Supplementary Information

Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences

Section I: Motif template database augmentation

The motif template database used in Vfold3D program is enlarged by opening canonical base pair in helices, forming non-canonical base pair in loops, and reassigning the 5'-end. We will illustrate the database augmentation process using an internal loop as an example.

(a) **Template augmentation by opening canonical base pairs in helices**.

As shown in Fig. S1, from left to right, one, one, and two canonical base pairs connected to the loops (blue) are opened and the internal loops are thus extended, forming three additional internal loop motifs. Up to one canonical base pair in a helix can be opened. For an *N*-way junction motif, motif templates can be enlarged 2^N times by opening canonical base pairs in helices.

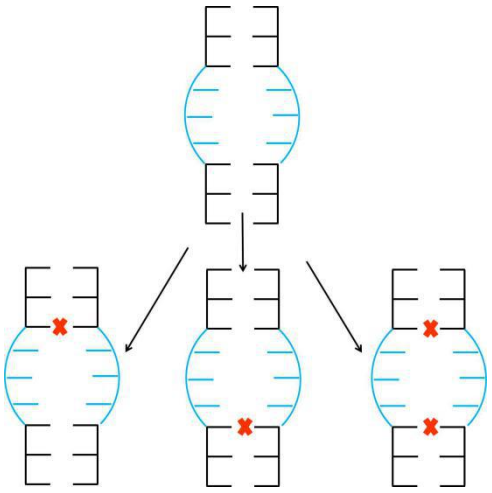

Figure S1. Internal loop motif template augmentation by opening canonical base pairs in helices. The black ladders represent the canonical base pairs in helices, and the blue ladders represent the internal loops. Only the canonical base pairs connected to the loops can be opened. From left to right, one, one, and two canonical base pairs are opened marked by the red cross.

(b) **Template augmentation by forming non-canonical base pairs in loops**.

As shown in Fig. S2, from left to right, one, one, and two non-canonical base pairs connected to the helices (black) are possibly formed and the helices are thus extended, forming three additional possible internal loop motifs. Only the non-canonical base pair connected to the helices can be possibly formed. If the presumed non-canonical base pairs do not exist in the experimentally solved structures, the added motif will be excluded. For an *N*-way junction motif, motif templates can be enlarged up to 2^N times by forming non-canonical base pairs in loops.

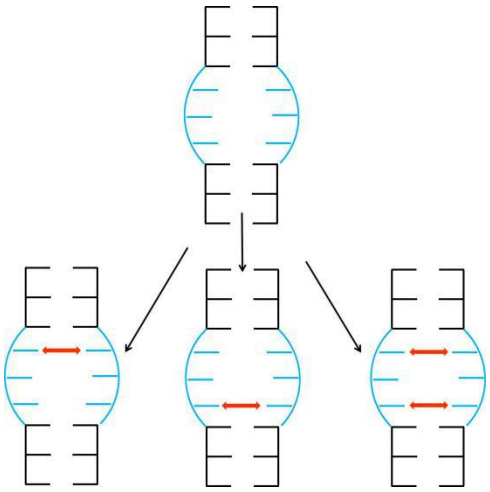

Figure S2. Internal loop motif template augmentation by forming non-canonical base pairs in loops. The black ladders represent the canonical base pairs in helices, and the blue ladders represent the internal loops. Only the non-canonical base pairs connected the helices can be formed. From the left to right, one, one, and two non-canonical base pairs are possibly formed marked by the red arrows.

(C) **Template augmentation by opening canonical base pairs in helices and forming non-canonical base pairs in loops simultaneously**.

The canonical base pair in helices can be opened and the non-canonical base pairs in loops can be formed at the same time as shown in Fig. S3. In both the lower left and right panels, one canonical base pair is opened and one non-canonical base pair is possibly formed, forming two additional possible internal loop motifs. If the presumed non-canonical base pairs do not exist in the experimentally solved structures, the added motif will be excluded. For an *N*-way junction motif, motif templates can be enlarged up to $(3^N-2^{N+1}+2)$ times by opening canonical base pairs in helices and forming non-canonical base pairs in loops at the same time.

