Open access • Posted Content • DOI:10.1101/2020.11.23.393488

LinearTurboFold: Linear-Time RNA Structural Alignment and Conserved Structure Prediction with Applications to Coronaviruses — Source link

Sizhen Li, He Zhang, He Zhang, Liang Zhang ...+7 more authors

Institutions: Oregon State University, Baidu, University of Rochester Medical Center

Published on: 24 Nov 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Structural alignment and Multiple sequence alignment

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

Sizhen Li[a], He Zhang[b,a], Liang Zhang[b,a], Kaibo Liu[b], Boxiang Liu[b], David H. Mathews[d,*], and Liang Huang[a,b,*]

[a]School of Electrical Engineering & Computer Science, Oregon State University, Corvallis, OR; [b]Baidu Research, Sunnyvale, CA; [d]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

Ribonucleic acid (RNA) plays 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 programming 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.*

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.

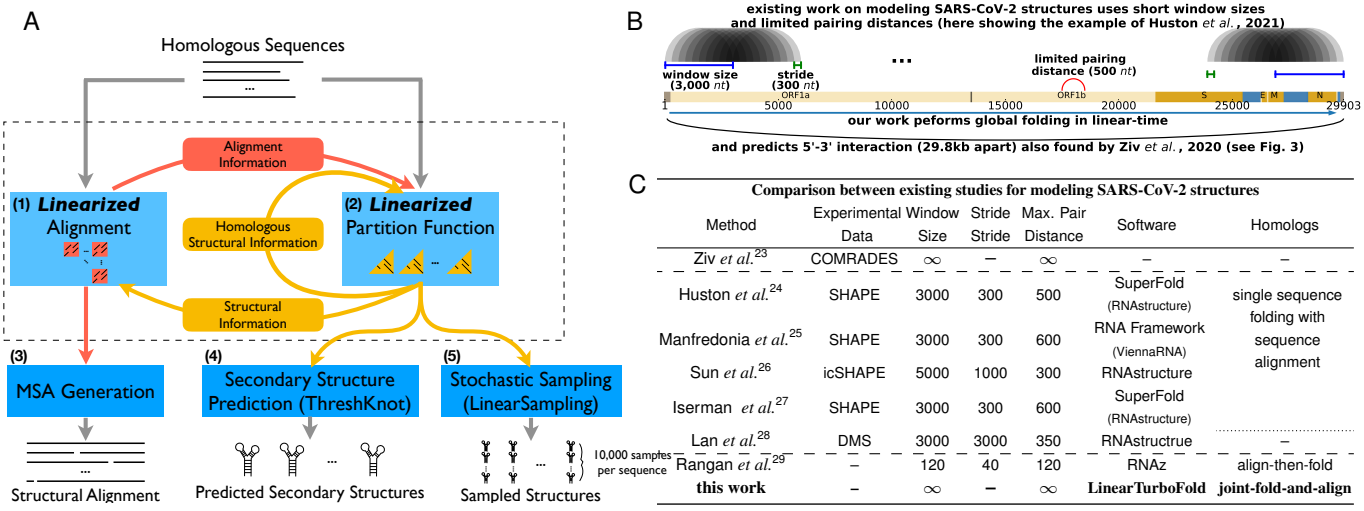
However, the cubic runtime and quadratic memory usage of TurboFold II prevent it from scaling to longer sequences such as full-length SARS-CoV-2 genomes, which contain ∼30,000 nucleotides; in fact, no joint-align-and-fold methods can scale to these genomes, which are the longest among RNA viruses. As a (not very principled) workaround, most existing efforts for modeling SARS-CoV-2 structures[29,24,25,27,28,26] resort to local folding methods[30,31] 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.



Method	Experimental Data	Window Size	Stride Stride	Max. Pair Distance	Software	Homologs
Ziv *et al.*[23]	COMRADES	∞	–	∞	–	–
Huston *et al.*[24]	SHAPE	3000	300	500	SuperFold (RNAstructure)	single sequence folding with sequence alignment
Manfredonia *et al.*[25]	SHAPE	3000	300	600	RNA Framework (ViennaRNA)	
Sun *et al.*[26]	icSHAPE	5000	1000	300	RNAstructure	
Iserman *et al.*[27]	SHAPE	3000	300	600	SuperFold (RNAstructure)	
Lan *et al.*[28]	DMS	3000	3000	350	RNAstructrue	–
Rangan *et al.*[29]	–	120	40	120	RNAz	align-then-fold
this work	–	∞	–	∞	**LinearTurboFold**	**joint-fold-and-align**

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-incidence 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 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** and Fig. 3).

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[32] and structures modeled with the aid of experimental data[24,25,27] for several well-studied regions. Thanks to global rather than local folding, LinearTurboFold discovers a long-range interaction involving 5' and 3' UTRs (∼29,800 *nt* apart), which is consistent with recent purely experimental work,[28] 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[33] and antisense oligonucleotides (ASOs).[26] 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,[34] small interfering RNA (siRNA),[35] CRISPR-Cas13 guide RNA (gRNA)[36] and reverse transcription polymerase chain reaction (RT-PCR) primers.[37] LinearTurboFold is a general technique that can also be applied to other RNA viruses (e.g., influenza, Ebola, HIV, Zika, etc.) and 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[38] (Fig. 1A module **2**), which approximates the RNA partition function[39] 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[40] used by LinearPartition. Finally, LinearTurboFold assembles the secondary structure from the final base pairing probabilities using an accurate and linear-time method named ThreshKnot[41] (module **4**). LinearTurboFold also integrates a linear-time stochastic sampling algorithm named LinearSampling[42] (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, for example, siRNA sequence design.[35] Therefore, the overall end-to-end runtime of LinearTurboFold scales linearly with sequence length (see **Methods §1–7** 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 genomes and SARS-CoV genomes, and the calculation for each family uses five homologous sequences (see **Methods §8** for more details). Fig. 2A compares the running times of

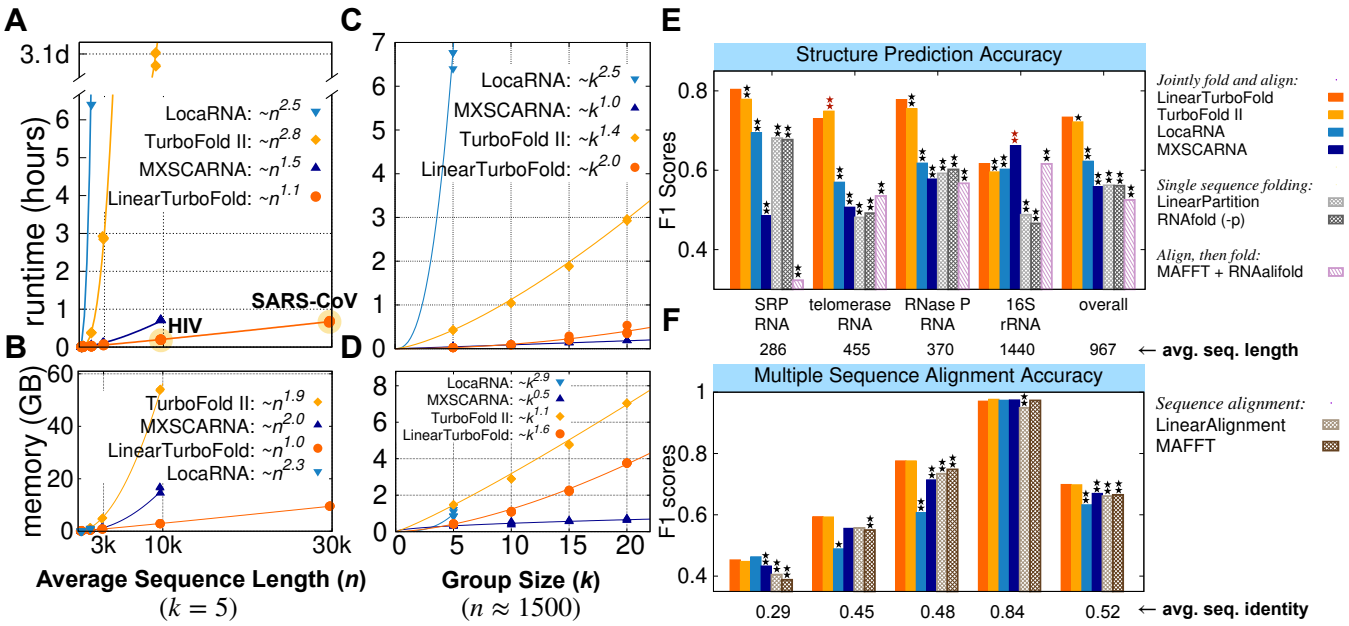


Fig. 2. End-to-end Scalability and Accuracy Comparisons. **A–B**: End-to-end runtime and memory usage comparisons between benchmarks and LinearTurboFold against the sequence length. **C–D**: End-to-end runtime and memory usage comparisons against the group size. LinearTurboFold is the first joint-fold-and-align algorithm to scale to full-length coronavirus genomes (∼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 ($\star$) on the top of the corresponding bars if $p < 0.05$ or two stars ($^{\star}_{\star}$) 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).

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 algorithms, 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 sets 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, was linearized in LinearTurboFold by LinearPartition (Fig. S10C).

