# Table of Contents

**Supplementary Table 1.** (in a separate Excel file)

**Experimental maps in the Cryo-READ benchmark dataset.** The file includes three tabs, “training”, “validation”, and “testing”, which provides information of clustered EM maps in the training, the validation, and the testing set, respectively. The 1st column, “Cluster” indicates the cluster ID. The 2nd column, “PDB-ID (EMD-ID)”, shows the structures that belong to the cluster with their PDB ID and EMD ID.

**Supplementary Table 2.** (in a separate Excel file)

**Detection performance of deep learning model in Cryo-READ for individual maps.** “grid stage 1” and “grid stage 2” tables provide grid-level performance from the corresponding network stage, which correspond to Extended Data 3. Tables named moiety stagex (x = 1, 2) are results for the nucleotide-moiety-based accuracy, which correspond to Fig. 2a. Based on the clustering information in Supplementary Table 1, we also further reported group-based performance by averaging metrics inside the group and then re-average across the different groups. “grid stage x group” and “moiety stage x group” report the clustering based performance of grid-level and nucleotide-moiety-based performance, respectively.

The EMD-ID column represents each entry’s EMD ID. x_recall indicates the fraction of correctly identified moieties, where the denominator is the total number of the moieties or grid points from the reference structures (i.e. the PDB entry). On the other hand, x_precision provides the fraction of correctly predicted moiety, where the denominator is the total number of predicted grid points of that type.  “nan” in the protein_recall indicates no protein structures are included in the reference PDB entry. The last column of “grid stage x” tables shows the overall accuracy of non-background grid points. The average values are shown at the bottom of each column in the tables.

### **Supplementary Table 3.** (in a separate Excel file)

**Atomic structure evaluation for the 68 cryo-EM maps in the testing set.** The "without sequence" table provides the performance of sugar and phosphate atoms; the "with sequence no refine" and "with sequence after refine" tables provide the performance of backbone atoms before and after the phenix structure refinement. This table corresponds to Fig. 2.

The EMD-ID, and PDB-ID columns represent each entry's EMD ID and PDB-ID, respectively. The resolution column shows the reported resolution of the deposited map. #Nuc indicates the number of nucleotides in the deposited structure. x_Recall indicates the fraction of correctly identified nucleotides, where the denominator is the total number of nucleotides in the reference structure. On the other hand, x_Precision provides the fraction of correctly predicted nucleotides in the predicted structure, where the denominator is the total number of nucleotides in the predicted structure. x here can be backbone, sequence, and sequence(match). RMSD is the root-mean-squared-deviation between backbone atoms of matched nucleotides. Base-RMSD is RMSD between base center of matched nucleotides. Matched nucleotides are pairs of predicted and native nucleotides that have an average atom pair distance less than 5 Å.

The average values are shown at the bottom of each column in the tables.

### **Supplementary Table 4.** (in a separate Excel file)

**Map and structure information of SARS-Cov-2 benchmark.** The EMD-ID and PDB-ID columns represent each entry's EMD ID, and PDB-ID, respectively. the Resolution column shows the reported resolution of the deposited map. Contour column provides the author suggested contour level of the EM map. Source column indicates how we collected the corresponding maps, "both" indicates it is identified in both queries from RCSB and EMDB, "RCSB" indicates it is only identified by a query from RCSB, "EMDB" indicates it is only identified by query from EMDB.

### **Supplementary Table 5.** (in a separate Excel file)

**Atomic structure evaluation in the 58 cryo-EM maps in SARS-Cov-2 benchmark.** "without sequence" table provides the performance of sugar and phosphate atoms; "with sequence no refine" and "with sequence after refine" tables provide the performance of backbone atoms before and after the phenix structure refinement. This table corresponds Fig. 4.

The EMD-ID (PDB-ID) column represents each entry's EMD ID (PDB-ID). The resolution column shows the reported resolution of the deposited map. #Nuc indicates the number of nucleotides in the deposited structure. x_Recall indicates the fraction of correctly identified nucleotides, where the denominator is the total number of nucleotides in the reference structure. On the other hand, x_Precision provides the fraction of correctly predicted nucleotides in the predicted structure, where the denominator is the total number of nucleotides in the predicted structure. x here can be backbone, sequence, and sequence(match). RMSD indicated the root-mean-squared-deviation between atoms of matched nucleotides. The average values are shown at the bottom of each column in the tables.