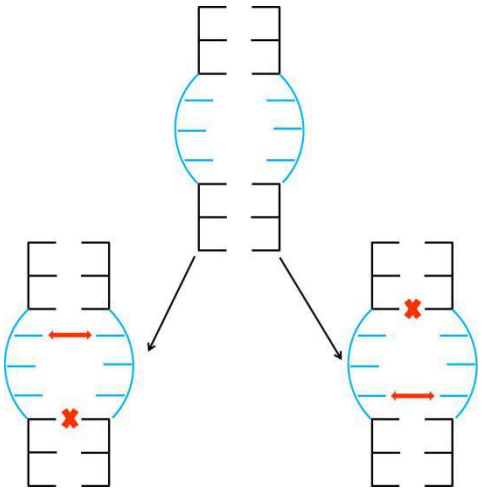

Figure S3. Internal loop motif template augmentation by opening canonical base pairs in helices and forming non-canonical base pairs in loops at the same time. The black ladders represent the canonical base pairs in helices, and the blue ladders represent the internal loops.

Only the canonical base pairs connected to the loops can be opened. Only the non-canonical base pairs connected the helices can be formed. From the left to right, one canonical base pair is opened and one non-canonical base pair is possibly formed marked by the red arrows.

(d) Template augmentation by reassigning the 5'-end

As shown in Fig. S4, a new internal loop motif can be created by reassigning the 5'-end. In an *N*-way junction motif, motif templates can be enlarged *N* times by reassigning the 5'-end.

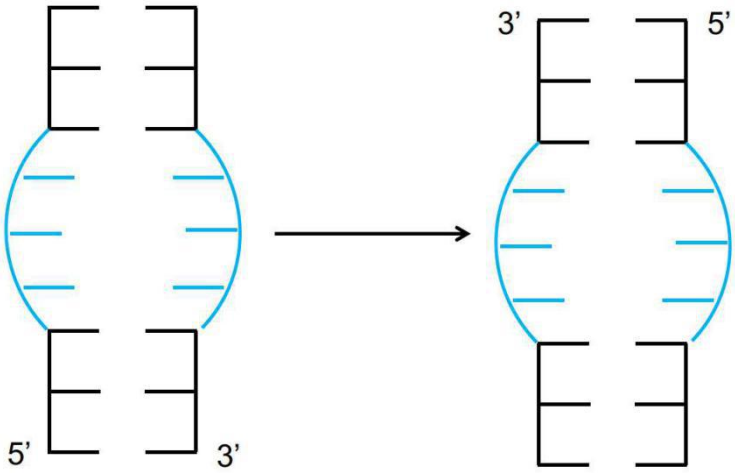

Figure S4. Internal loop motif template augmentation by reassigning the 5'-end.

Theoretically, for an *N*-way junction motif, motif templates can be enlarged up to $N*3^N$ times by the motif augmentation methods described above. The practical augmentation is listed in Table S1.

Table S1. The number of motifs used in Vfold3D program before and after motif template augmentation.

Motif name	Number before augmentation	Number after augmentation	Enlarged # times
Internal/bulge loop	9006	117,647	13.06
3-way junction	2385	101,986	42.76
4-way junction	1393	177,090	127.13
5-way junction	709	244,805	345.28
6-way junction	142	143,096	1007.72
7-way junction	143	562,760	3935.38
Pseudoknot	213	2,485	11.67
HP-HP-Kissing	324	12,404	38.28

Section II: Performance comparison with the RNAComposer server

To test the performance of our server and compare it with the RNAComposer server, we compiled a dataset of 92 single-stranded RNAs that cover different structural topologies including hairpin/internal loops, 3-, 4-, and 5-way junctions, and pseudoknots (PK). The native structures were excluded from the template library in both servers. The default secondary structure prediction method in the RNAComposer server, CentroidFold, was used for the non-PK RNAs and the method IPknot was used for the PK RNAs. Fig. S5 shows the RMSDs of the predicted 3D structures relative to their native structures. We can see that, in general, our server (red line) can give smaller RMSDs than the RNAComposer server (blue line). Table S2 shows the detailed information and results for the 92 RNAs. We can see that the average RMSDs for the hairpin/internal loop structures are 6.0 and 8.4 Å obtained by the Vfold-Pipeline and RNAComposer servers, respectively. For the 3-way junction structures, they are 9.7 and 19.9 Å. For the 4-way junction structures, they are 7.4 and 21.7 Å. For the 5-way junction structures, they are 15.0 and 21.6 Å. For the PK structures, they are 15.6 and 19.4 Å. From the above results, we can see that for the different topologies, our server works better than the RNAComposer server on average.

