# Rethinking Performance Measures of RNA Secondary Structure Problems

**Frederic Runge**
Department of Computer Science
University of Freiburg
runget@cs.uni-freiburg.de

**Jörg K.H. Franke**
Department of Computer Science
University of Freiburg
frankej@cs.uni-freiburg.de

**Daniel Fertmann**
Department of Computer Science
University of Freiburg
fertmand@cs.uni-freiburg.de

**Frank Hutter**
Department of Computer Science
University of Freiburg
fh@cs.uni-freiburg.de

## Abstract

Accurate RNA secondary structure prediction is vital for understanding cellular regulation and disease mechanisms. Deep learning (DL) methods have surpassed traditional algorithms by predicting complex features like pseudoknots and multi-interacting base pairs. However, traditional distance measures can hardly deal with such tertiary interactions and the currently used evaluation measures (F1 score, MCC) have limitations. We propose the Weisfeiler-Lehman graph kernel (WL) as an alternative metric. Embracing graph-based metrics like WL enables fair and accurate evaluation of RNA structure prediction algorithms. Further, WL provides informative guidance, as demonstrated in an RNA design experiment.

## 1 Introduction

Ribonucleic acid (RNA) is one of the major regulators in cells and has been connected to multiple diseases like cancer [1] and Parkinson's [2]. Since the function of RNAs is dominated by their structure [3], accurate prediction of these structures appears as a fundamental problem in computational biology [4]. RNA folds hierarchically and the formation of the final 3-dimensional shape strongly depends on the formation of a *secondary structure*, which describes nucleotide pairings (base pairs) of the RNA sequence via hydrogen bonds [5]. The secondary structure already defines the sites for interactions with other cellular compounds [3], and improvements in secondary structure prediction, therefore, could have a substantial impact on RNA-related research.

The potential improvements in the field of RNA structure prediction by sophisticated learning algorithms recently attracted the interest of the deep learning (DL) community and led to an explosion of DL-based approaches in the field [6–16], and new state-of-the-art results. The typical output of these algorithms is a squared $L \times L$ binary adjacency matrix, where $L$ is the length of the input nucleotide sequence, indicating positions where base pairs form [6, 16]. Therefore, they can predict all possible base pairs, including pseudoknots [17] and multi-interacting bases (multiplets) which is an advantage over more traditional, dynamic programming-based algorithms.

RNA secondary structure predictions are typically evaluated for the accuracy of predicted base pairs compared to a known structure. Traditional algorithms typically use tree representations, either defining structure distance via edit operations [18–20] or tree alignment approaches [21]. However, while there exist methods that are capable of comparing pseudoknotted structures [22], to the best of our knowledge, there exists currently no algorithm that considers base multiplets, an important type of nucleotide interactions e.g. for the formation of G-quadruplex structures [23]. Therefore,

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

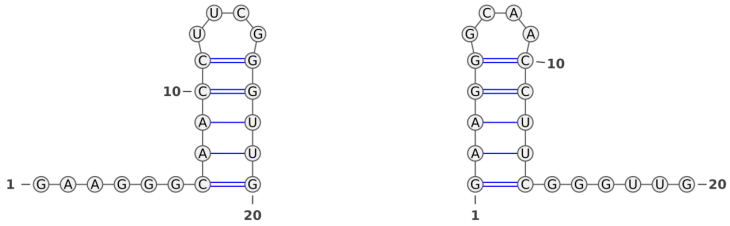


Figure 1: **Bi-stable RNA 20mer.** The F1 score and MCC when comparing both folds is $0.0$ and $-0.026$, respectively. The Weisfeiler-Lehman graph kernel provides a score of $0.25$.

current state-of-the-art DL approaches evaluate their predictions using performance measures derived from the confusion matrix. Two performance measures are well established, the F1 score and the Matthews Correlation Coefficient (MCC) [24]. The main difference is that the F1 score is independent of the true negatives and that it is not symmetric with respect to class swapping which makes the MCC generally preferable with respect to binary classification [25]. However, for scoring secondary structure prediction, both have their flaws. As an example, Figure 1 shows a bi-stable RNA 20mer taken from Wenter et al. [26] in its two conformations. The respective F1 score between both conformations is zero, while the MCC score of $-0.026$ is even below the score expected for a random structure. Obviously, both structures share a common feature (the hairpin structure) with the exact same base pair pattern which should be reflected by the score. Mathews [27] recently proposed a derivative of the F1 score, named F1-shift for the remainder of this paper, to account for RNA structural dynamics like bulge migration [28]. While this measure accounts for certain shifts in the base pairing scheme of the secondary structure, it still cannot capture the similarity between secondary structures as shown in Figure 1, similarly resulting in a score of zero.

