Supplementary Data for An Extended Dual Graph Library and Partitioning Algorithm Applicable to Pseudoknotted RNA Structures

Swati Jain[1], Sera Saju[1], Louis Petingi[2], and Tamar Schlick[1,3,4,*]

[1]Department of Chemistry, New York University, New York, NY, USA
[2]Computer Science Department, College of Staten Island, City University of New York, New York, NY, USA
[3]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[4]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai, China
[*]Corresponding author: schlick@nyu.edu

S1 Differences between prior and expanded dual graph libraries for RNA structures

S1.1 Differences in dual graph IDs for existing RNA structures

To test our expanded dual graph library against known RNA structures, we updated our dataset of RNA structures (as used previously in our recent study of dual graphs [1]) to include all experimentally determined RNA 3D structures available in the PDB as of August 31, 2018 (see Subsection 2.4 of the main paper for details). We then assigned dual graph IDs to 2495 RNA 2D structure files with RNA graphs between 2 and 9 vertices (our labels identify topologies between 2 and 9 vertices) using both the prior and expanded dual graph libraries. For fair comparison, the 9 dual graph topologies that were manually added recently [1] were not considered as part of the prior library. Table S2 shows the distribution of dual graphs corresponding to 2495 full RNA 2D structures. For $V = 2$ to 6, there are no graph ID differences between using the prior and the new dual graph library (indicated by 0 in the Unassigned and Misclassified graphs columns). Below we elaborate the differences found in the graph ID assignment for existing RNA structures.

Different assignments occur to two RNA structures: short ribosomal RNA 1 (srRNA 1) of the 60S ribosomal subunit for *Trypanosoma cruzi* (PDB ID: 5T5H chain E): 7_1091 before and 7_2550 now (Figure S4a); and srRNA 1 of the large ribosomal subunit of *Trypanosoma*

brucei (PDB ID: 4V8M, chain BE): 8_3576 before and 8_14143 now (Figure S4b). Dual graph topologies 7_1091 and 8_3576 are non-isomorphic but with identical eigenvalue spectra as 7_2550 and 8_14143, respectively. Since the prior dual graph library retained only one dual graph topology per eigenvalue spectrum, it misclassified the above two structures. This is corrected in the expanded dual graph library.

The prior dual graph library also did not assign any graph IDs to 19 structures: one 8-vertex dual graph for bacteriophage MS2 genome fragments (PDB ID: 5TC1 chain R), and 18 9-vertex RNA dual graphs (indicated in the Unassigned graphs column in Table S2). The 8-vertex graph is now 8_12185. Of these 18 9-vertex graphs, 16 correspond to 4 graph IDs manually added in our recent study (9_38596 - 1 structure, 9_38597 - 13 structures, 9_38598 - 1 structure, 9_38599 - 1 structure); and the two remaining structure were now assigned graph IDs of 9_49214 and 9_86359. Twenty one previously unassigned or misclassified RNA dual graphs are now labeled correctly.

S1.2 Differences in dual graph IDs for RNA subgraphs

We applied our dual graph partitioning algorithm (Subsection 2.3 of the main paper) to all existing 3853 RNA 2D structure files with more than 1 vertex. Each subgraph between 2 and 9 vertices was then assigned a graph ID following the same procedure as for full RNA dual graphs, except that now the query Laplacian corresponds to the vertices and edges in the subgraph. Table S2 shows the distribution of the 311,294 subgraphs (excluding the full graphs corresponding to 2495 RNA 2D structure files) between 2 and 9 vertices from the 3853 RNA structure files. Similar to full graphs, there are no differences of graph IDs from 2 to 6 vertices. Described below are the differences found in the graph ID assignment between for RNA subgraphs.