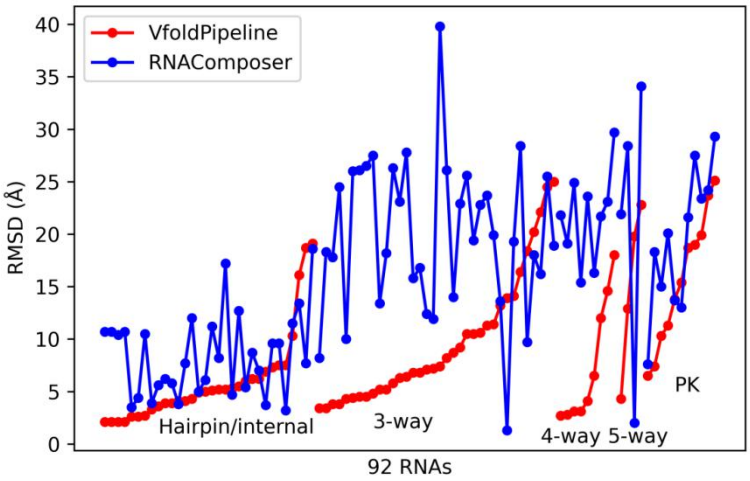

Figure S5. The RMSDs for the predicted 3D structures relative to the native structures for the 92 RNAs. The red and blue lines represent the results of the Vfold-Pipeline and RNAComposer servers, respectively. The five separate red lines stand for five structural topologies indicated by the labels below. Hairpin/internal means that there are only hairpin or/and internal loops in the structures. 3-way, 4-way, and 5-way mean that there are junctions in the structures. PK means that there are pseudoknots in the structures.

Table S2. The detailed information and RMSDs for the 92 RNAs. H/I is short for hairpin/internal loops, and PK is short for pseudoknots.

PDB code	Topology	Size	RMSD (Å) by Vfold-Pipeline	RMSD (Å) by RNAComposer
1dul	H/I	48	2.1	10.7
2pxb	H/I	49	2.1	10.7
2pxf	H/I	49	2.1	10.4
2pxt	H/I	49	2.1	10.7
1zc5	H/I	41	2.6	3.5
1z2j	H/I	45	2.6	4.4
2pxe	H/I	49	2.7	10.5
2kx8	H/I	42	3.3	3.9
2ke6	H/I	48	3.6	5.6
1s03	H/I	47	3.9	6.2
2kuw	H/I	48	3.9	5.8
2n4l	H/I	53	4.1	3.8
5lyu	H/I	57	4.1	7.7
1cq5	H/I	43	4.3	12
2nbz	H/I	40	4.9	5
2l2j	H/I	42	5	6.1
2hua	H/I	40	5.1	11.2
4c7o	H/I	48	5.2	8.2
2mqt	H/I	68	5.2	17.2
5m0h	H/I	42	5.3	4.7
2l3j	H/I	71	5.5	12.7
5v16	H/I	41	6	5.4
4pmi	H/I	40	6.2	8.7
2n6t	H/I	42	6.2	7
1mnx	H/I	42	6.9	3.7
2fey	H/I	43	7.3	9.6
1xjr	H/I	46	7.5	9.6
4k27	H/I	55	7.5	3.2
1kxk	H/I	70	10.3	11.5
2kzl	H/I	55	16.1	13.4
2au4	H/I	41	18.7	7.7