RNAs can be represented as graphs and, therefore, graph metrics could be considered for scoring the distance between two RNA secondary structures. To our knowledge, graph distance metrics have not yet been used for scoring RNA secondary structure prediction. Hence, we propose to use the Weisfeiler-Lehman graph kernel [29] for more accurate evaluations in RNA secondary structure prediction tasks, instead of the commonly used F1 score, MCC, or their derivatives for binary classification. In particular, our main contributions are as follows:

- We show that the commonly used F1 score and MCC as metrics for evaluating the quality of the secondary structure are misleading and error-prone.
- We propose the Weisfeiler-Lehman graph kernel (WL) as an alternative metric to closer align the biological motivation with the metric scale.
- We provide real-world examples to showcase the benefit of WL evaluation in contrast to F1 score, shifted F1 score, and MCC.
- In two application settings, we show the practical usefulness of WL for analysis and to guide an RNA design algorithm with improved results compared to traditional distance measures.

## 2 RNA Secondary Structure Measures

We review the commonly used measures for RNA secondary structure prediction, F1 score, MCC, and F1-shift before we briefly introduce the Weisfeiler-Lehman (WL) kernel, a powerful graph similarity measure widely used for comparing graphs in various domains. The WL kernel leverages the concept of graph isomorphism and employs a label propagation approach to capture the structural information of graphs.

### 2.1 Confusion Matrix Based Measures

The commonly used performance measures for RNA secondary structure prediction are based on a confusion matrix, which describes the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) of a given prediction.

**F1 Score** The F1 score describes the harmonic mean of precision and recall and can be described as $F1 = 2 \cdot TP/(2 \cdot TP + FP + FN)$.

**Matthews Correlation Coefficient** While the F1 score emphasizes on positives, the MCC is a more balanced measure. The MCC can be calculated as follows.

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{1}$$

**F1-shift** The F1-shift is a measure to account for structural dynamics in RNAs [27]. The F1-shift is computed as the F1 score, but for a given pair $(i, j)$ all pairs $(i, j + 1)$, $(i + 1, j)$, $(i, j - 1)$, and $(i - 1, j)$ are also considered correct.

### 2.2 Graph Isomorphism

Graph isomorphism refers to the notion of two graphs being structurally identical. Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, an isomorphism between them is a bijective mapping $f : V_1 \to V_2$ that preserves the adjacency relationship between vertices. Formally, for any two vertices $u, v \in V_1$, $(u, v) \in E_1$ if and only if $(f(u), f(v)) \in E_2$. Determining graph isomorphism is a computationally challenging problem with significant implications in various domains.

### 2.3 The Weisfeiler-Lehman Kernel

The Weisfeiler-Lehman kernel is a graph kernel that captures the structural information of graphs by iteratively refining node labels based on their local neighborhoods. It operates in two main steps: label propagation and hash function computation. The kernel assigns each node in the graph a label representing the node's local structural information and then computes a hash function that aggregates these labels to generate a feature vector.

#### 2.3.1 Label Propagation

The label propagation step involves iterating over the nodes of the graph and updating their labels based on the labels of their neighboring nodes. Initially, each node is assigned a unique label. In each iteration, the kernel collects the labels of a node's neighbors, sorts them lexicographically, and appends the node's own label to the list. This combined list of labels serves as the input for a hash function in the next step. Let $L_i(u)$ denote the label of node $u$ at iteration $i$. The label propagation process can be defined as follows:

$$L_{i+1}(u) = \text{hash}\left(L_i(u), \text{sort}\left(\left(L_i(v_1), \ldots, L_i(v_{|N(u)|})\right)\right)\right), v_j \in N(u), j \in \{1, \ldots, |N(u)|\}, \tag{2}$$

where $N(u)$ represents the set of neighboring nodes of $u$, sort$(\cdot)$ sorts the labels lexicographically, and hash$(\cdot)$ computes a hash value.