We next compare the accuracies of predicted secondary structures and MSAs between LinearTurboFold and several benchmark methods (see **Methods §9**). Besides Sankoff-style LocARNA and MXSCARNA, we also consider three types of negative controls: (a) single sequence folding (partition function-based): Vienna RNAfold[31] (-p mode) and LinearPartition; (b) sequence-only alignment: MAFFT[21] and LinearAlignment (a standalone version of the alignment method developed for this work, but without structural information in LinearTurboFold); and (c) an align-then-fold method that predicts consensus structures from MSAs (Fig. S6): 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

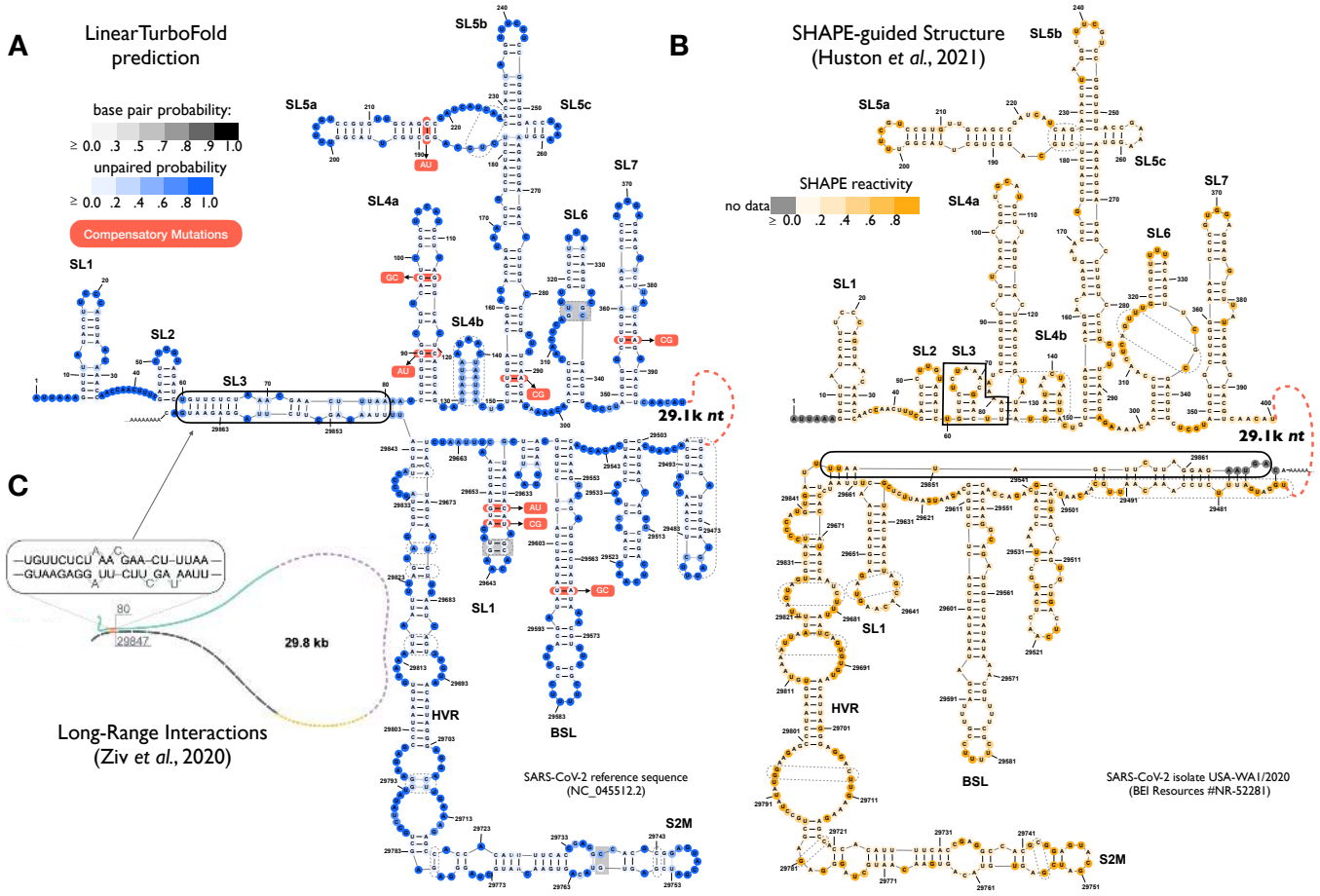


Fig. 3. Secondary structures predictions of SARS-CoV-2 extended 5' and 3' UTRs. **A**: LinearTurboFold prediction.The nucleotides and base pairs are colored by unpaired probabilities and base-pairing probabilities, respectively. The compensatory mutations extracted by LinearTurboFold are annotated with 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.*[24] (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[23] (**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 (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 model and Ziv *et al.*'s experimental result.

Initiative on Sharing Avian Influenza Data (GISAID)[43] up to December 2020 (**Methods §11**). We further extended the group by adding 9 SARS-related homologous genomes (5 human SARS-CoV-1 and 4 bat coronaviruses).[44] 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.[26,25] 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.*[23] 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[45,46,47] 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.[24,25,27,28] 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

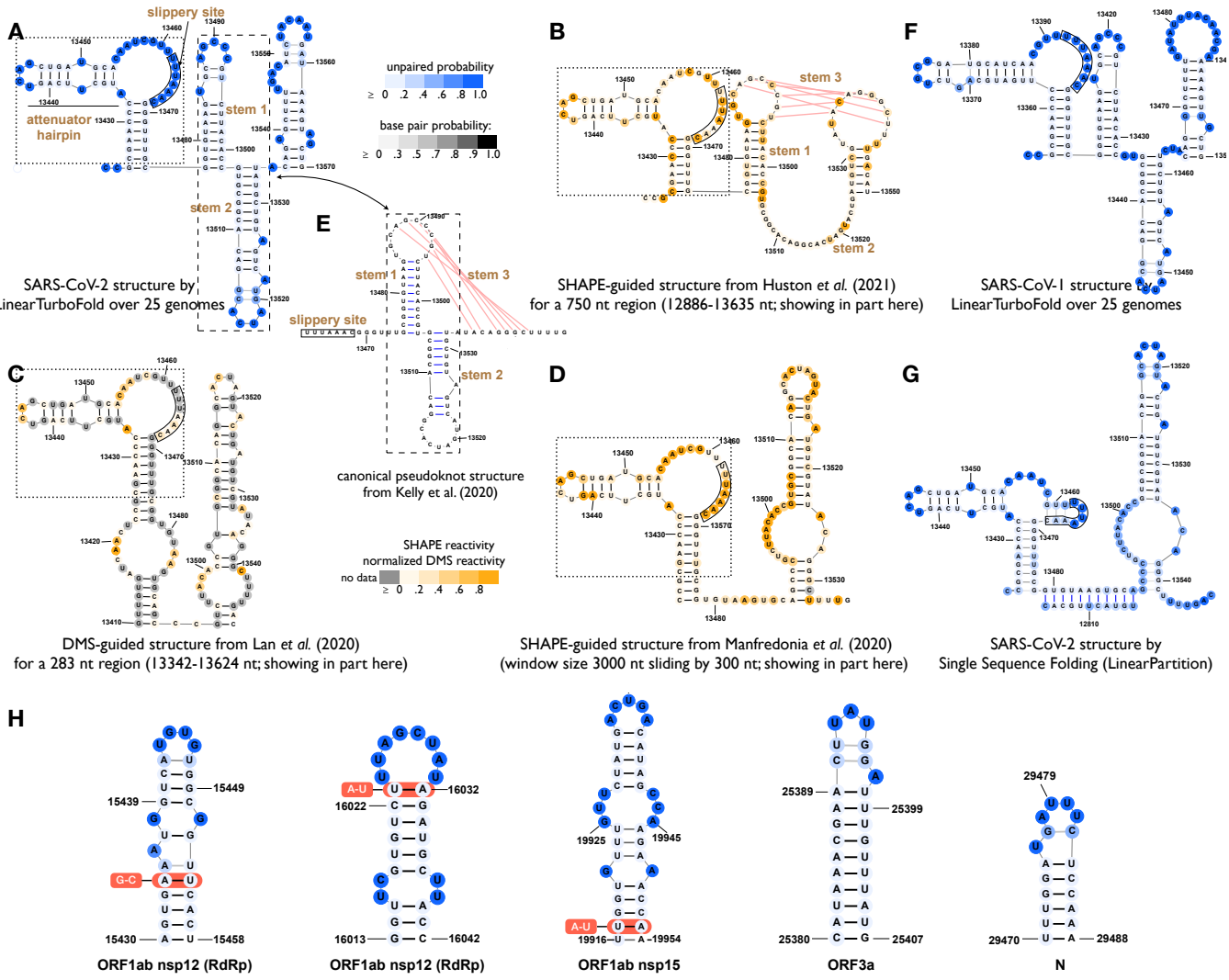
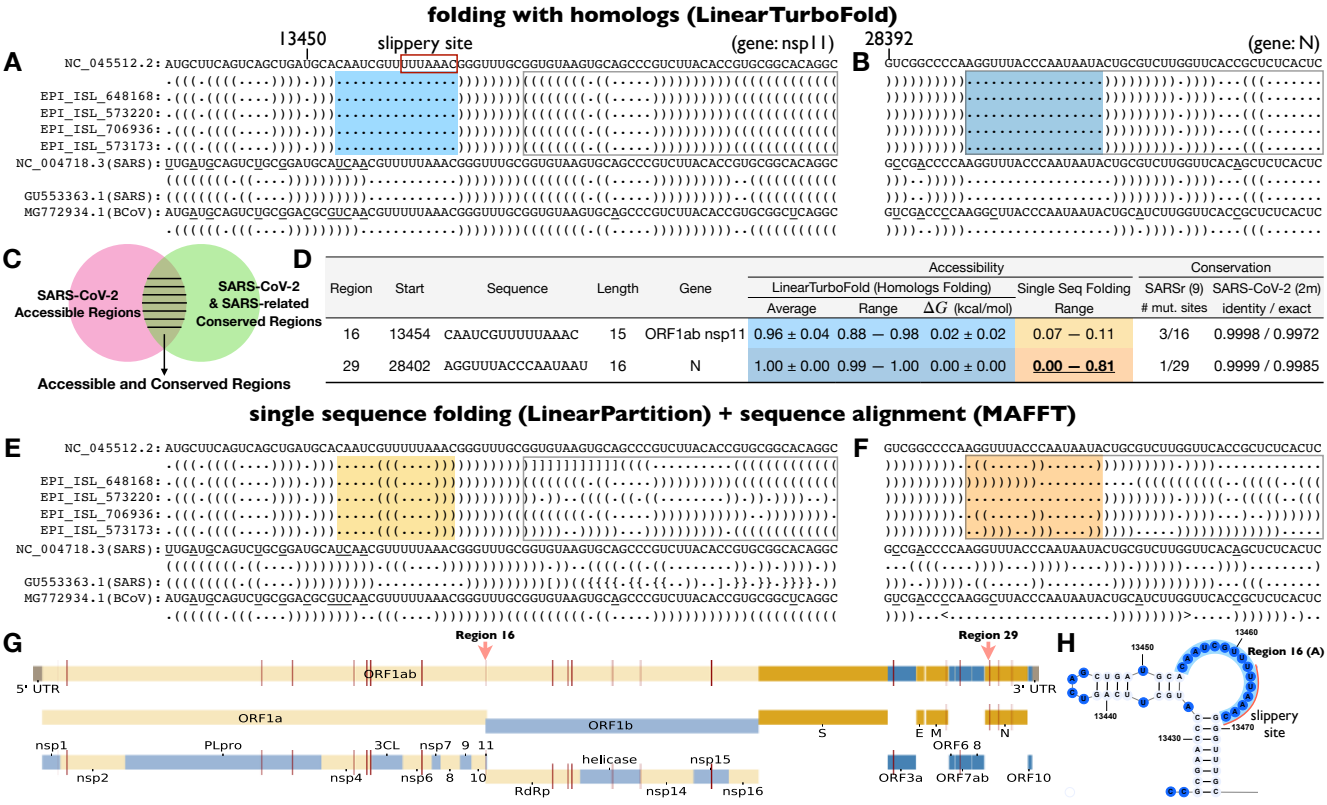


