovered by both our work and Ziv et al. (See Fig. 3)
C

Comparison between existing studies of modeling SARS-CoV-2 structures

Method	Exp. data	Window Size	Stride	Max. Pair Distance	Software	Homologs
Ziv *et al.* (43)	COMRADES	∞	–	∞	–	–
Huston *et al.* (24)	SHAPE	3000	300	500	SuperFold (RNAstructure)	Single Seq Folding w/ Sequence Alignment
Manfredonia *et al.* (25)	SHAPE	3000	300	600	RNA Framework (ViennaRNA)	
Sun *et al.* (28)	icSHAPE	5000	1000	300	RNAstructure	
Iserman *et al.* (26)	SHAPE	3000	300	600	SuperFold (RNAstructure)	
Lan *et al.* (27)	DMS	3000	3000	350	RNAstructrue	–
Rangan *et al.* (23)	–	120	40	120	RNAz	align-then-fold
this work	–	∞	–	∞	LinearTurboFold	joint-fold-and-align

Figure 1: **A**: The LinearTurboFold framework. Like TurboFold, LinearTurboFold also takes multiple unaligned homologous sequences as input and then outputs a multiple sequence alignment and structures for each sequence, but unlike TurboFold, it employs two linearizations to ensure linear runtime: a *linearized* alignment computation (module **1**) to predict posterior co-incidence probabilities (red squares) for all pairs of sequences and a *linearized* partition function computation (module **2**) to estimate base-pairing probabilities (yellow triangles) for all the sequences. These two modules take advantage of information from each other and iteratively refine predictions (see Fig. S2 for details). After several iterations, module **3** generates the final multiple sequence alignments, and module **4** predicts secondary structures. Module **5** is an optional output to stochastically sample structures. **B–C**: Most prior studies (expect for a 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.* Some work also used homologous sequences to identify conserved structures, but they only predicted structures for one genome and utilized sequence alignments to extract mutations. By contrast, LinearTurboFold is a global folding method without any limitations on sequence length or paring 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**, Fig. 3).