The prior library misclassifies 910 subgraphs for $V = 7$, 697 subgraphs for $V = 8$, and 2651 subgraphs for $V = 9$. As with full graphs (see above), this misclassification is the result of non-isomorphic graphs with identical eigenvalue spectra being ignored in the prior dual graph library. These subgraphs correspond to 9, 12, and 13 distinct dual graph topologies for 7, 8, and 9 vertices, respectively, in the expanded dual graph library.

The number of unassigned graphs with the prior library increases significantly. For $V = 7$, the prior dual graph library does not assign any graph ID to 1737 subgraphs, 568 of which correspond to 7_2389 the dual graph topology for a 7-way junction manually identified in our recent study [1]. For $V = 8$, this number increases to 4120, and forms almost 50% (22,022 of 45,842) of the subgraphs for $V = 9$, 35 of which correspond to 9-vertex topologies manually identified in our recent study. Overall, the expanded dual graph library contains a total of 848 distinct dual graph topologies (8 for 7 vertices, 110 for 8 vertices, and 730 for 9 vertices) corresponding to existing RNA subgraphs that were missed by the prior library.

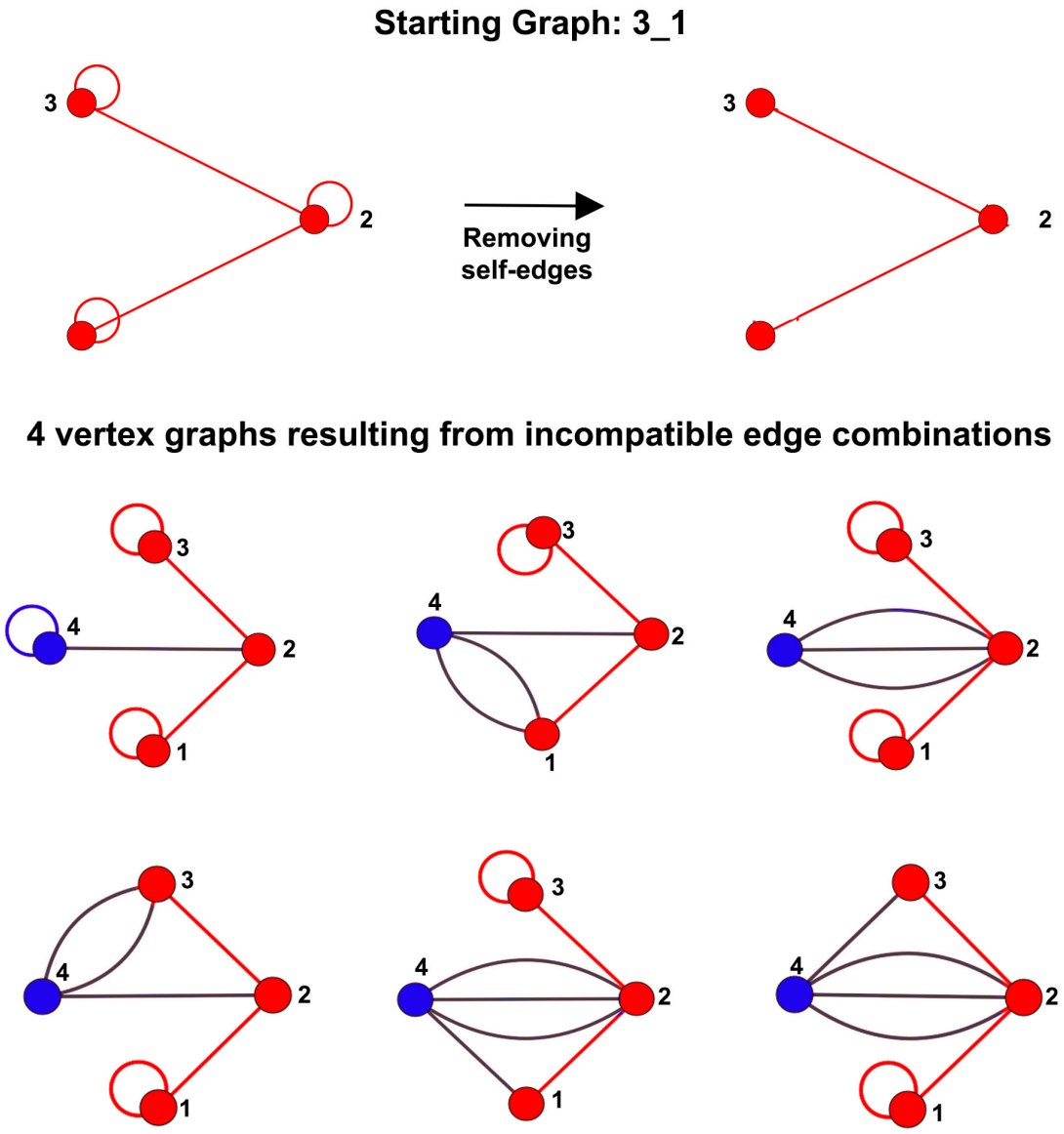

Figure S1: Incompatible 4-vertex graphs resulting from connecting one new vertex (blue segment) to the starting graph 3_1 (red). Of the 31 graphs generated using the edge combinations listed in Figure 3 of the main paper, 6 graphs shown here do not follow all dual graph rules (listed in Subsection 2.2 of the main paper). The first three graphs have all degree-3 vertices (exactly two degree-3 vertices are allowed), and the last three graphs have one degree-5 vertex each (the maximum degree allowed for a vertex is 4). These graphs are discarded by our graph enumeration algorithm. See Figure 4 of the main paper for the remaining 25 graphs that do follow the RNA dual graph rules.