Fig. 4. A–D: Secondary structure predictions of SARS-CoV-2 extended frameshifting stimulation element (FSE) region (13425–13545 *nt*). **A**: LinearTurboFold prediction. **B–D**: Experimentally-guided predictions from the literature,[24,28,25] which are sensitive to the context and region boundaries 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.[32] 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.[29,24]

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. S11B).

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[24,25,28] (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.[33] 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[32] (Fig. 4E). Our prediction is closer to the canonical structure compared to the experimentally-guided models[24,25,28] (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



Region	Start	Sequence	Length	Gene	Accessibility: LinearTurboFold (Homologs Folding) Average	Range	ΔG (kcal/mol)	Accessibility: Single Seq Folding Range	Conservation: SARSr (9) # mut. sites	Conservation: SARS-CoV-2 (2m) identity / exact
16	13454	CAAUCGUUUUUAAAC	15	ORF1ab nsp11	0.96 ± 0.04	0.88 — 0.98	0.02 ± 0.02	0.07 — 0.11	3/16	0.9998 / 0.9972
29	28402	AGGUUUACCCAAUAAU	16	N	1.00 ± 0.00	0.99 — 1.00	0.00 ± 0.00	**0.00 — 0.81**	1/29	0.9999 / 0.9985

Fig. 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**: Two examples out of 33 accessible and conserved regions found by LinearTurboFold. Region 16 and Region 29 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**). The conservation of these regions on 9 SARS-related genomes is the number of mutated sites. The conservation on the ∼2M SARS-CoV-2 dataset is shown in both average sequence identity with the reference sequence and the percentage of exact matches, respectively. **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 33 regions (red bars) across the whole genome (see Tab. S6 for more details). All the accessible and conserved regions are potential targets for siRNAs, ASOs, CRISPR-Cas13 gRNAs and RT-PCR primers.

for the FSE region were estimated for specific local regions. As a result, the models are sensitive to the context and region boundaries[28,24,48] (see Fig. S12D–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.[32] Finally, as negative controls, both the single sequence folding algorithm (LinearPartition in Fig. 4G) and the align-then-fold method (RNAalifold in Fig. S12G) 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).

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[29,24] (Fig. 4H and Tab. S4). These novel structures are potential targets for small-molecule drugs[33] and antisense oligonucleotides.[26,49] 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. We also provide the whole multiple sequence alignment and predicted structures for 25 genomes from LinearTurboFold (see Fig. S13 for the format and link).

Highly Accessible and Conserved Regions in SARS-CoV-2 and SARS-related Betacoronaviruses. Studies show that the siRNA silencing efficiency, ASO inhibitory efficacy, CRISPR-Cas13 knockdown efficiency, and RT-PCR primer binding efficiency, all correlate with the target region's *accessibility*,[37,35,36,50] which is the probability of a target site being fully unpaired. However, most existing work for designing siRNAs, ASOs, CRISPR-Cas13 gRNAs, and RT-PCR primers does not take this feature into consideration[51,52] (Tab. S5). Here LinearTurboFold is able to provide more principled design candidates by identifying accessible regions of the target genome. In addition to accessibility, the emerging variants around the world reduce effectiveness of existing vaccines and test kits (Tab. S5), which indicates sequence conservation is another critical aspect for therapeutic and diagnostic design. LinearTurboFold, being a tool for both structural alignment and homologous folding, can identify regions that are both (sequence-wise) conserved and (structurally)

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

To get unstructured regions, Rangan *et al.*[29] imposed a threshold on unpaired probability of each position, which is a crude approximation because the probabilities are not independent of each other. By contrast, the widely-used stochastic sampling algorithm[53,42] 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[42] 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 define *accessible regions* that are with at least 0.5 accessibility among all 16 SARS-CoV-2 genomes (Fig. 5A–B). We also measure the free energy to open a target region $[i, j]$,[54] notated: $\Delta G_{\mathrm{u}}[i, j] = -RT(\log Z_{\mathrm{u}}[i, j] - \log Z) = -RT \log P_{\mathrm{u}}[i, j]$ where Z is the partition function which sums up the equilibrium constants of all possible secondary structures, $Z_{\mathrm{u}}[i, j]$ is the partition function over all structures in which the region $[i, j]$ is fully unpaired, R is the universal gas constant and T is the thermodynamic temperature. Therefore $P_{\mathrm{u}}[i, j]$ is the unpaired probability of the target region and can be approximated via sampling by s_0/s, where s is the sample size and s_0 is the number of samples in which the target region is single-stranded. The regions whose free energy changes are close to zero need less free energy to open, thus more accessible to bind with siRNAs, ASOs, CRISPR-Cas13 gRNAs and RT-PCR primers.

Next, to identify *conserved regions* that are highly conserved among both SARS-CoV-2 and SARS-related genomes, we require that these regions contain at most three mutated sites on the 9 SARS-related genomes compared to the SARS-CoV-2 reference sequence because historically conserved sites are also unlikely to change in the future,[55] and the average sequence identity with reference sequence over a large SARS-CoV-2 dataset is at least 0.999 (here we use a dataset of ∼2M SARS-CoV-2 genomes submitted to GISAID up to June 30, 2021[†]; see **Methods §11**). Finally, we identified 33 *accessible and conserved regions* (Fig. 5G and Tab. S6), which are not only structurally accessible among SARS-CoV-2 genomes but also highly conserved among SARS-CoV-2 and SARS-related genomes (Fig. 5C). Because the specificity is also a key factor influencing siRNA efficiency,[56] we used BLAST against the human transcript dataset for these regions (Tab. S6). Finally, we also listed the GC content of each region. 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). 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). By removing the constraint of conservation on SARS-related genomes, we identified 38 additional candidate regions (Tab. S7) that are accessible but only highly conserved on SARS-CoV-2 variants.

We also designed a negative control by analyzing the SARS-CoV-2 reference sequence alone using LinearSampling, which can also predict 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. S8). The reason for this difference 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).

Discussion

The constant emergence of new SARS-CoV-2 variants is reducing the effectiveness of exiting vaccines and test kits. To cope with this issue, there is an urgent need to identify conserved structures as promising targets for therapeutics and diagnostics that would work in spite of current and future mutations. Here we 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 on SARS-CoV-2 using local folding and single-sequence folding workarounds, LinearTurboFold enables unprecedented global structural analysis on 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 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.

Methods

Detailed description of our algorithms, datasets, and evaluation metrics are available in the online version of the paper.

[†]The average sequence identity is 0.9987 on that ∼2M dataset (downloaded on July 25, 2021).

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] Robert W Holley, Jean Apgar, 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.

[8] 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.

[9] 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.

[10] KP Williams and DP Bartel. Phylogenetic analysis of tmRNA secondary structure. *RNA*, 2(12):1306–1310, 1996.

[11] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810—-825, 1985.

[12] 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.

[13] Jakob H Havgaard, Elfar Torarinsson, and Jan Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Computational Biology*, 3(10):1896–1908, 2007.

[14] 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.

[15] Zhenjiang Xu and David H Mathews. Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics*, 27(5):626–632, 2011.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Ilaria Manfredonia, Chandran Nithin, Almudena Ponce-Salvatierra, 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.

[26] Lei Sun, Pan Li, Xiaohui Ju, Jian Rao, Wenze 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.

[27] 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.