1p5p	H/I	77	19.1	18.6
Average	**H/I**	**49**	**6.0**	**8.4**
1e8o	3-way	49	3.4	8.2
5dar	3-way	74	3.4	18.3
1mms	3-way	58	3.8	17.8
5axm	3-way	72	3.8	24.5
2hgh	3-way	55	4.3	10
3ds7	3-way	67	4.4	26
2ees	3-way	67	4.5	26.1
3rkf	3-way	67	4.5	26.5
1y27	3-way	66	4.8	27.5
2nc1	3-way	67	5.2	13.4
3owi	3-way	86	5.2	18.2
3la5	3-way	71	5.8	26.3
1y26	3-way	71	6.3	23.1
3ivn	3-way	69	6.4	27.8
3egz	3-way	65	6.8	15.8
2cky	3-way	77	6.8	16.8
1dk1	3-way	57	7.1	12.4
3r4f	3-way	66	7.2	11.9
5e54	3-way	65	7.4	39.8
4wfm	3-way	101	8.2	26.1
2mhi	3-way	53	8.7	14
3ski	3-way	66	9.2	22.9
3iwn	3-way	93	10.5	25.6
3ndb	3-way	136	10.5	19.4
3sd3	3-way	89	10.6	22.8
3k0j	3-way	87	11.3	23.7
5tpy	3-way	71	11.4	19.9
2nbx	3-way	108	13.2	13.6
3e5c	3-way	52	13.9	1.3
4pqv	3-way	68	14.1	19.3
4uyk	3-way	133	16.4	28.4
2oiu	3-way	71	18.4	9.7
4yb0	3-way	83	20.2	18

5m73	3-way	144	22.1	16.2
2n1q	3-way	155	24.5	25.5
3pdr	3-way	161	25	18.9
Average	**3-way**	**82**	**9.7**	**19.9**
2du3	4-way	71	2.7	21.8
1j1u	4-way	74	2.8	19.1
1ffy	4-way	75	3.1	24.9
2zue	4-way	75	3.1	15.4
3wqy	4-way	75	4.1	23.6
1u0b	4-way	74	6.5	16.3
1h4q	4-way	65	12	21.7
4aob	4-way	94	14.6	23.1
3d0u	4-way	161	18	29.7
Average	**4-way**	**85**	**7.4**	**21.7**
3w3s	5-way	98	4.3	21.9
1h3e	5-way	79	12.9	28.4
3am1	5-way	81	19.8	2
2zzm	5-way	84	22.8	34.1
Average	**5-way**	**86**	**15.0**	**21.6**
1a60	PK	44	6.5	7.6
1e95	PK	36	7.4	18.3
2n8v	PK	70	10.3	15
5kh8	PK	47	11.3	20.1
3vrs	PK	52	13.8	13.7
4jf2	PK	76	15.4	13
4znp	PK	73	18.7	21.6
2qwy	PK	52	19	27.5
1sjf	PK	74	19.9	23.4
4qk8	PK	120	23.7	24.2
3q3z	PK	74	25.1	29.3
Average	**PK**	**65**	**15.6**	**19.4**

Fig. S6 shows five examples of the well-predicted 3D structures (green) aligned over the native structures (red) for the five structural topologies.

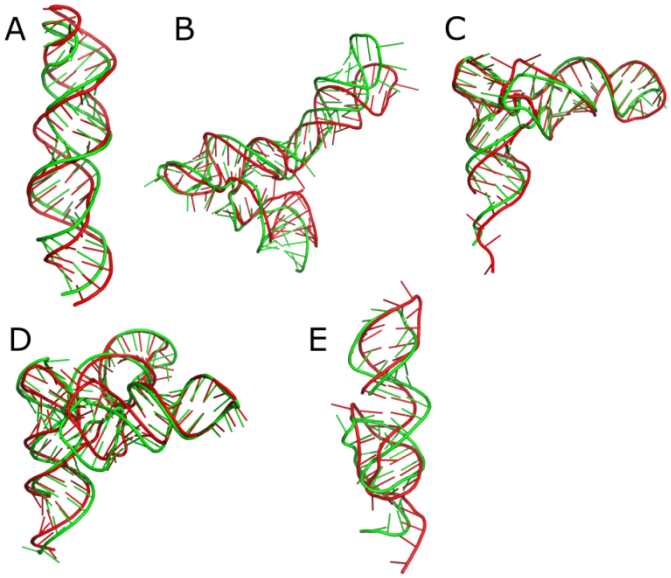

Figure S6. The aligned native (red) and predicted (green) 3D structures for five RNAs with five different topologies. (A) RNA 2ke6 with only hairpin/internal loops. The RSMD is 3.6 Å. (B) RNA 3owi with the 3-way junction structure. The RMSD is 5.2 Å. (C) RNA 3wqy with the 4-way junction structure. The RMSD is 4.1 Å. (D) RNA 3w3s with the 5-way junction structure. The RMSD is 4.3 Å. (E) RNA 1a60 with the pseudoknotted structure. The RMSD is 6.5 Å.