#### 2.3.2 Hash Function Computation

The hash function computes a hash value based on the lexicographically sorted list of labels obtained from the label propagation step. This hash value captures the local structural information of a node's neighborhood and is used to refine the node's label in the subsequent iterations. The hash function is typically implemented using a simple and efficient algorithm, such as the Rabin-Karp hash function [30]. The WL kernel computes a similarity score between two graphs $G_1$ and $G_2$ by comparing their corresponding label distributions obtained from the label propagation process. The similarity score can be computed as the dot product of the feature vectors generated by the hash functions:

$$\text{WL-Similarity}(G_1, G_2) = \Phi(G_1) \cdot \Phi(G_2), \tag{3}$$

where $\Phi(G)$ represents the feature vector of graph $G$ obtained by aggregating the labels through the hash functions.

The WL kernel is a powerful metric for comparing graphs, as it captures the structural information while being computationally efficient. It has been widely adopted in graph classification, pattern

Table 1: Evaluation of RNA secondary structure prediction algorithms. Note that we are not interested in finding the state-of-the-art algorithm but in a comparison of the different performance measures.

| Model | F1 | F1-shift | MCC | WL | F1 Rank | F1-shift Rank | MCC Rank | WL Rank |
|---|---|---|---|---|---|---|---|---|
| RNAformer | 0.712 | 0.721 | 0.731 | 0.771 | 1 | 1 | 1 | 1 |
| SPOT-RNA | 0.672 | 0.691 | 0.687 | 0.707 | 2 | 3 | 2 | 3 |
| SPOT-RNA2 | 0.668 | 0.701 | 0.674 | 0.705 | 3 | 2 | 4 | 4 |
| RNA-FM | 0.665 | 0.691 | 0.683 | 0.709 | 4 | 3 | 3 | 2 |
| RNAFold | 0.636 | 0.659 | 0.641 | 0.695 | 5 | 4 | 5 | 5 |
| LinearFold-V | 0.633 | 0.659 | 0.638 | 0.694 | 6 | 4 | 6 | 6 |
| ContraFold | 0.625 | 0.659 | 0.636 | 0.692 | 7 | 4 | 7 | 7 |
| pKiss | 0.613 | 0.64 | 0.62 | 0.676 | 8 | 5 | 10 | 9 |
| ipKnot | 0.611 | 0.617 | 0.624 | 0.675 | 9 | 8 | 9 | 10 |
| LinearFold-C | 0.61 | 0.63 | 0.628 | 0.686 | 10 | 7 | 8 | 8 |
| RNAstructure | 0.606 | 0.633 | 0.611 | 0.667 | 11 | 6 | 11 | 11 |
| REDfold | 0.487 | 0.502 | 0.501 | 0.619 | 12 | 9 | 12 | 12 |

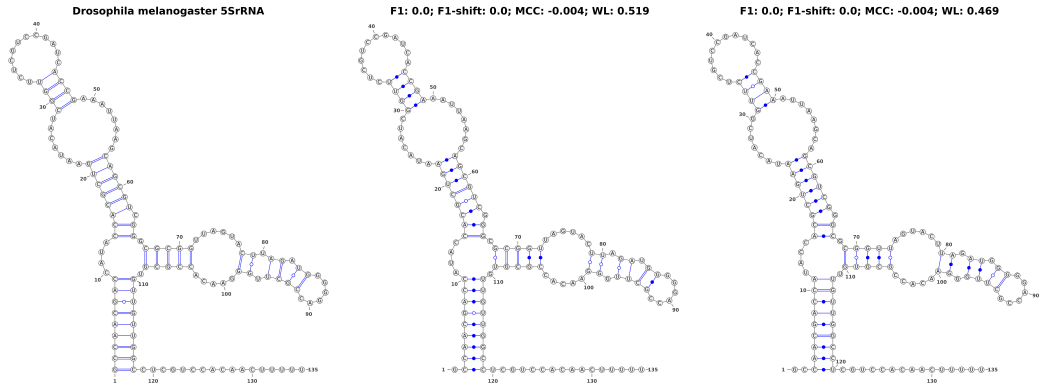


Figure 2: **Example of structural shift.** (Left) We show a 5SrRNA of Drosophila melanogaster (Middle) The same structure shifted by one position. (Right) The same structure shifted by two positions.

recognition, and graph mining tasks, showcasing its effectiveness across various domains. In the following, we use five iterations of WL for all numbers reported.

## 3 The Weisfeiler-Lehman Kernel for RNA Secondary Structure Measure

Before discussing potential misleading properties of the currently used performance measures for RNA secondary structure prediction, F1 score, F1-shift, and MCC, and displaying the benefits of WL, we start our discussion with an overview of the results, given by the different measures.