[28] Tammy CT Lan, Matthew F Allan, Lauren Malsick, Stuti Khandwala, Sherry SY Nyeo, Mark Bathe, Anthony Griffiths, and Silvi Rouskin. Structure of the full SARS-CoV-2 RNA genome in infected cells. *BioRxiv*, 2020.

[29] Ramya Rangan, Ivan N Zheludev, Rachel J Hagey, Edward A Pham, Hannah K Wayment-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.

[30] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):1–9, 2010.

[31] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):1, 2011.

[32] 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.

[33] Hafeez S Haniff, 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] John S. McCaskill. The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, 29:11105–1119, 1990.

[40] Liang Huang and Kenji Sagae. Dynamic programming for linear-time incremental parsing. In *Proceedings of ACL 2010*, page 1077–1086, Uppsala, Sweden, 2010. ACL.

[41] Liang Zhang, He Zhang, David H. Mathews, and Liang Huang. ThreshKnot: Thresholded probknot for improved RNA secondary structure prediction. *BioRxiv*, 2019.

[42] 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.

[43] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017.

[44] Carmine Ceraolo and Federico M Giorgi. Genomic variance of the 2019-nCoV coronavirus. *Journal of Medical Virology*, 92(5):522–528, 2020.

[45] Matthew G Seetin and David H Mathews. RNA structure prediction: an overview of methods. In *Bacterial Regulatory RNA*, pages 99–122. Springer, 2012.

[46] Thomas JX Li and Christian M Reidys. The rainbow spectrum of RNA secondary structures. *Bulletin of Mathematical Biology*, 80(6):1514–1538, 2018.

[47] 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.

[48] 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.

[49] Valeria Lulla, Michal P Wandel, Katarzyna J Bandyra, Rachel Ulferts, 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.

[50] Zhi J. Lu and David H. Mathews. Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Research*, 36:640–647, 2008.

[51] 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.

[52] 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.

[53] Ye Ding and Charles E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.

[54] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.

[55] Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.

[56] Elham Fakhr, 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.

[57] 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.

[58] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[59] Ivo L Hofacker, Stephan HF Bernhart, and Peter F Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227, 2004.

[60] Stanislav Bellaousov and David H Mathews. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, 16(10):1870–1880, 2010.

[61] 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.

[62] 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.

[63] Nima Aghaeepour and Holger H Hoos. Ensemble-based prediction of RNA secondary structures. *BMC Bioinformatics*, 14(1):139, 2013.

[64] 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.

[65] Paul P Gardner and Robert Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5(1):1–18, 2004.

[66] 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.

[67] 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.

[68] Adeel Afzal. Molecular diagnostic technologies for COVID-19: Limitations and challenges. *Journal of Advanced Research*, 2020.

Methods

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

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

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

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

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

where x_i and y_i are the ith and jth nucleotides of sequences $\mathbf{x}$ and $\mathbf{y}$, respectively. Note that the first step $a_0 = (\text{ALN}, 0, 0)$ and the last $a_{m+1} = (\text{ALN}, |\mathbf{x}|+1, |\mathbf{y}|+1)$ do not have emissions.

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

$$\begin{aligned} \alpha^h_{i,j} &= \sum_{\substack{a \in \mathcal{A}(\mathbf{x},\mathbf{y}) \\ \exists k, a_k = (h,i,j)}} p(\mathbf{x},\mathbf{y},a[:k]) \\ &= p_\pi(h_0) \cdot \prod_{l=1}^{k} p_\text{t}(h_l \mid h_{l-1}) p_\text{e}(e(\mathbf{x},\mathbf{y},a_l) \mid h_l) \end{aligned}$$

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

$$\begin{aligned} \beta^h_{i,j} &= \sum_{\substack{a \in \mathcal{A}(\mathbf{x},\mathbf{y}) \\ \exists k, a_k = (h,i,j)}} p(\mathbf{x},\mathbf{y},a[k+1:]) \\ &= \left\{ \prod_{l=k+1}^{m} p_\text{t}(h_l \mid h_{l-1}) p_\text{e}(e(\mathbf{x},\mathbf{y},a_l) \mid h_l) \right\} \cdot p_\text{t}(h_{m+1} \mid h_m) \end{aligned}$$

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

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

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

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:

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

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

$$p(i \sim j \mid \mathbf{x},\mathbf{y}) = \frac{\sum_h \alpha^h_{i,j} \cdot \beta^h_{i,j}}{\alpha^{\text{ALN}}_{|\mathbf{x}|+1,|\mathbf{y}|+1}}$$

§3 LinearAlignment. Unlike a previous method[57] 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 $\alpha^h_{i,j}$ 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 there states, and updates the corresponding forward probabilities $\alpha^{\text{INS1}}_{i+1,j}$, $\alpha^{\text{INS2}}_{i,j+1}$ and $\alpha^{\text{ALN}}_{i+1,j+1}$, 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[40] 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 $\alpha^h_{i,j}$ (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 encourage the pairwise alignment conforming with estimated secondary structures, LinearTurboFold predicts structural alignments by incorporating the secondary structural conformation. PMcomp[59] 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 corresponding 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 $(t-1)$:

$$\begin{aligned} m^{(t)}_{\mathbf{x},\mathbf{y}}(i,j) = \alpha_1 &\left[\sqrt{P^{(t-1)}_{\mathbf{x},>}(i) \cdot P^{(t-1)}_{\mathbf{y},>}(j)} \sqrt{P^{(t-1)}_{\mathbf{x},<}(i) \cdot P^{(t-1)}_{\mathbf{y},<}(j)} \right] \\ &+ \alpha_2 \sqrt{P^{(t-1)}_{\mathbf{x},o}(i) \cdot P^{(t-1)}_{\mathbf{y},o}(j)} + \alpha_3 \end{aligned}$$

where α_1, α_2 and α_3 are weight parameters trained in TurboFold II. The forward-backward phrases integrate the match score as a prior when aligning two nucleotides (Fig. S9 line 10, and Fig. S9 line 12).

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

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

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

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

§6 LinearPartition for Base Pairing Probabilities Estimation with Extrinsic Information. The classical partition function algorithm scales cubically with sequence length. The slowness limits its extension to longer sequences. To address this bottleneck, our recent LinearPartition[38] algorithm approximates the partition function and base paring probability matrix computation in linear time. LinearPartition is significantly faster, and correlates better with the ground truth structures than the traditional cubic partition function calculation. Thus LinearTurboFold uses LinearPartition to predict base pair probabilities instead of the traditional $O(n^3)$-time partition function.

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

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

§7 MSA Generation and Secondary Structure Prediction. After several iterations, TurboFold II builds the multiple sequence alignment using a probabilistic consistency transformation, generating a guide tree and performing progressive alignment over the pairwise posterior co-incidence probabilities.[22] The whole procedure is accelerated in virtue of the sparse matrix by discarding alignment pairs of probability smaller than a threshold (0.01 by default). Since LinearAlignment uses the beam search method and only saves a linear number of co-incident pairs, the MSA generation in LinearTurboFold costs linear runtime against the sequence length straightforwardly.

Estimated base pair probabilities are fed into downstream methods to predict secondary structures. To maintain the end-to-end linear-time property, LinearTurboFold uses ThreshKnot,[41] which is a thresholded version of ProbKnot[60] and only considers base pairs of probability exceeding a threshold θ ($\theta = 0.3$ by default). We evaluate the performance of ThreshKnot and MEA with different hyperparameters (θ and γ). On a sampled RNAStrAlign training set, ThreshKnot is closer to the upper right-hand than MEA, which indicates that ThreshKnot always has a higher Sensitivity than MEA at a given PPV (Fig. S10B).

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

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

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

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

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

LocARNA (local alignment of RNA) costs $O(n^2(n^2 + k^2))$ time and $O(n^2 + k^2)$ space by restricting the alignable regions. MXSCARNA progressively aligns multiple sequences as an extension of the pairwise alignment algorithm SCARNA[62] with improved score functions. SCARNA first aligns stem fragment candidates, then removes the inconsistent matching in the post-processing to generate the sequence alignment. MXSCARNA reduces runtime to $O(k^3 n^2)$ and space to $O(k^2 n^2)$ with a limited searching space of folding

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

§10 Significance Test. We use a paired, two-tailed permutation test[63] to measure the significant difference. Following the common practice, the repetition number is 10,000, and the significance threshold α is 0.05.

§11 SARS-CoV-2 Datasets. We used two large SARS-CoV-2 datasets. The first dataset is used to draw a representative sample of most diverse SARS-CoV-2 genomes. We downloaded all the genomes submitted to GISAID[43] by December 29, 2020 (downloaded on December 29, 2020), and filtered out low-quality genomes (with more than 5% unknown characters and degenerate bases, shorter than 29,500 *nt*, or with framing error in the coding region), and we also discard genomes with more than 600 mutations compared with the SARS-CoV-2 reference sequence (NC_0405512.2).[64] After preprocessing, this dataset includes about 258,000 genomes. To identify a representative group of samples with more variable mutations, we designed a greedy algorithm to select 16 most diverse genomes genomes found at least twice in the 258,000 genomes. The general idea of the greedy algorithm is to choose genomes one by one with the most new mutations compared with the selected samples, which consists of only the reference sequence at the beginning.