We examined the cases with large RMSDs and found that many failures were due to the wrong predicted 2D structures. Take tRNA 3wqy as an example. Its native 2D structure is a 4-way junction in Fig. S7 (left). Our vfold2D model predicts that it is a 4-way junction, but the junction loop is wrong as shown in Fig. S7 (middle), leading to a wrong selection of motif templates. The default secondary structure prediction method, CentroidFold, in the RNAComposer server predicts that it has a long internal loop as shown in Fig. S8 (right), leading to a completely wrong 3D structure with an RMSD of 23.6 Å. Using the RNA 3wqy sequence to search in the Rfam database, it is easy to know that this RNA belongs to the tRNA family and to get the right 2D constraints as shown in Fig. S8. With the rigFigure S10 shows the running time for the 92 test cases. The 2D structure prediction for the non-PK structures is very quick (almost done in 30 seconds) in Fig. S10 (top). The 2D structure prediction for the PK structures is very quick (almost done in 50 seconds) for sequence lengths less than 120 nts, and it takes longer for larger RNAs (almost done in 1,000 seconds for sequence lengths less than 160 nts) in Fig. S10 (top). For 3D structure prediction, the Vfold-Pipeline server tries to first run the Vfold3D method, and if the Vfold3D method fails, then the server starts to run the VfoldLA method. The Vfold3D method is faster than the VfoldLA method since in the VfoldLA method each loop needs to be assembled instead of the whole motifs in the Vfold3D method, and after assembly, short-time coarse-grained molecular dynamics simulations need to be performed. Therefore, the time for running 3D structure prediction depends on which prediction engine is launched. As shown in Fig. S10 (middle), the 3D structure prediction for the PK structures takes more time than that for the non-PK structures in that PK 3D structure prediction more likely uses the VfoldLA method due to the lack of motif templates for the Vfold3D method. The 3D structure prediction for the non-PK structures is almost done in 500 seconds, while that for the PK structures is almost done in 1,000 seconds. Generally, the total time for the prediction of the non-PK and PK structures shorter than 160 nts in our server is less than 1,000 and 2,000 seconds, respectively, as shown in Fig. S10 (bottom).

ht 2D constraints, the RMSD of the predicted 3D structure by our server is 4.1 Å. In the above test dataset of 92 RNAs, half of them can be assigned to a family according to the Rfam database.

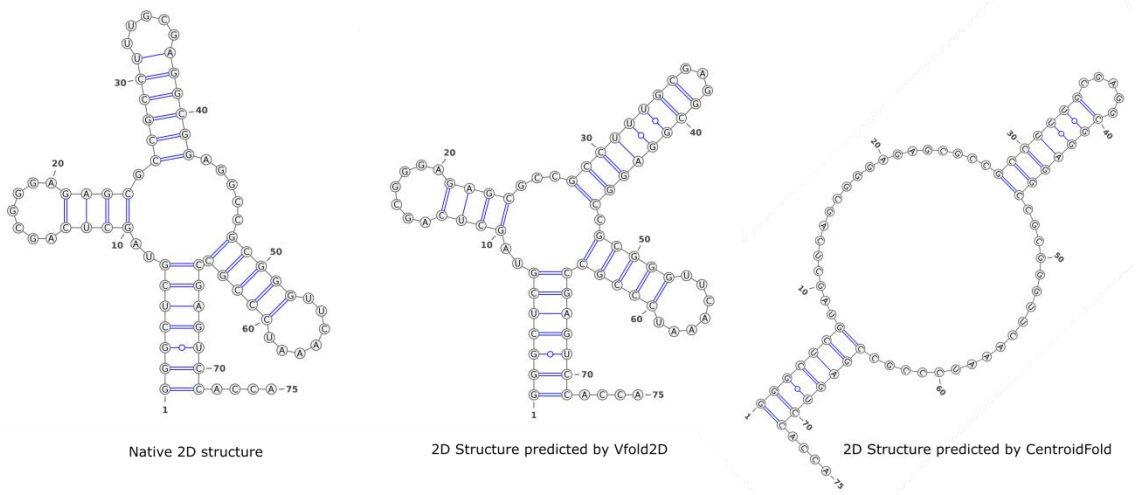

Figure S7. The native (left) and predicted 2D structures by Vfold2D (middle) and CentroidFold (right) for the tRNA 3wqy.