**Evaluation of RNA Folding Algorithms** We evaluate several RNA secondary structure prediction algorithms[31–36, 6, 9, 14–16] on the most recently proposed benchmark dataset for RNA secondary structure prediction, TS-hard [9]. The TS-hard test set is derived from 3D structures of the Protein Data Bank (PDB) [37] and contains a total of 28 samples; 7 samples without pseudoknots and nucleotides that pair with more than one other nucleotide (multiplets), 1 sample without pseudoknots but with multiplets, and 20 samples with both, pseudoknots and multiplets. We rank all algorithms with respect to the different performance measures. Table 1 shows the results of our evaluation. We observe that the choice of the performance measure can change the order of algorithms drastically.

**Properties of Performance Measures** We continue with analyzing, what we believe to be, a misleading property of the currently used performance measures F1 score, F1-shift, and MCC while showing that the WL approach can provide more informative scores for the specific setting. As an example, we use the secondary structure of a 5SrRNA of *Drosophila melanogaster* as provided by RNAcentral [38] (RNAcentral ID: URS00003B4856_7227). We then introduce a shift in the

structure by one and two positions. As shown in Figure 2, the structures all look similar, except for the positional shift. However, due to the binary nature of the scores, the F1 score as well as the F1-shift drop to zero for the shifted structures, while the MCC even shows a negative score. Further, the scores remain unchanged between the one positional and two positional shifts. In contrast, the WL captures the shift very well, resulting in a score of $0.519$ for the single positional shift and $0.469$ for the shift by two positions. Similarly, we simulate a bulge migration process, shown in Figure 4 in Appendix A.1. The F1-shift measure was introduced to capture such events and shows a score of $1.0$ for all structures. However, there exist structures that do not show the migrating phenomenon. For example the HIV-1 TAR RNA shows a specific stabilized tri-nucleotide bulge that is essential for Tat protein binding and obligatory in virus replication [39–41]. Thus, we think that a score should still be able to quantify a difference between the structures. We observe that the score of WL gradually decreases with the bulge moving further away from the original point similar to F1 score and MCC, however, the decrease is smaller, and less strict than for the other measures. We believe that these are very valuable and desirable properties of WL, which we can also leverage during training: While all of the measures, including WL, are not differentiable, there exist surrogate models capable of accurately modeling graph distances based on graph neural networks (GNNs) [42]. Such a surrogate model could be used during training to inform a learning algorithm for secondary structure prediction while being fully differentiable.

**Evolutionary Distance** A strong advantage of WL is the inclusion of sequence information into structure evaluation. We demonstrate this by simulating mutation events on the sequence level. The results are shown in Figure 5 in Appendix A.2. While all other measures cannot capture the mutation information, the WL decreases with the amount of sequence changes. This could be a useful feature of WL e.g. when applying it in evolutionary studies to determine distances on the sequence and structure level.

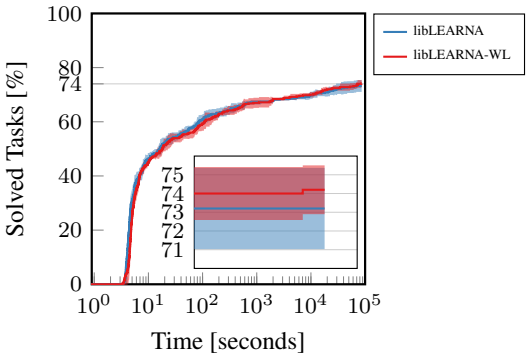


Figure 3: RNA Design guided by Hamming distance (libLEARNA) or WL (libLEARNA-WL).

**Application to RNA Design** We see another application of WL in the field of RNA design. Here, we use WL to guide the design of the most recently proposed learning-based algorithm, libLEARNA [43], an improved version of the automated reinforcement learning approach, LEARNA [44]. Specifically, we use a model that was trained using the Hamming distance to measure the distance between the target structure and the folding of the predicted candidate sequence and exchange this distance measure with WL during evaluation. We use version two [45] of the commonly used Eterna100 benchmark [46], with the proposed evaluation scheme of five independent runs for 24 hours. Figure 3 shows that without training on it, WL seems to improve the guidance of libLEARNA during evaluation, resulting in improved performance.