## **Supplementary Table 6.** (in a separate Excel file)

**Atomic structure evaluation in the 68 cryo-EM benchmark for Phenix.** It includes two tables, “with mask” table includes the evaluation of Phneix based on our masked map (removing protein density regions by our detection) for core atoms, “without mask” table includes the evaluation of pure Phenix. This table corresponds Fig. 5 and Extended Data 8.

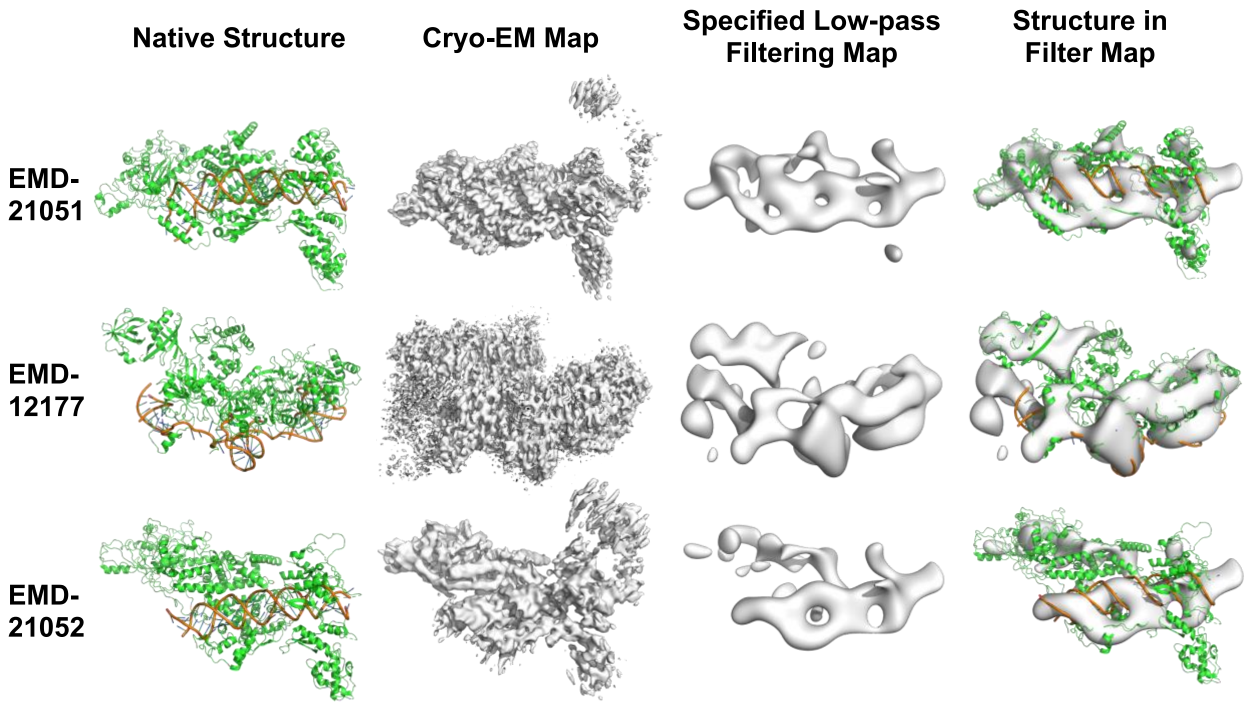
The EMD-ID (PDB-ID) column represents each entry’s EMD ID (PDB-ID). Resolution column shows the reported resolution of the deposited map. #Nuc indicates the number of nucleotides in the deposited structure. x_Recall indicates the fraction of correctly identified nucleotides, where the denominator is the total number of nucleotides in the reference structure. On the other hand, x_Precision provides the fraction of correctly predicted nucleotides in the predicted structure, where the denominator is the total number of nucleotides in the predicted structure. x here can be backbone, sequence, and sequence(match). RMSD indicated the root-mean-squared-deviation between atoms of matched nucleotides. The average values are shown at the bottom of each column in the tables.