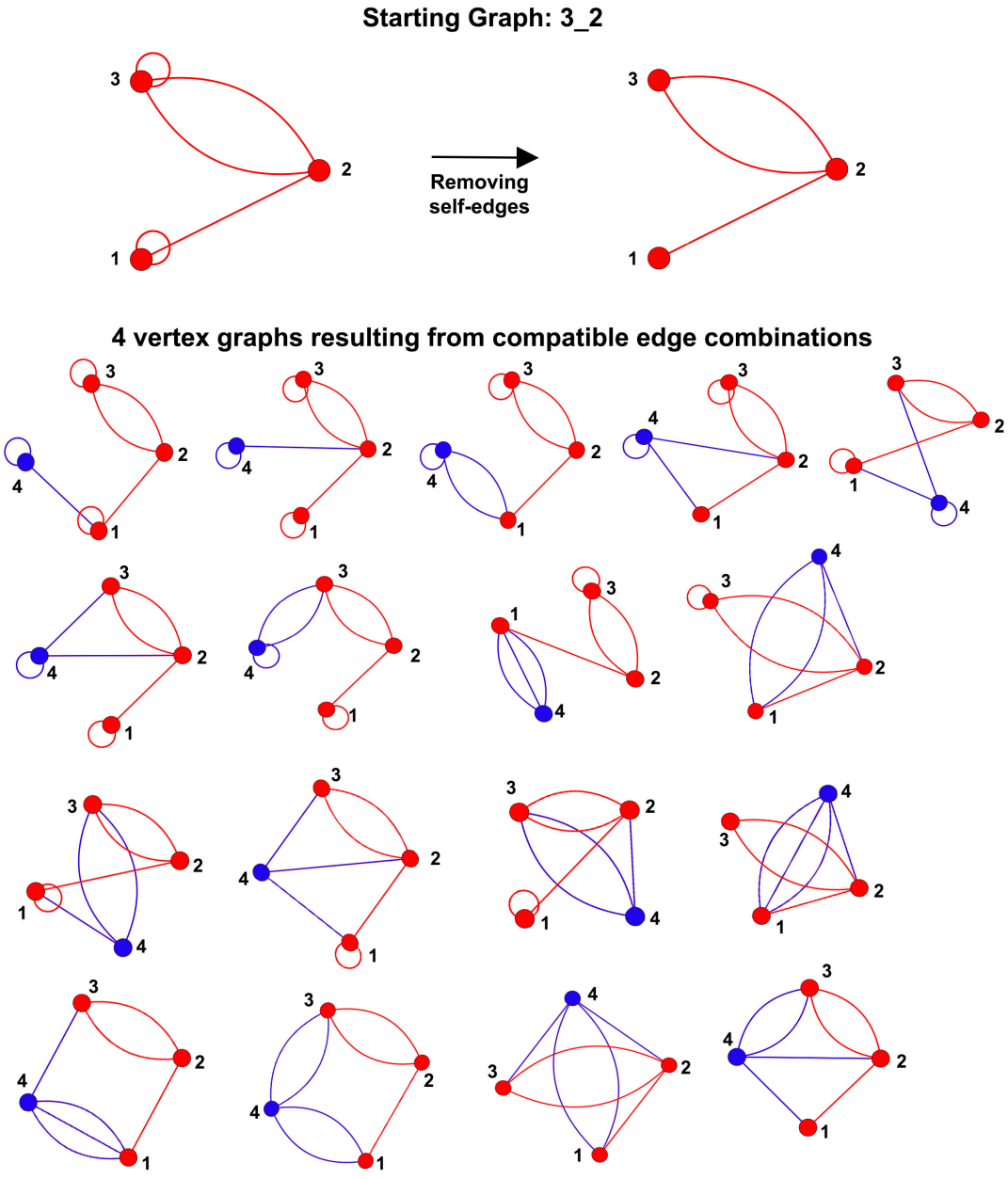

Figure S2: Compatible dual graphs generated by our enumeration algorithm by connecting the new vertex (blue segment) to the starting graph 3_2 (red). Of the 31 graphs generated using the edge combinations listed in Figure 3 of the main paper, 17 graphs shown here follow all dual graph rules.