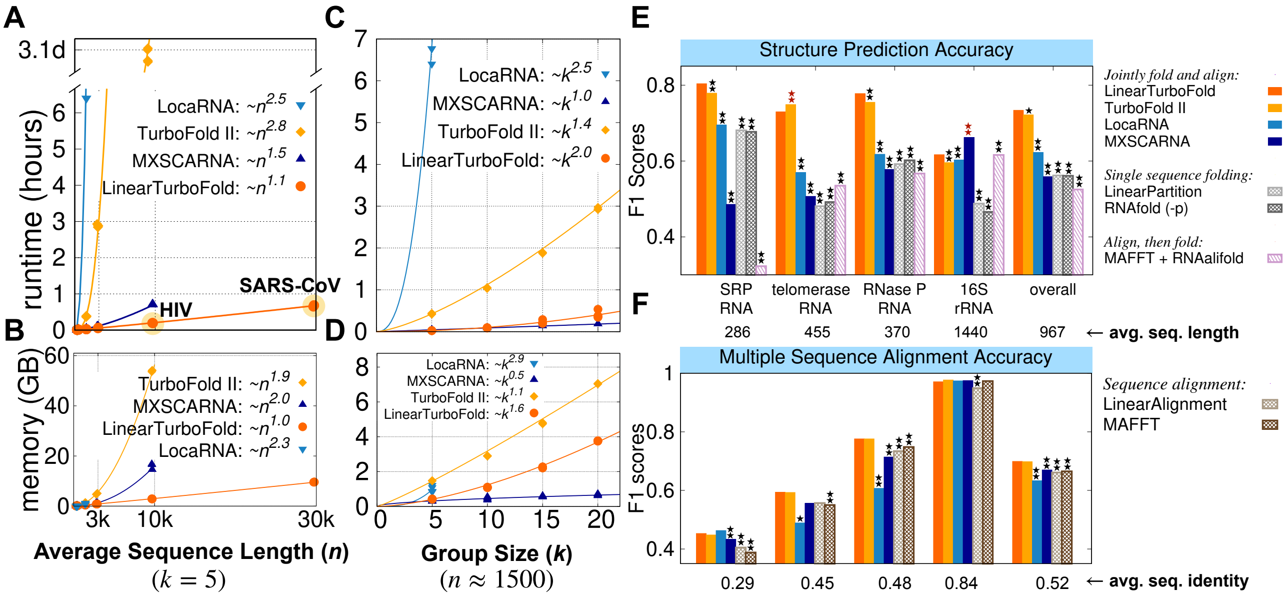


Figure 2: End-to-end Scalability and Accuracy Comparisons. **A–B**: End-to-end runtime and memory usage comparisons between benchmarks and LinearTurboFold against the sequence length. **C–D**: End-to-end runtime and memory usage comparisons against the group size. LinearTurboFold is the first joint-fold-and-align algorithm to scale to full-length coronavirus genomes (∼30,000 *nt*) due to 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 ($\overset{\star}{\star}$) 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).