## **Supplementary Table 7.** (in a separate Excel file)

**Comparison between CryoREAD and auto-DRRAFTER.** It includes two tabs, the “RNA” tab compares the performance of CryoREAD and auto-DRRAFTER under 4 different settings, while the “DNA” tab compares the performance of CryoREAD and auto-DRRAFTER (provided DNA-only map region and ground truth secondary structure information). For RNA, we selected 11 maps from our benchmark dataset that includes RNA of less than 400 nucleotides. Out of them, results of the 4 maps are shown because auto-DRRAFTER did not run for the rest of 7 maps. Similarly, results of 6 maps are shown for DNA. In the testing set, we selected maps with DNA less than 400 nucleotides. Out of the 10 maps selected, auto-DDRAFTER was able to process these 4 maps reported in the table.

For “RNA” tabs, it includes five tables. “CryoREAD” table includes the evaluation of CryoREAD with full map (no prior map segmentation). 4 results for auto-DRRAFTER: “auto-DRRAFTER: RNA-only map region + ground truth secondary structure”, we ran Auto-DRRAFTER with a RNA-only map region (segmented by a proper low-pass filter threshold) and ground truth secondary structure information (secondary structure information taken from the PDB entry of the RNA); “auto-DRRAFTER: RNA-only map region + predicted secondary structure”, auto_DRRAFTER ran on a RNA-only map region with predicted secondary information by RNAFold; “Auto-DRRAFTER: Full map + ground truth secondary structure auto-DRRAFTER ran with full map and ground truth secondary structure information; “auto-DRRAFTER: Full map + predicted secondary structure”, auto-DRRAFTER ran on a full map with predicted secondary information.

The EMD-ID (PDB-ID) column represents each entry’s EMD ID (PDB-ID). Resolution column shows the reported resolution of the deposited map. #Nuc indicates the number of nucleotides in the deposited structure. x_Recall indicates the fraction of correctly identified nucleotides, where the denominator is the total number of nucleotides in the reference structure. x_Precision provides the fraction of correctly predicted nucleotides in the predicted structure, where the denominator is the total number of nucleotides in the predicted structure. x here can be backbone, sequence, and sequence(match) (See Methods for definition). RMSD indicated the root-mean-squared-deviation between atoms of matched nucleotides. The average values are shown at the bottom of each column in the tables.