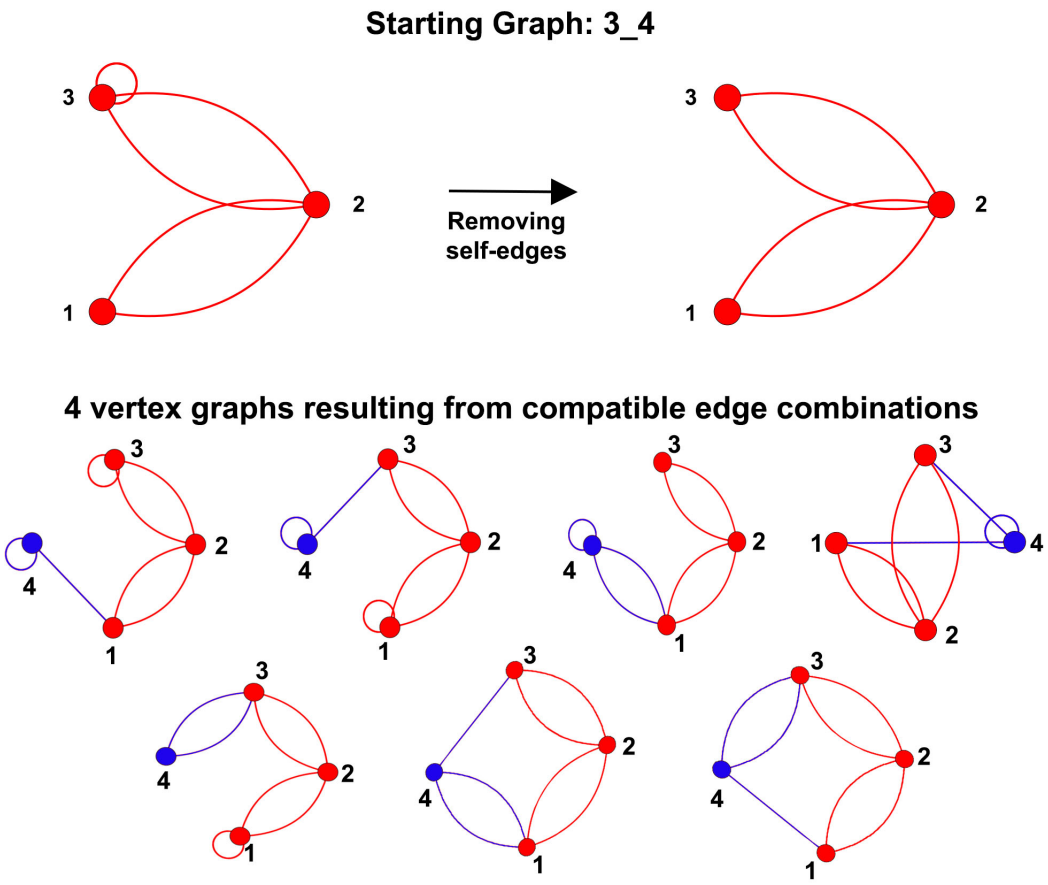

Figure S3: Compatible dual graphs generated by our enumeration algorithm by connecting the new vertex (blue segment) to the starting graph 3_4 (red). Of the 31 graphs generated using the edge combinations listed in Figure 3 of the main paper, 7 graphs shown here follow all dual graph rules.

Table S1: Number of pairs, triples, quadruples, and quintuples of non-isomorphic dual graph topologies (ignoring self-edges) between 5 and 9 vertices with identical eigenvalue spectra in the new dual graph library. Total number of graphs indicates the number of graphs in the new library that have at least one non-isomorphic graph with the same eigenvalue spectrum in the library.

Number of Vertices	Number of Pairs	Number of Triples	Number of Quadruples	Number of Quintuples	Total Number of Graphs
5	2	0	0	0	4
6	12	0	0	0	24
7	104	2	0	0	214
8	425	22	4	1	937
9	2622	149	12	2	5749