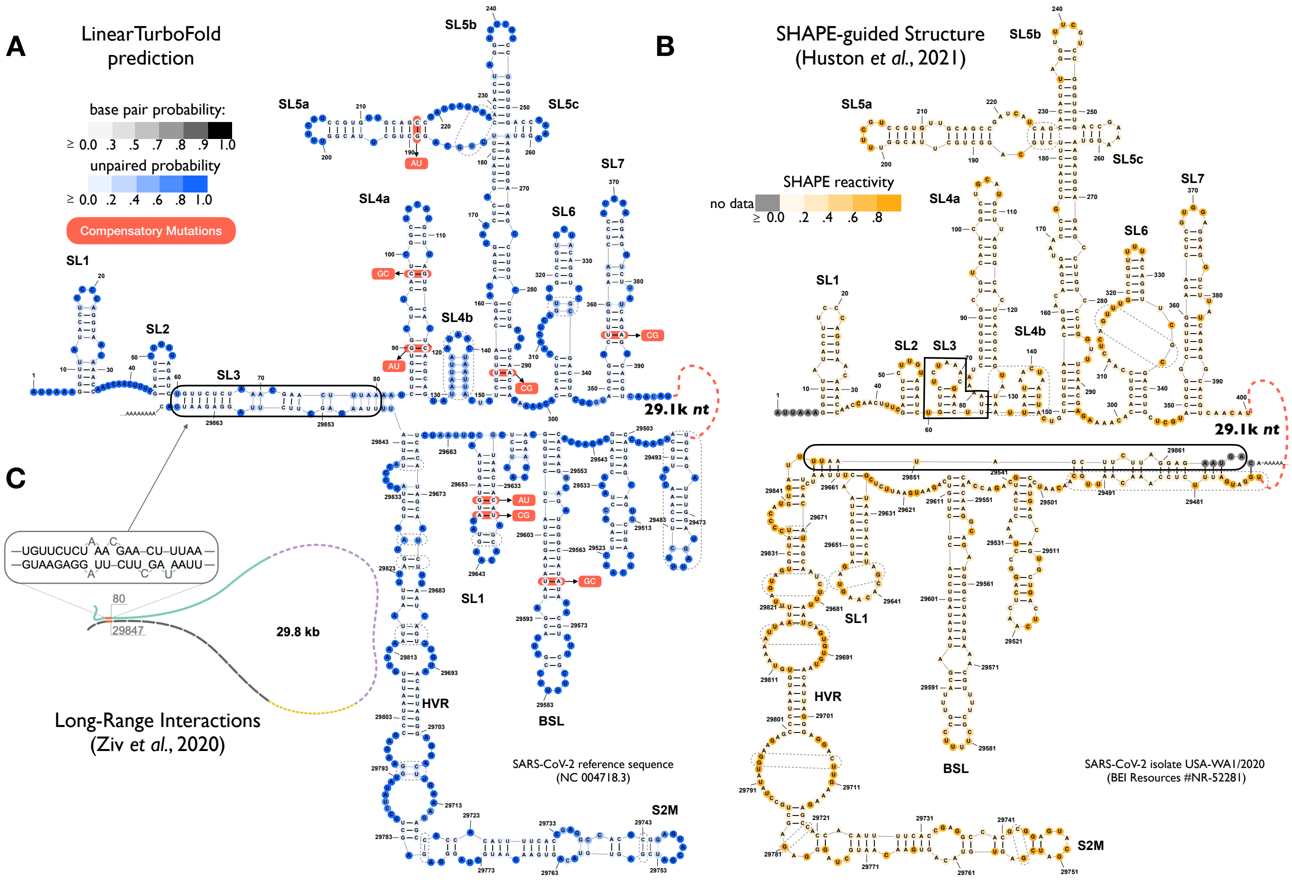

A
LinearTurboFold prediction
base pair probability:
≥ 0.0 .3 .5 .7 .8 .9 1.0
unpaired probability
≥ 0.0 .2 .4 .6 .8 1.0
Compensatory Mutations
SL1
SL2
SL3
SL4a
SL4b
SL5a
SL5b
SL5c
SL6
SL7
29.1k nt
HVR
BSL
S2M
SARS-CoV-2 reference sequence (NC 004718.3)
B
SHAPE-guided Structure (Huston et al., 2021)
SHAPE reactivity
no data
≥ 0.0 .2 .4 .6 .8
SARS-CoV-2 isolate USA-WA1/2020 (BEI Resources #NR-52281)
C
—UGUUCUCU AA GAA–CU–UUAA—
—GUAAGAGG UU–CUU GA AAUU—
80
29847
29.8 kb
Long-Range Interactions (Ziv et al., 2020)

Figure 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.* (24) (window size 3000 *nt* sliding by 300 *nt* with maximum pairing distance 500 *nt*). The nucleotides are colored by SHAPE reactivities. Dash boxes circle 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 (43) (**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. S6), the align-then-fold (RNAalifold) 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