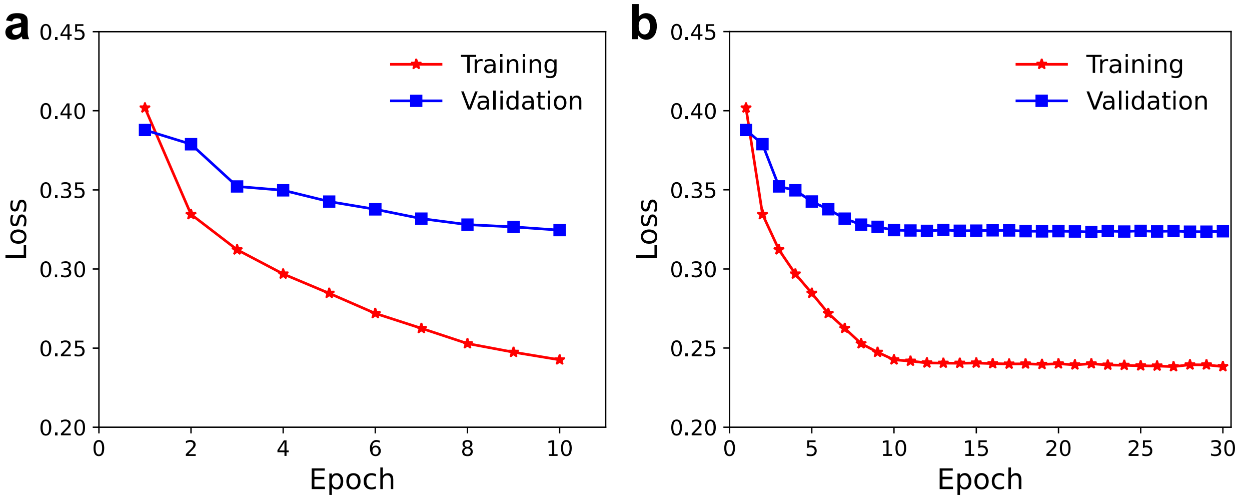


## Supplementary Fig 1

**Input EM maps used to run auto-DRRAFTER**. We ran auto-DRRAFER on 11 maps, among which four of them were processed and produced RNA structure models. The results are reported in Supplementary Table 7. Among the four maps, these three maps include a protein-RNA complex. Therefore, following the auto-DRRAFTER's protocol, we applied a low-pass filter to extract a map region which contains the RNA. The low-pass filter was applied with a threshold of 5%, 10%, …, 90%, and 95% of the max density of the map and the threshold that yielded the best overlap and distinction for RNA was chosen.

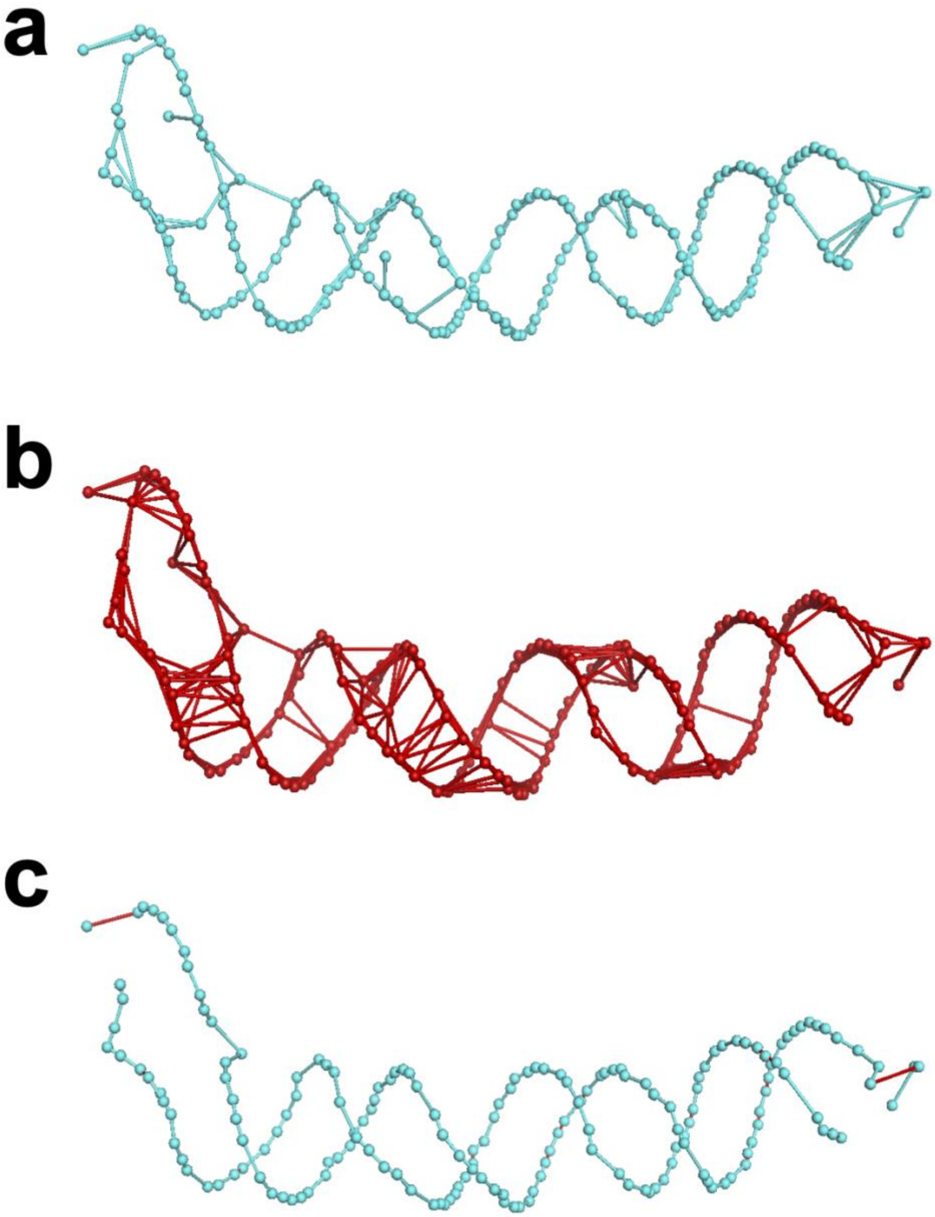
In this figure, the three columns from left to right are 1) the structure of the protein-RNA complex; 2) the input cryo-EM map; 3) the map region processed by the low-pass filter.