The second, larger, dataset is to evaluate the conservation of regions with respect to more up-to-date variants. We downloaded all the genomes submitted to GISAID by June 30, 2021 (downloaded on July 25, 2021), and did the same preprocessing as the first dataset. This resulted in a dataset of $\sim$2M genomes, 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

-

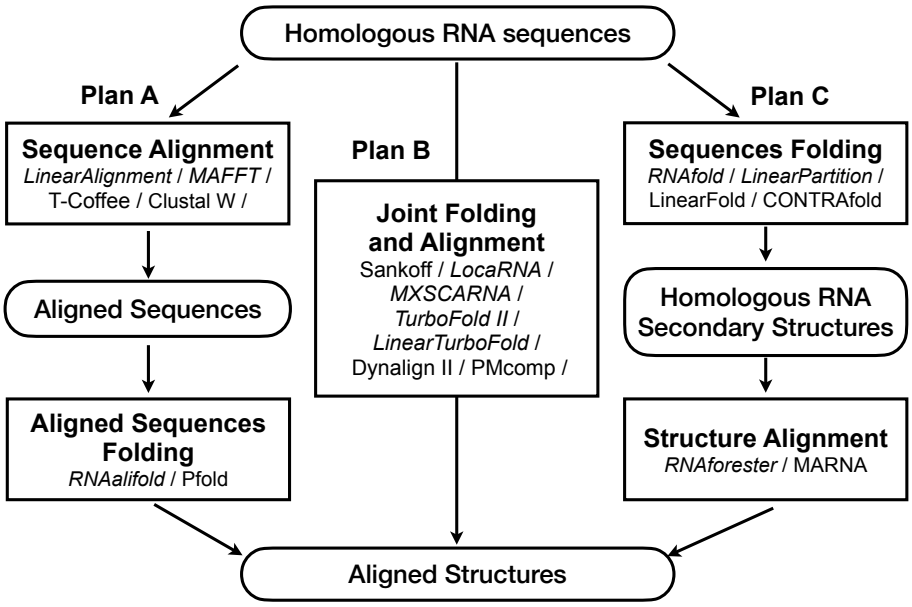


Fig. S6. Approaches for analyzing homologous sequence can be categorized into three plans[65] (related to Fig. 1). Plan A involves two steps: first aligning sequences and then folding aligned multiple sequences. This line works well for homologs with a high sequence identity. Plan B employs joint folding 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).

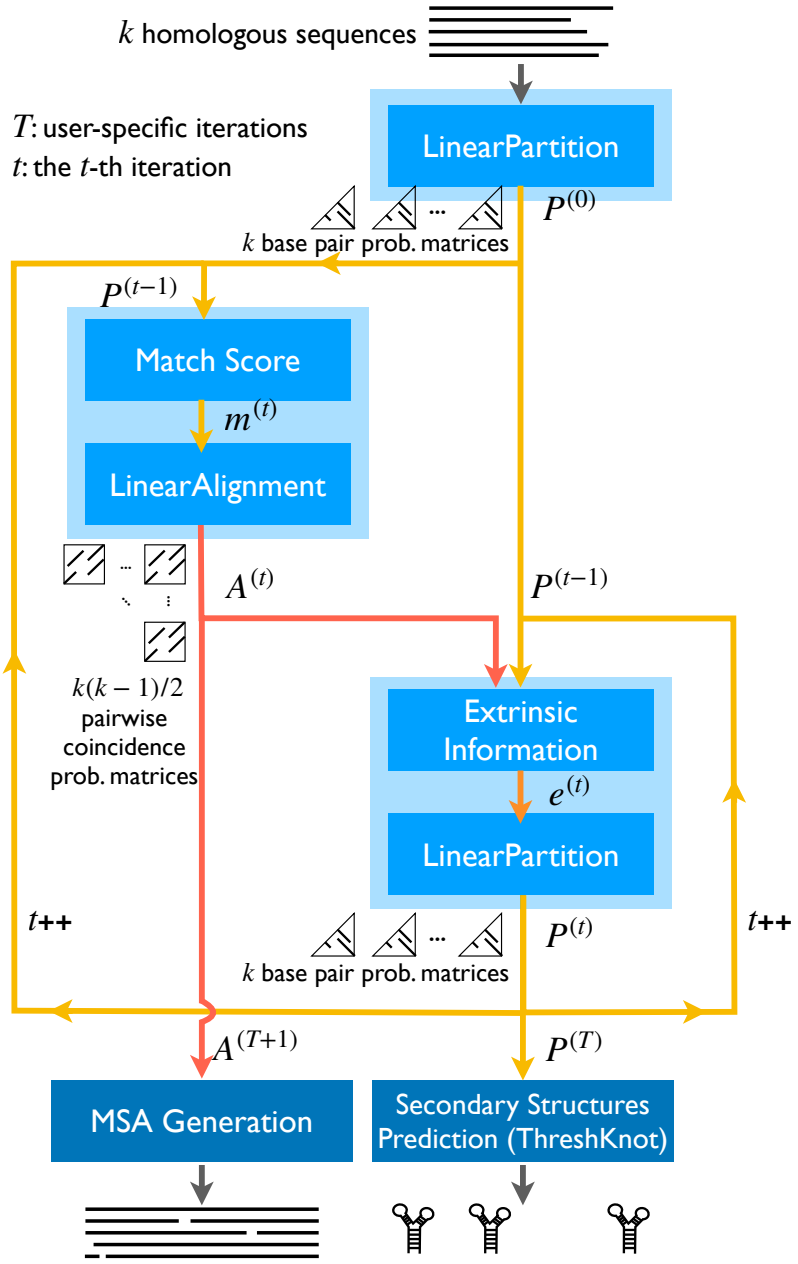


Fig. S7. The flowchart of LinearTurboFold with more detailed information (related to Fig. 1). At iteration 0, LinearPartition calculates the partition function and estimates the base pair probabilities for each sequence. From iteration 1 to T, the two major modules LinearAlignment and LinearPartition 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.

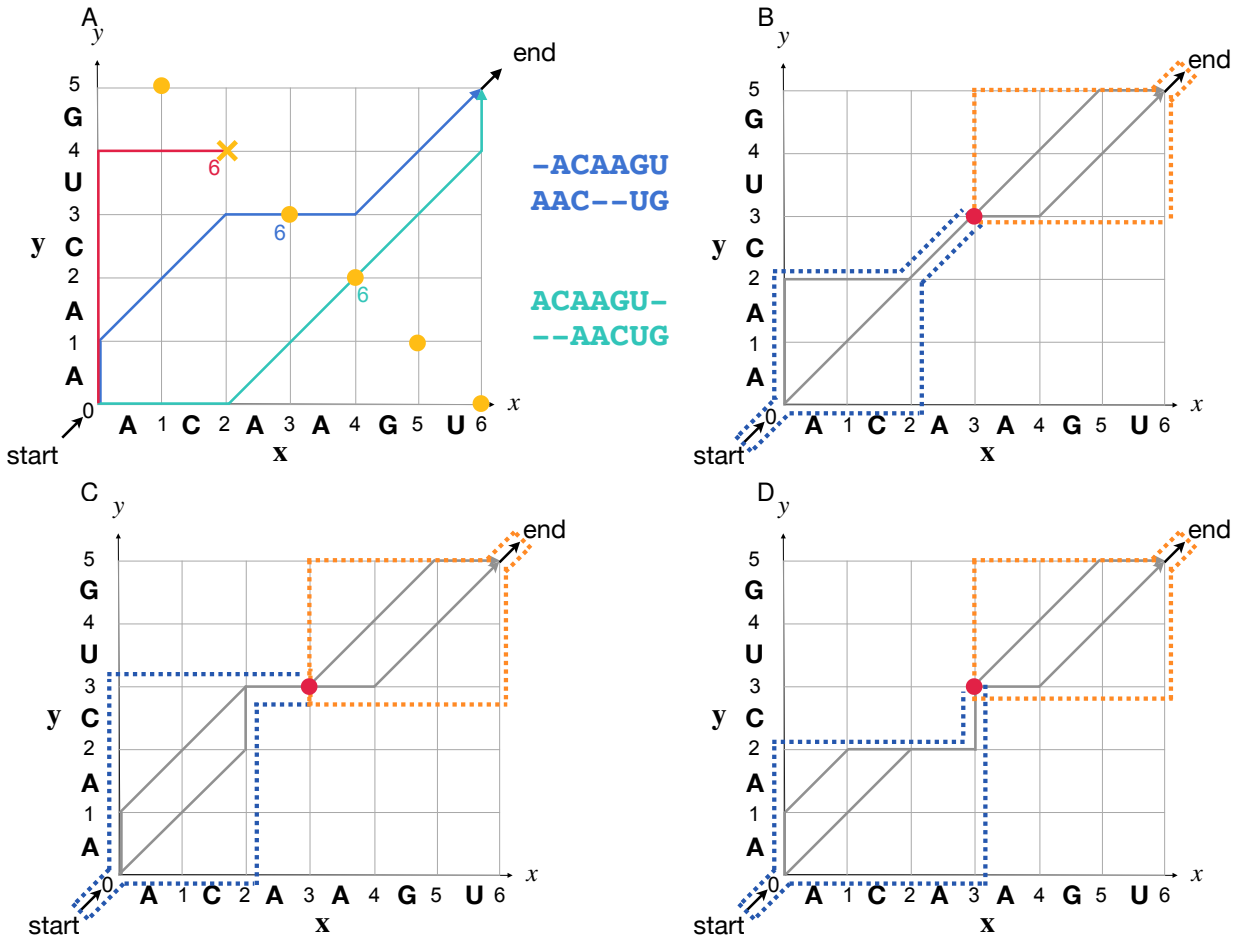


Fig. S8. Illustrations of LinearAlignment. **A**: An example of aligning two sequences and the beam search method based on the step count. The x-axis and y-axis of the matrix represent two sequences $\mathbf{x}$ and $\mathbf{y}$. Yellow notes have the same step count 6. At step count 6, the red path 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 $\alpha^{\text{ALN}}_{3,3}$, 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 $(\text{ALN}, 3, 3)$ ($\beta^{\text{ALN}}_{3,3}$). **C** and **D**: The regions circled by the blue dashed lines are $\alpha^{\text{INS1}}_{3,3}$ and $\alpha^{\text{INS2}}_{3,3}$, and regions circled by the orange dashed lines are $\beta^{\text{INS1}}_{3,3}$ and $\beta^{\text{INS2}}_{3,3}$, respectively.