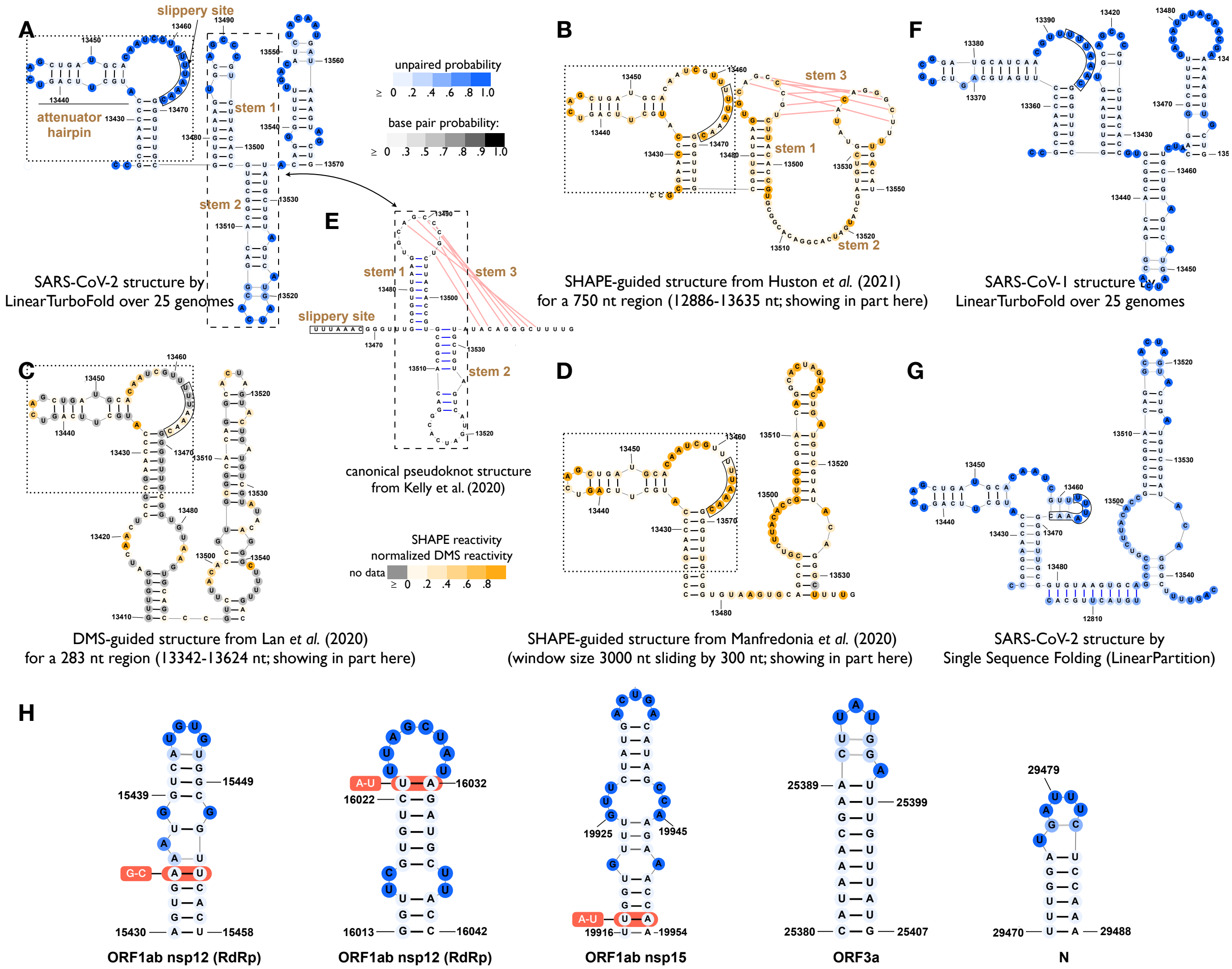
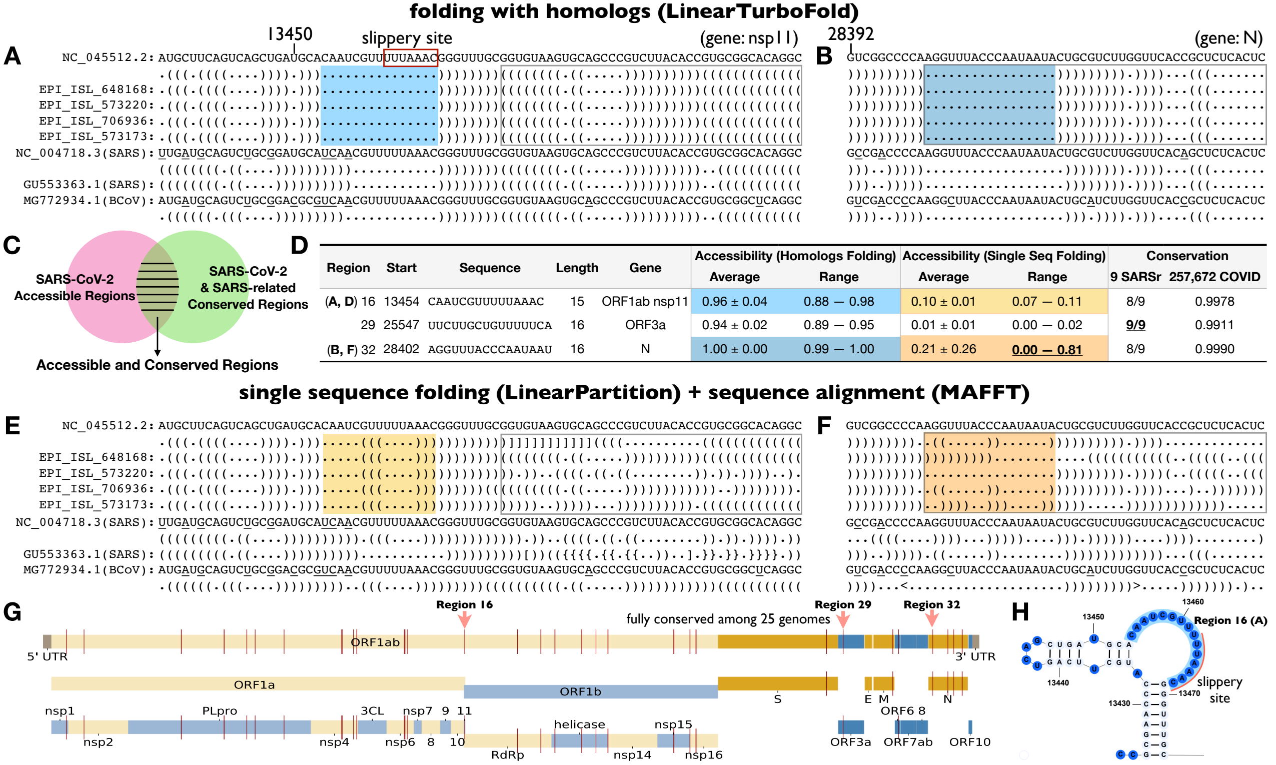