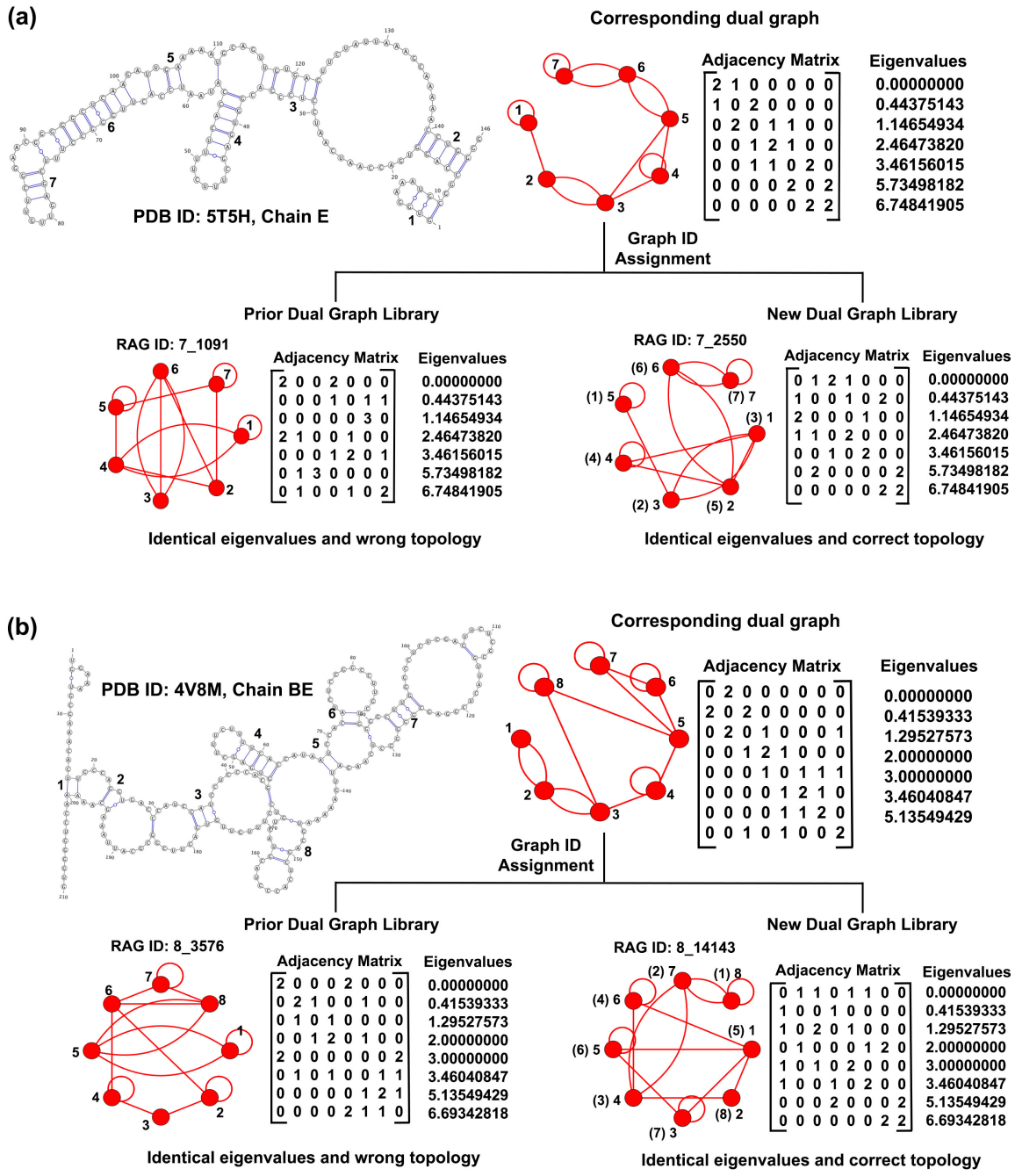

Figure S4: Differences between the graph IDs assigned to two RNA 2D structures using the prior and the expanded dual graph library. (a) short ribosomal RNA 1 of the 60S ribosomal subunit for *Trypanosome cruzi* (PDB ID: 5T5H chain E) was assigned graph ID 7_1091 using the prior library and 7_2550 with the expanded library. The numbers in parenthesis show the vertex number of the RNA graph on top mapped onto the 7_2550 graph from the library on right as they are isomorphic, and hence 7_2550 is the correct ID. (b) short ribosomal RNA 1 of the large ribosomal subunit of *Trypanosoma brucei* (PDB ID: 4V8M, chain BE) was assigned graph ID 8_3576 using the prior library and 8_14143 with the expanded library. The numbers in parenthesis show the vertex number of the RNA graph on top mapped onto the 8_14143 graph from the library on right as they are isomorphic, and hence 8_14143 is the correct ID.

Table S2: Number of dual graphs and subgraphs between 2-9 vertices for RNA 2D structure files (see Subsection 2.4 of the main paper) that were either unassigned or misclassified previously.

Number of Vertices	Full Graphs			Subgraphs		
	Total Graphs	Unassigned Graphs	Misclassified Graphs	Total Subgraphs	Unassigned Subgraphs	Misclassified Subgraphs
2	433	0	0	52506	0	0
3	309	0	0	34033	0	0
4	828	0	0	33021	0	0
5	189	0	0	35546	0	0
6	476	0	0	34951	0	0
7	201	0	1	36271	1737	910
8	28	1	1	39124	4120	697
9	31	18	0	45842	22022	2651

References

[1] S. Jain, C. S. Bayrak, L. Petingi, T. Schlick, Dual graph partitioning highlights a small group of pseudoknot-containing RNA submotifs, Genes 9 (8) (2018) 371. doi:10.3390/genes9080371.