function FORWARD($\mathbf{x}$, $\mathbf{y}$, b_1) ▷ the forward phase
  $\alpha^{\text{ALN}}_{0,0} \leftarrow 1$ ▷ initial probability distribution
  for $s = 0 \ldots |\mathbf{x}| + |\mathbf{y}| - 1$ **do** ▷ topological order
    for each h **in** $\mathcal{H}$ **do**
      $B(s,h) \leftarrow$ all the nodes (i,j) such that $\alpha^h_{i,j}$ exists and $i + j = s$
      BEAMPRUNE($B(s,h), b_1$) ▷ keep top b_1 nodes in $B(s,h)$ by $\alpha^h_{i,j}$
      for each node (i,j) **in** $B(s,h)$ **do** ▷ transitions to next states
        $\alpha^{\text{INS1}}_{i+1,j} \mathrel{+}= \alpha^h_{i,j} \cdot p_{\text{t}}(\text{INS1} \mid h) \cdot p_{\text{e}}((x_{i+1}, -) \mid \text{INS1})$
        $\alpha^{\text{INS2}}_{i,j+1} \mathrel{+}= \alpha^h_{i,j} \cdot p_{\text{t}}(\text{INS2} \mid h) \cdot p_{\text{e}}((-, y_{j+1}) \mid \text{INS2})$
        $\alpha^{\text{ALN}}_{i+1,j+1} \mathrel{+}= \alpha^h_{i,j} \cdot p_{\text{t}}(\text{ALN} \mid h) \cdot p_{\text{e}}((x_{i+1}, y_{j+1}) \mid \text{ALN})$
  return α

1. Pseudocode of the LinearAlignment algorithm forward phase

function BACKWARD($\mathbf{x}$, $\mathbf{y}$, α, B) ▷ the backward phase
  $\beta \leftarrow$ hash()
  $\beta^{\text{ALN}}_{|\mathbf{x}|+1,|\mathbf{y}|+1} \leftarrow 1$ ▷ initialization
  $p_{\mathbf{x},\mathbf{y}} \leftarrow \alpha^{\text{ALN}}_{|\mathbf{x}|+1,|\mathbf{y}|+1}$ ▷ probability of observing two sequences
  $p_{i,j} \leftarrow 0$ ▷ co-incident probability initialization
  for $s = |\mathbf{x}| + |\mathbf{y}| \ldots 0$ **do**
    for each h **in** $\mathcal{H}$ **do**
      for each node (i,j) **in** $B(s,h)$ **do** ▷ $B(s,h)$ saves b_1 entries during the forward phase
        if $i = |\mathbf{x}|$ **and** $j = |\mathbf{y}|$ **then** ▷ boundary conditions
          $\beta^h_{i,j} = p_{\text{t}}(\text{ALN} \mid h) \cdot \beta^{\text{ALN}}_{|\mathbf{x}|+1,|\mathbf{y}|+1}$
        else
          $\beta^h_{i,j} = p_{\text{t}}(\text{ALN} \mid h) \cdot p_{\text{e}}((x_{i+1}, y_{j+1}) \mid \text{ALN}) \cdot \beta^{\text{ALN}}_{i+1,j+1}$
              $+ p_{\text{t}}(\text{INS1} \mid h) \cdot p_{\text{e}}((x_{i+1}, -) \mid \text{INS1}) \cdot \beta^{\text{INS1}}_{i+1,j}$
              $+ p_{\text{t}}(\text{INS2} \mid h) \cdot p_{\text{e}}((-, y_{j+1}) \mid \text{INS2}) \cdot \beta^{\text{INS2}}_{i,j+1}$
        $p_{i,j} \mathrel{+}= \frac{\alpha^h_{i,j} \cdot \beta^h_{i,j}}{p_{\mathbf{x},\mathbf{y}}}$ ▷ update co-incident probabilities

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

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

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

	Structure Prediction Accuracy						
	LinearTurboFold	TurboFold II	LocARNA	MXSCARNA	LinearPartition	Vienna RNAfold	RNAalifold
	PPV						
SRP	0.801	**0.819**	0.698	0.485	0.662	0.673	0.629
telomerase	0.650	**0.685**	0.516	0.465	0.409	0.430	0.602
RNase P RNA	0.734	**0.752**	0.606	0.573	0.543	0.571	0.698
16S rRNA	0.615	0.608	0.586	**0.662**	0.467	0.464	0.628
overall	0.700	**0.716**	0.602	0.546	0.520	0.534	0.639
	Sensitivity						
SRP	**0.806**	0.743	0.693	0.488	0.700	0.682	0.218
telomerase	**0.832**	0.826	0.637	0.558	0.584	0.576	0.482
RNase P RNA	**0.828**	0.758	0.630	0.584	0.650	0.636	0.478
16S rRNA	0.620	0.584	0.622	**0.663**	0.511	0.469	0.605
overall	**0.772**	0.728	0.645	0.573	0.611	0.591	0.446
	F1 scores						
SRP	**0.804**	0.779	0.695	0.486	0.681	0.677	0.323
telomerase	0.730	**0.749**	0.570	0.507	0.481	0.492	0.535
RNase P RNA	**0.778**	0.755	0.618	0.578	0.592	0.602	0.567
16S rRNA	0.617	0.596	0.603	**0.662**	0.488	0.466	0.616
overall	**0.734**	0.722	0.623	0.559	0.562	0.561	0.525

	Multiple Sequence Alignment Accuracy						
	LinearTurboFold	TurboFold II	LocARNA	MXSCARNA	LinearAlignment	MAFFT	RNAforester [66]
	PPV						
SRP	**0.463**	0.458	0.305	0.387	0.414	0.393	0.263
telomerase	**0.617**	0.615	0.311	0.554	0.575	0.572	0.239
RNase P RNA	**0.788**	0.787	0.615	0.692	0.744	0.759	0.258
16S rRNA	0.971	**0.977**	0.647	0.971	0.947	0.974	0.239
overall	**0.710**	0.709	0.470	0.651	0.670	0.675	0.250
	Sensitivity						
SRP	**0.443**	0.438	0.452	0.384	0.396	0.382	0.271
telomerase	**0.573**	0.572	0.470	0.523	0.540	0.529	0.262
RNase P RNA	**0.765**	**0.765**	0.596	0.684	0.724	0.738	0.286
16S rRNA	0.971	**0.977**	0.974	0.971	0.951	0.973	0.298
overall	**0.688**	**0.688**	0.623	0.641	0.653	0.656	0.280
	F1 scores						
SRP	**0.453**	0.448	0.364	0.385	0.405	0.388	0.267
telomerase	**0.594**	0.593	0.375	0.538	0.557	0.550	0.250
RNase P RNA	**0.776**	**0.776**	0.605	0.688	0.734	0.748	0.271
16S rRNA	0.971	**0.977**	0.778	0.971	0.949	0.973	0.265
overall	**0.699**	0.698	0.535	0.646	0.661	0.665	0.264

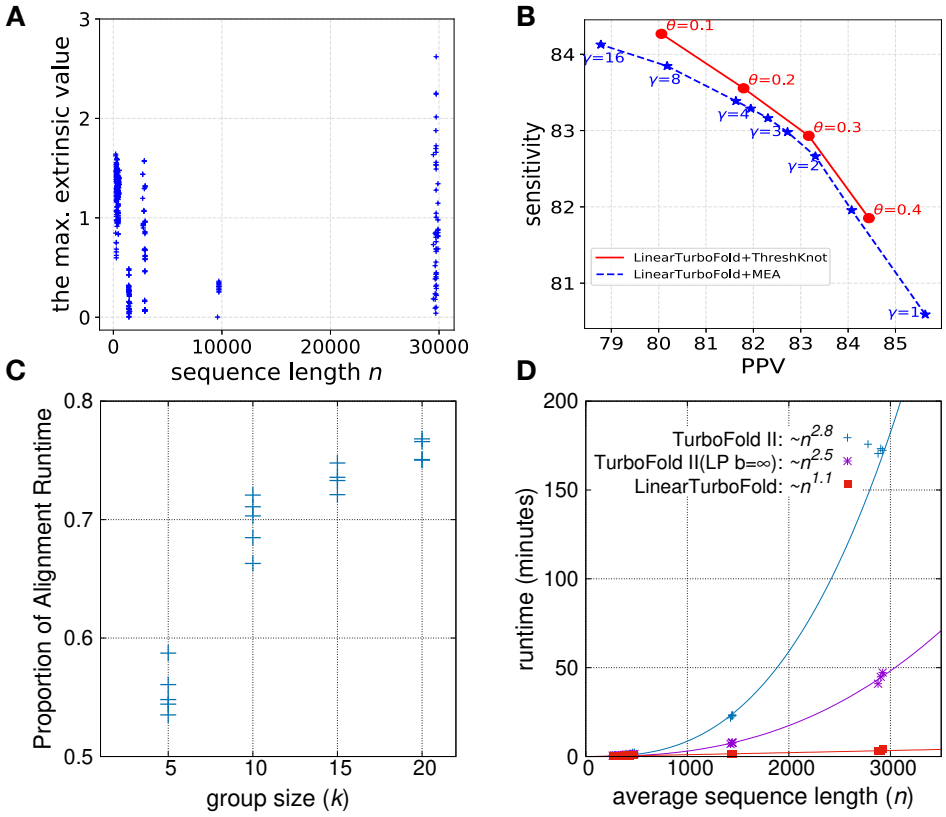


Fig. S10. **A**: The maximum values of the extrinsic information as a function of sequence length. The maximal value for each sequence is recorded when running LinearTurboFold on the collected dataset of sequence length ranging from 200 nt to 30,000 nt. **B**: Accuracy comparison between ThreshKnot and 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,[31] which is a subset of the RNAstructure[67] 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[21] 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	-