A
slippery site
attenuator hairpin
stem 1
stem 2
SARS-CoV-2 structure by LinearTurboFold over 25 genomes
unpaired probability
≥ 0 .2 .4 .6 .8 1.0
base pair probability:
≥ 0 .3 .5 .7 .8 .9 1.0
B
stem 3
stem 1
stem 2
SHAPE-guided structure from Huston et al. (2021) for a 750 nt region (12886-13635 nt; showing in part here)
F
SARS-CoV-1 structure by LinearTurboFold over 25 genomes
E
stem 1
stem 3
slippery site
stem 2
canonical pseudoknot structure from Kelly et al. (2020)
C
DMS-guided structure from Lan et al. (2020) for a 283 nt region (13342-13624 nt; showing in part here)
SHAPE reactivity
normalized DMS reactivity
no data
≥ 0 .2 .4 .6 .8
D
SHAPE-guided structure from Manfredonia et al. (2020) (window size 3000 nt sliding by 300 nt; showing in part here)
G
SARS-CoV-2 structure by Single Sequence Folding (LinearPartition)
H
ORF1ab nsp12 (RdRp)
ORF1ab nsp12 (RdRp)
ORF1ab nsp15
ORF3a
N

Figure 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 (24, 27, 25), which are sensitive to the context and region boundaries due to the use of local folding methods (Fig. S7). **E**: The canonical pseudoknot structure by the comparative analysis between SARS-CoV-1 and SARS-CoV-2 genomes (31). 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. S7G). **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 (23, 24).



Region	Start	Sequence	Length	Gene	Accessibility (Homologs Folding)		Accessibility (Single Seq Folding)		Conservation	
					Average	Range	Average	Range	9 SARSr	257,672 COVID
(**A, D**) 16	13454	CAAUCGUUUUUAAAC	15	ORF1ab nsp11	0.96 ± 0.04	0.88 — 0.98	0.10 ± 0.01	0.07 — 0.11	8/9	0.9978
29	25547	UUCUUGCUGUUUUUCA	16	ORF3a	0.94 ± 0.02	0.89 — 0.95	0.01 ± 0.01	0.00 — 0.02	**9/9**	0.9911
(**B, F**) 32	28402	AGGUUUACCCAAUAAU	16	N	1.00 ± 0.00	0.99 — 1.00	0.21 ± 0.26	**0.00 — 0.81**	8/9	0.9990

Figure 5: An illustration of accessible and conserved regions that LinearTurboFold identifies. **A–B**: Identified structurally-conserved accessible regions by LinearTurboFold with the help of considering alignment and folding simultaneously. The regions at least 15 *nt* long with accessibility of at least 0.5 among all the 16 SARS-CoV-2 genomes are shaded on blue background. Structures are encoded in dot-bracket notation. "(" and ")" indicates nucleotides pairing in the 3' and 5' direction, respectively. "." indicates an unpaired nucleotide. The positions with mutations compared to the SARS-CoV-2 reference sequence among three different subfamilies (SARS-CoV-2, SARS-CoV-1 and BCoV) are underlined. **C**: Accessible and conserved regions are not only *accessible* among SARS-CoV-2 genomes (pink circle) but also *conserved* (at sequence level) among both SARS-CoV-2 and SARS-related genomes (green circle). **D**: Three examples out of 35 accessible and conserved regions found by LinearTurboFold. Region 16 and Region 32 correspond to the accessible regions in **A** and **B**, respectively. Region 16 is also the long internal loop including the slippery site in the FSE region (**H**). Region 29 is fully conserved among all 25 genomes. **E–F**: Single sequence folding algorithms predict greatly different structures even if the sequence identities are high (grey boxes). These two regions, fully conserved among SARS-CoV-2 genomes, still fold into different structures due to mutations outside the regions. **G**: The positions of these 35 regions (red bars) across the whole genome (see Tab. S5 for more details). All the accessible and conserved regions are potential targets for siRNAs, ASOs, CRISPR-Cas13 gRNAs and RT-PCR primers.