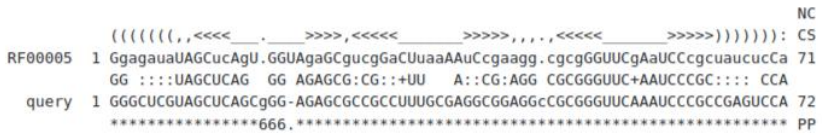

Figure S8. The sequence search result in the Rfam database for the tRNA 3wqy.

In addition to the wrong 2D structures, the lack of the right motif templates in our template database also leads to the large RMSDs. Take the longest RNA, 3pdr, as an example. The predicted 2D structure having two three-way junctions is right with the help of the Rfam database. But the wrong motif template for the first 3-way junction loop makes the predicted 3D structure an extended structure, while the native structure is bent over this 3-way junction as shown in Fig. S9.

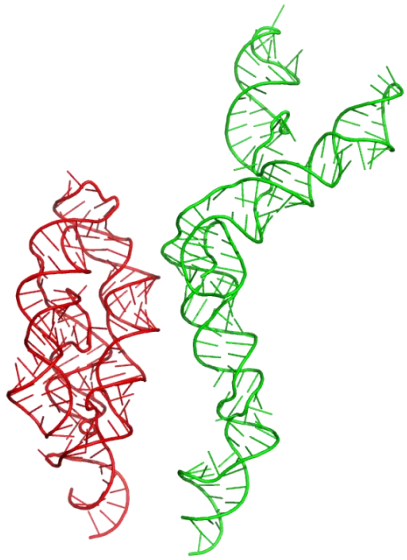

Figure S9. The native (red) and predicted 3D structure (green) by the Vfold-Pipeline server for the RNA 3pdr.

Figure S10 shows the running time for the 92 test cases. The 2D structure prediction for non-PK structures is very quick (almost completed within 30 seconds) in Fig. S10 (top). The 2D structure prediction for PK structures is fast (almost completed within 50 seconds) for sequence lengths less than 120 nts, and it takes longer for larger RNAs, e.g., in 1,000 seconds for sequence lengths less than 160 nts as shown in Fig. S10 (top). For 3D structure prediction, the Vfold-Pipeline server first uses the Vfold3D method, and if the Vfold3D method fails (due to lack of templates), then the server starts to run the VfoldLA method. The Vfold3D method is faster than the VfoldLA method. This is because unlike Vfold3D, which searches for templates for motifs, the VfoldLA method needs to assemble a motif (loop) from templates of individual strands. Therefore, the time for running 3D structure prediction depends on which prediction engine is launched. As shown in Fig. S10 (middle), the 3D structure prediction for the PK structures takes longer than that for the non-PK structures because PK 3D structure prediction more likely uses the VfoldLA method due to the lack of motif templates. The 3D structure prediction for the non-PK structures can almost be finished in 500 seconds, while that for the PK structures would require 1,000 seconds. Generally, the total time for the prediction pipeline for the non-PK and PK structures is less than 1,000 and 2,000 seconds, respectively, as shown in Fig. S10 (bottom).

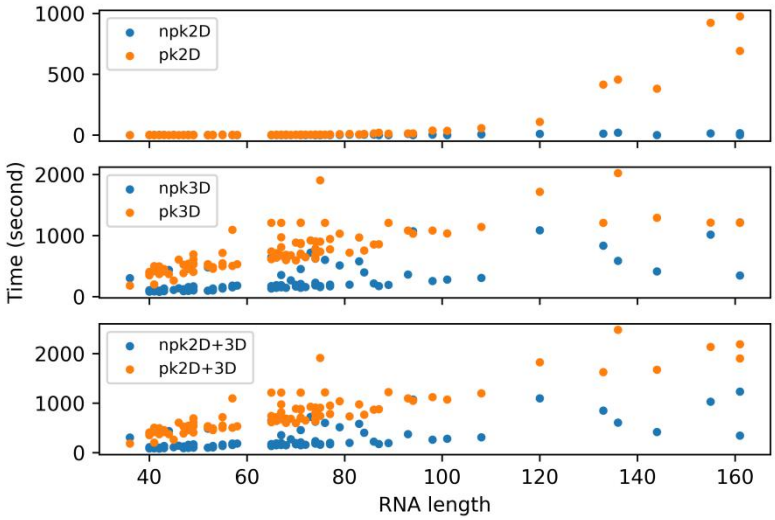

Figure S10. The running time for the 2D and 3D structure prediction and the pipeline prediction of the non-PK and PK structures for the dataset of 92 RNAs. Only one 2D structure is used for 3D structure prediction.