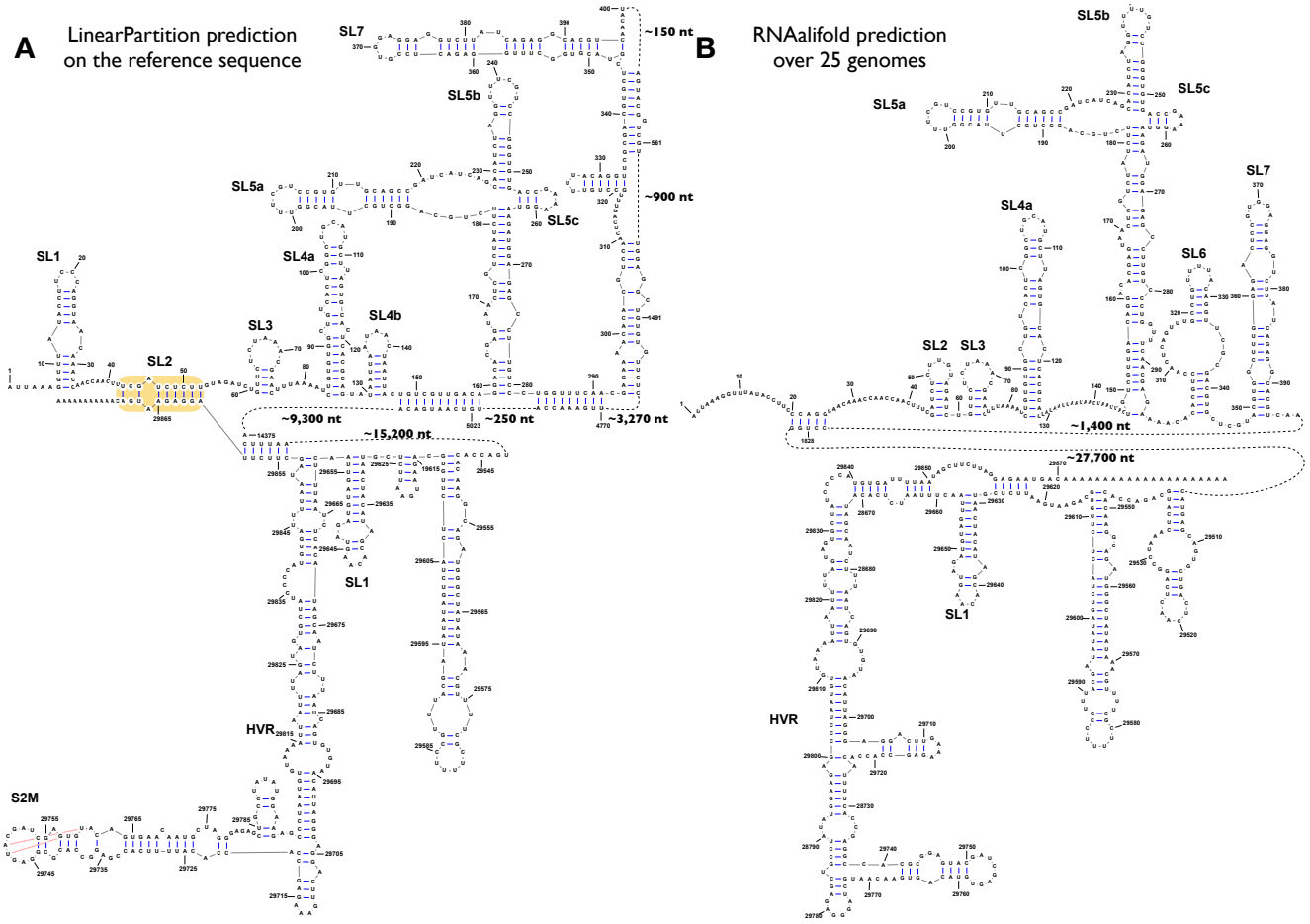


Fig. S11. Secondary structure prediction of SARS-CoV-2 for extended 5' and 3' UTRs (related to Fig. 3). **A**: LinearPartition prediction of the SARS-CoV-2 reference sequence (NC_0405512.2) alone (single sequence folding). LinearPartition also predicts a long-range interaction between 5' and 3' UTRs. 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.