To apply low-pass filter, we used the suggested command in Auto-DRRAFTER documentation (https://www.rosettacommons.org/docs/latest/application_documentation/rna/auto-drrafter) "python $ROSETTA/main/source/src/apps/public/DRRAFTER/auto-DRRAFTER_setup.py -map_thr 30 -full_dens_map {input_map} -full_dens_map_reso {resolution} -fasta {input_fasta} -secstruct {input_secondary_info} -out_pref mini_example -rosetta_directory $ROSETTA/main/source/bin/ -nstruct_per_job 100 -cycles 1000 -fit_only_one_helix -rosetta_extension .static.linuxgccrelease -just_low_pass". to generate low-pass filtering map and determine a threshold to focus on RNA map regions. The three are 1) EMD-21051, PDB-ID: 6v5b, Resolution: 3.7 Å, Threshold: 0.52; 2)EMD-12177, PDB-ID: 7BGB, Resolution: 3.4 Å, Threshold: 0.018; 3) EMD-21052, PDB-ID: 6v5C, Resolution: 4.4 Å, Threshold: 0.57.



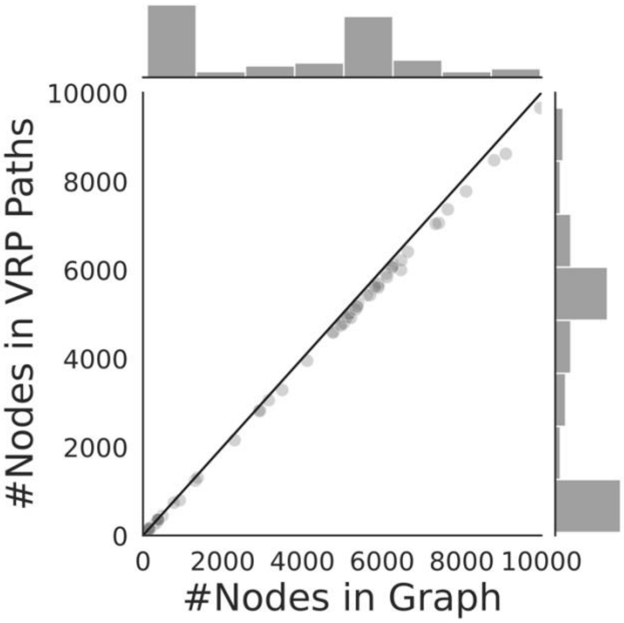
### Supplementary Fig 2

**Training and validation loss.** The red curve shows the training loss, and the blue curve is the validation loss. **a,** training and validation loss plot up to 10 epochs. **b,** training and validation extended to 30 epochs.



**Supplementary Fig 3**

**Illustration of the main chain tracing solved as a VRP.** The map used is PDB-ID: 6v5b; EMD-ID: 21051, Resolution: 3.7 Å. **a.** A sugar graph with edges where only adjacent grid point clusters are connected. **b.** Graph formed with edges with a length less than R (= 10 Å) and other associated conditions mentioned in Methods. **c.** The traced paths by the VRP solver. Edges in cyan are included in the graph in panel a, while red edges are only included in the graph in panel b.



**Supplementary Fig 4**

**Comparison of number of representative nodes before and after VRP process.** For most targets, the drop ratio of nodes is within 5%. Only 8 small targets (#Nuc<=200) have drop ratio larger than 10%. With visual inspection on those targets, we found all those dropped nodes are reasonable.