Section III: Input and output Snapshots of the Vfold-Pipeline Server

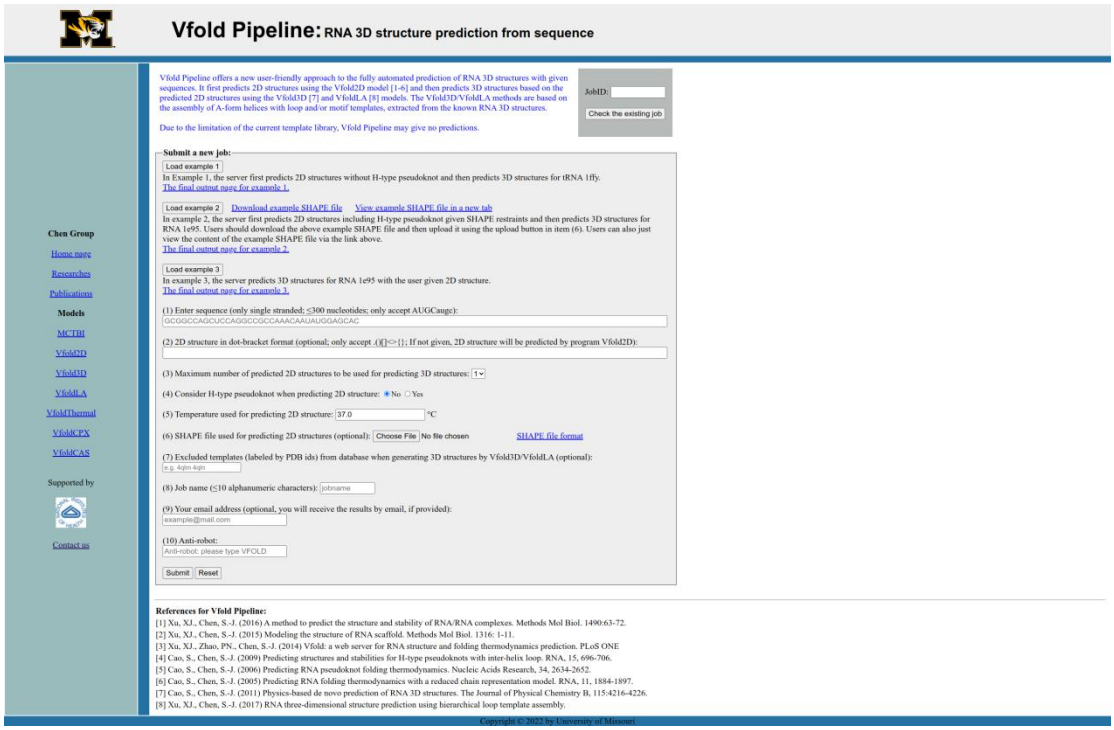

Figure S11. Input snapshot of the Vfold-Pipeline server, including the following input information: (1) Sequence, (2) 2D structure (optional), (3) the maximum number of predicted 2D structures to be used for predicting 3D structures, (4) if considering H-type pseudoknots when predicting 2D structures, (5) temperature used for 2D structure prediction, (6) optional SHAPE file for 2D structure prediction and the SHAPE file format, (7) excluded motif/loop templates used in the Vfold3D/VfoldLA programs, (8) job name, (9) email address, (10) anti-robot code. Moreover, it provides three input examples and their corresponding result pages. In Example 1, the server first predicts 2D structures without H-type pseudoknots and then predicts 3D structures for tRNA 1ffy. In example 2, the server first predicts 2D structures including H-type pseudoknots given SHAPE restraints and then predicts 3D structures for RNA 1e95. In example 3, the server predicts 3D structures for RNA 1e95 with the user given 2D structure.