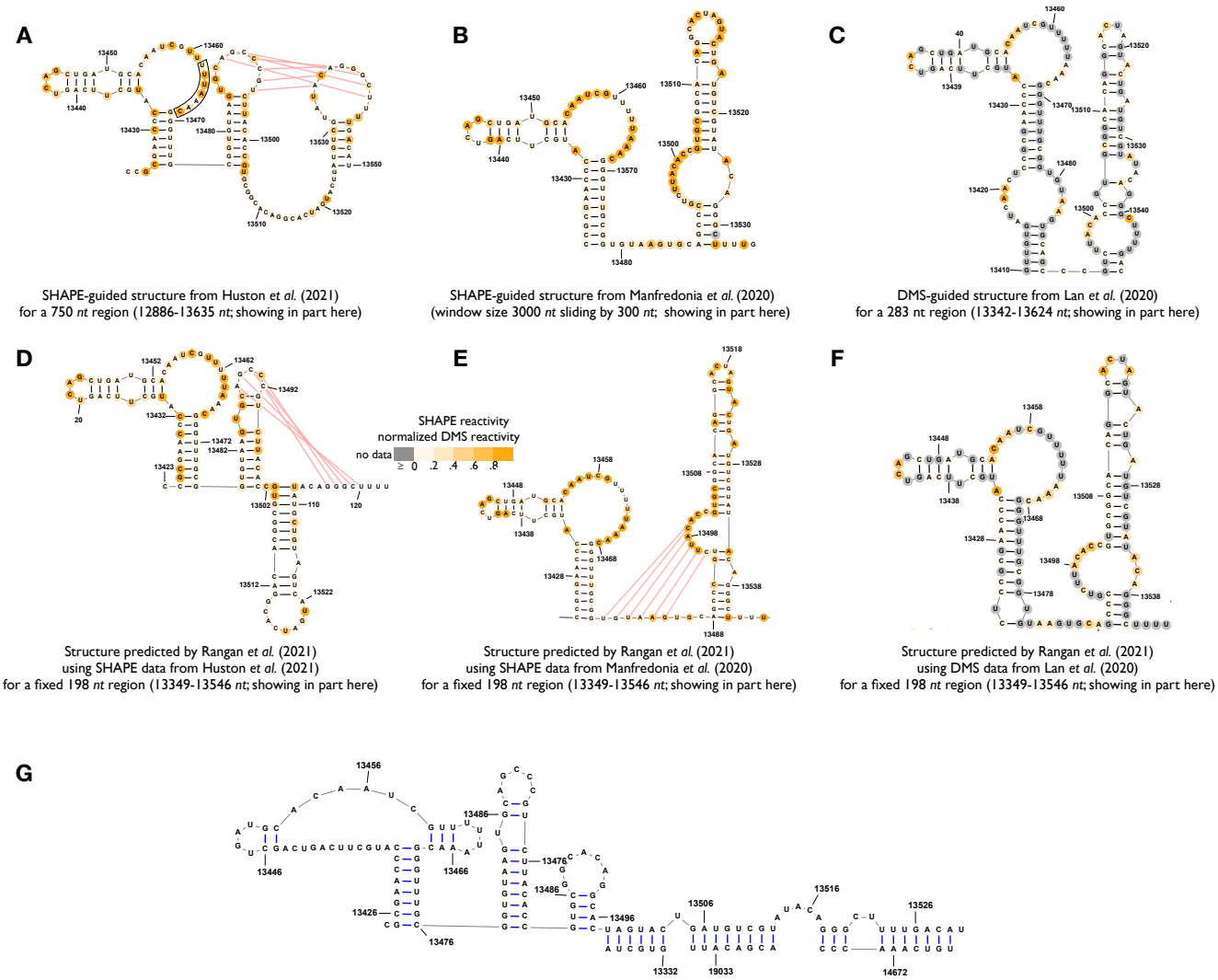


Fig. S12. Secondary structure predictions of SARS-CoV-2 for extended frameshifting stimulation element (13425-13545 *nt*) (related to Fig. 4). **A–F**: Experimentally-guided structures with different probing data for different regions. The structures in each column were estimated with the same experimental data but different regions. The structures in the second row were predicted by Rangan *et al.* for a fixed region of 198 *nt*[48] . 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.*[29,24] 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*	GUAAUUUUAAAGUUAC ((((((....))))))	
2818-2837*	AGUACUUAAUGAGAAGUGCU (((((((.......)))))))	
4973-4993	GUGUUUACAACAGUAGACAAC (((((((....))))))).)	UA ->CG
5021-5033	AUGUCAAUGACAU (((((...)))))	
6450-6498	UUGAGUGUAAUGUGAAAACUACCGAAGUUGUAGGAGACAUUAUACUUAA ((((((((((((.....((((.......))))...))))))))))))	AU ->GC UA ->CG
8078-8084	CCAAUGG ((...))	
10494-10509*	GUGUUGGUUUUAACAU ((((((....))))))	
11131-11140	UGCUUUUGCA (((....)))	
12203-12218*	UUGAAGAAGUCUUUGA ((((((....))))))	
12257-12268	CAACGUAAGUUG ((((....))))	
12386-12412	GAUAAUGAUGCACUCAACAACAUUAUC (((((((.((.......)).)))))))	
12672-12685	CUGUCAAAUUACAG ((((......))))	
12904-12926*	UAGGUUUGUUACAGACACACCUA (((((.((((...)))).)))))	
12970-12988*	CAACCUAAAUAGAGGUAUG (((((......)))).))	
13409-13422*	AGUUGUGAUCAACU (((((....)))))	
14729-14769	AGGAAGGAAGUUCUGUUGAAUUAAAACACUUCUUCUUUGCU ((.(((((((...((((.......))))...))))))).))	
14773-14790	GAUGGUAAUGCUGCUAUC (((((((....)))))))	
14794-14818	GAUUAUGACUACUAUCGUUAUAAUC (((((((((.......)))))))))	
15254-15269*	UUUAUAGUGAUGUAGA ((((((....))))))	
15430-15458*	AGUGAAAUGGUCAUGUGUGGCGGUUCACU ((((((.(.((((....)))).))))))	AU ->GC
15502-15509*	GCUUAUGC ((....))	
15618-15628	ACUUUAUGAGU ((((...))))	
15775-15801*	AUAAAGAACUUUAAGUCAGUUCUUUAU ((((((((((.......))))))))))	
16013-16042*	GGUUCGUGUCUUUAGCUAUAGAUGCUUACC (((..((((((.......))))))..)))	UA ->AU
16180-16194	AGGUAUUGGGAACCU ((((.......))))	
16955-16973*	UAGUGCCACAAGAGCACUA (((((((....).)))))))	

Table S4 continued from previous page

Region	Sequence & Structure	Compensatory Mutations
17236-17262	AUACCUGCACGUGCUCGUGUAGAGUGU	UA ->CG
	((((((((((....)))))))).))))	AU ->GC
17462-17470	GCACAUUGC	
	(((...)))	
18547-18563*	UUAAGUGACACACUUAA	AU ->CG
	(((((((...)))))))	
18660-18674*	UUGUCUAUGUGAUAG	
	(((((.....)))))	
18848-18872*	CUAGUUGUGAUGCAAUCAUGACUAG	
	((((((((((...))))))))))	
19036-19043*	GGUAACCC	
	((....))	
19973-20002*	CACUCACUGUCUUUUUUGAUGGUAGAGUUG	
	((((((((((......)))))).)))).)	
20103-20127*	UCUUAAUGGAGUCACAUUAAUUGGA	
	(((((((.....))))))...)))	
20709-20719*	GCAAAGAAUGC	
	(((.....)))	
23101-23110*	UUCUUUUGAA	
	(((....)))	
23119-23128	UGCACCAGCA	
	(((....)))	
23796-23807	AUUCAACUGAAU	UA ->CG
	((((....))))	
24305-24319	GUUCUCUAUGAGAAC	
	((((((...))))))	
25380-25407*	CAUAAACGAACUUAUGGAUUUGUUUAUG	
	((((((((((...)).))))))))))	
25937-25959	AUGACUACCAGAUUGGUGGUUAU	
	((((((((((...))))))))))	
26209-26224	GUGCCUUUGUAAGCAC	
	((((........))))	
26358-26379	AUUGUGUGCGUACUGCUGCAAU	
	((((((.(((...))))))))))	
26581-26600	GGAACCUAGUAAUAGGUUUC	
	(((((((....)))))))	
26713-26740	GUUUUGUGCUUGCUGCUGUUUACAGAAU	
	(((((((...((....)).)))))))	
27635-27651*	CAGUUUCACCUAAACUG	
	((((((.....))))))	
28755-28766	UUCCUCAAGGAA	
	((((....))))	
29145-29166	UAAUCAGACAAGGAACUGAUUA	
	(((((((........)))))))	
29210-29225	GCGUUCUUCGGAAUGU	
	((((((....))))))	
29240-29253	GAAGUCACACCUUC	
	((((......))))	
29288-29306	UUGGAUGACAAAGAUCCAA	
	((((((.......))))))	
29321-29347	GUCAUUUUGCUGAAUAAGCAUAUUGAC	
	((((...((((.....))))...))))	
29470-29488*	UUUGGAUGAUUUCUCCAAA	
	(((((((.(....)))))))	
29548-29613	CACAAGGCAGAUGGGCUAUAUAAACGUUUUCGCUUUUCCGUUUACGAUAUAUAGUCUACUCUUGUG	UA ->CG
	((((((..((.((((((((((..(((...((......))...))).))))))))))))))))))	

Fig. S13. A glimpse of the whole MSA and aligned predicted structures for 25 genomes from LinearTurboFold, which are available in https://github.com/LinearFold/LinearTurboFold/blob/main/sars-cov-2_results/. Each genome corresponds to three lines: name, aligned sequence and aligned structure, respectively.

Table S5. Accessibility and conservation of target regions for public RT-PCR forward/reverse primers and probes.[68] 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				
		Start	Length	Accessibility	Conservation	
					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 / 0.0160	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	Institut Pasteur (2)	14079 / 14105 / 14166	19 / 19 / 20	0.0000 / 0.0001 / 0.0000	4 / 8 / 9	0.9978 / 0.9332 / 0.9941
(RdRp)	Charite Germany (1)	15430 / 15469 / 15504	22 / 25 / 26	0.0000 / 0.0000 / 0.0000	0 / 6 / 1	0.9167 / 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
N	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 / 0.0026 / 0.0000	3 / 6 / 6	0.9908 / 0.9953 / 0.9927
	CDC (2)	28680 / 28704 / 28731	22 / 24 / 21	0.0000 / 0.0000 / 0.0010	4 / 4 / 2	0.9862 / 0.9796 / 0.9895
	Charite Germany (3)	28705 / 28753 / 28813	19 / 25 / 20	0.0000 / 0.0003 / 0.0000	2 / 0 / 6	0.9914 / 0.9920 / 0.9858
	China CDC (2)	28880 / 28934 / 28957	22 / 20 / 22	0.0710 / 0.0000 / 0.0000	5 / 7 / 4	**0.2734** / 0.9911 / **0.4844**
	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	ΔG(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	AUUAUAUGCCUUAUUU	ORF1ab nsp3 (PLpro)	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	UAUAUGCCUUAUUUC		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	CAGUUUAAAAGACCAAUAAAUC		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	ORF1ab nsp4	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		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	UACAAGUAUUUUAGU		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	UAUUGUUUUUAUGUGUGUUGAGUAU	ORF1ab nsp6	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	GUUUUUAUGUGUGUUGAGUAUU		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	UUUAUGUGUGUUGAGUAUUG		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	UGUGUGUUGAGUAUUGCCC		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	CAAUAGACAGUUUCAUCAAAAA	ORF1ab nsp12 (RdRp)	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		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	ACACUUAUGAUUGAAC		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	AAGGCAUUAAAAUAUUUGCCUA	ORF1ab nsp13 (helicase)	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		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		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	UAAAAACACAGUUCAAUUAUU		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
Region 27	**27132**	**15**	**UAUAAAUUAAACACA**	**M**	$\mathbf{0.70 \pm 0.01}$	**0.68 – 0.72**	$\mathbf{0.22 \pm 0.01}$	**0.65 – 0.79**	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	N	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	GAAUACACCAAAAGA		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	AAUACACCAAAAGAUCACAU		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	ACACCAAAAGAUCACAUU		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	UACAAUGUAACACAA		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	ΔG(kcal/mol)	Folding Range	# Mut. Sites	Identity / Exact		
Region 1	1094	19	UUAAAUUCCAUAAUCAAGA		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	ACUGAGAAUUUGACUAAAG		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	UUGUUAAAAUUUAUUGUC	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	AUUUACAACCAUUAGAACA		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	UUCAACUUCUUAAGA		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	UGACCCUAUACAUUCUUUA		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	ACAUGCAGAAGAAACAC	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	CAUGCAGAAGAAACACGCAAA	(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	AAGUGUAUAUUACACUAGUAAUCCUA		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	UAAACAUAUAACUUCUAAAGAAA		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	ACCAGCAAAUAAUAGUUUAAAAAUUACAGAAG		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	UUUUCUUAACAAAGU		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	CUUCUUACAUCGUUGAUAG		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	UAAAUUUACUUACUA	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	UUUACUCAUUCUUAC		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	UUUUUGUACUUGUUACUU	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	AUAUUAACCUUGACCA	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	UUAAAAUGUUUUAUAA	(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	AUUACAAUCUAGUCAA	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	CUAUAAUAUGAGAGUUAUACAUUU	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	CAGAACUCAAUUACCC		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	UAAUAACGCUACUAAU		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	GUCUGUGAAUUUCAA		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	UAACAAUCUUGAUUCUAAGGUUGG		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	UUCCUUUACAAUCAUA	S	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	CUUCAUCAAACAAUA		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	AAUGAUUGCUCAAUACACU		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	GACUCACUUUCUUCCAC		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	GUUAGAUAAAUAUUUUAA		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	UAAAAAUUUAUCUAAG	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	UUUACUCAUUCUUAC	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	CCCACAAAUUUUACUA	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
27129	**18**	**AACUAUAAAUUAAACACA**	**M**	**0.77**	**0.76 ± 0.04**	**0.62 – 0.78**
28433	15	ACCGCUCUCACUCAA	N	0.55	0.27 ± 0.27	0.00 – 0.70
28691	17	AAUACACCAAAAGAUCA	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