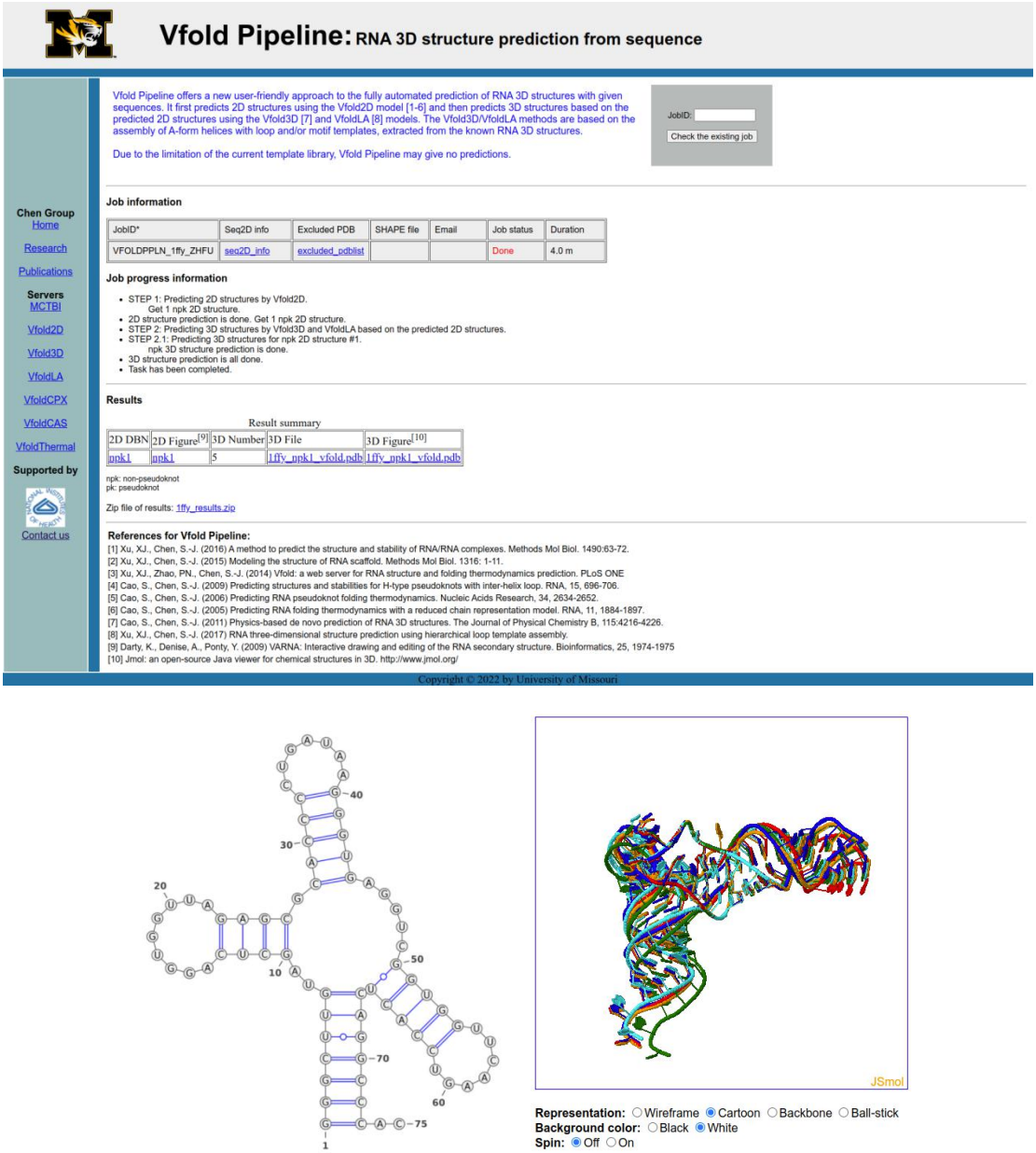

Figure S12. Output snapshot of the Vfold-Pipeline server for the Example 1 (tRNA 1ffy) on the main web page, including the job information, the progress information updated during calculation, the predicted 2D structures in DBN (dot-bracket notation) format and the corresponding figures drawn by the VARNA applet [1], the number of predicted 3D models for each predicted 2D structure, and the predicted 3D structures in PDB format and the corresponding 3D figures shown in JSmol [2], and a downloadable compressed file containing all the results. The predicted 2D and 3D structures (bottom) can be viewed in a new tab by clicking the links to the 2D and 3D Figures in the "Result summary box".

References:

[1] Darty, K. et al. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974-1975.

[2] Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/