**Limitations** A common phenomenon in RNA is base-stacking [47], which cannot be captured by any of the performance measures compared here. Further, while the WL allows to include sequence information which makes it aware of changes in the base pairing pattern, vanilla WL is unaware of the value of exchanging specific base pairs (e.g. Watson-Crick and non-canonical base pairs). For example, it might be desirable to penalize the introduction of certain base interactions (e.g. pseudoknots) when these are not formed in the ground-truth structure. However, both, the base-stacking and base pair penalties, could generally be introduced via weights of edges to inform the WL about such changes, but such an adapted version of WL demands in-depth analysis in the future.

## 4 Conclusion

We show that currently used performance measures for the evaluation of RNA secondary structure prediction have misleading properties due to their focus on the confusion matrix. We propose to

use graph distance metrics for the evaluation of the secondary structure prediction and suggest the Weisfeiler-Lehman graph kernel (WL) as a competent measure of graph similarity. Subsequently, we compare the Weisfeiler-Lehman graph kernel to current measures in different settings, indicating its benefits and limitations. Finally, we suggest that GNN-based surrogate models can be used to train DL algorithms more informed for RNA secondary structure prediction and show that WL can improve the performance when applied to RNA design.

## Acknowledgments and Disclosure of Funding

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG.

## References

[1] John R. Prensner, Matthew K. Iyer, O. Alejandro Balbin, Saravana M. Dhanasekaran, Qi Cao, J. Chad Brenner, Bharathi Laxman, Irfan A. Asangani, Catherine S. Grasso, Hal D. Kominsky, Xuhong Cao, Xiaojun Jing, Xiaoju Wang, Javed Siddiqui, John T. Wei, Daniel Robinson, Hari K. Iyer, Nallasivam Palanisamy, Christopher A. Maher, and Arul M. Chinnaiyan. Transcriptome sequencing across a prostate cancer cohort identifies pcat-1, an unannotated lincrna implicated in disease progression. *Nature Biotechnology*, 29(8):742–749, 2011.

[2] Bingqing Cao, Tao Wang, Qiumin Qu, Tao Kang, and Qian Yang. Long noncoding rna snhg1 promotes neuroinflammation in parkinson's disease via regulating mir-7/nlrp3 pathway. *Neuroscience*, 388:118 – 127, 2018. ISSN 0306-4522.

[3] Minakshi Gandhi, Maiwen Caudron-Herger, and Sven Diederichs. Rna motifs and combinatorial prediction of interactions, stability and localization of noncoding rnas. *Nature Structural & Molecular Biology*, 25:1070–1076, 2018.

[4] Édouard Bonnet, Paweł Rzażewski, and Florian Sikora. Designing rna secondary structures is hard. *Journal of Computational Biology*, 27(3):302–316, 2020.

[5] Ignacio Tinoco Jr and Carlos Bustamante. How rna folds. *Journal of molecular biology*, 293(2): 271–281, 1999.

[6] Jaswinder Singh, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications*, 10(1):1–13, 2019.

[7] Hao Zhang, Chunhe Zhang, Zhi Li, Cong Li, Xu Wei, Borui Zhang, and Yuanning Liu. A new method of rna secondary structure prediction based on convolutional neural network and dynamic programming. *Frontiers in genetics*, 10:467, 2019.

[8] FA Rezaur Rahman Chowdhury, He Zhang, and Liang Huang. Learning to fold rnas in linear time. *bioRxiv*, page 852871, 2019.

[9] Jaswinder Singh, Kuldip Paliwal, Tongchuan Zhang, Jaspreet Singh, Thomas Litfin, and Yaoqi Zhou. Improved rna secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics*, 37, 2021.

[10] Mehdi Saman Booy, Alexander Ilin, and Pekka Orponen. Rna secondary structure prediction with convolutional neural networks. *BMC bioinformatics*, 23(1):58, 2022.

[11] Hannah K Wayment-Steele, Wipapat Kladwang, Alexandra I Strom, Jeehyung Lee, Adrien Treuille, Alex Becka, Eterna Participants, and Rhiju Das. Rna secondary structure packages evaluated and improved by high-throughput experiments. *Nature Methods*, 19(10):1234–1242, 2022.

[12] Andrew J Jung, Leo J Lee, Alice J Gao, and Brendan J Frey. Rtfold: Rna secondary structure prediction using deep learning with domain inductive bias.

[13] Jörg Franke, Frederic Runge, and Frank Hutter. Probabilistic transformer: Modelling ambiguities and distributions for rna folding and molecule design. *Advances in Neural Information Processing Systems*, 35:26856–26873, 2022.

[14] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Irwin King, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

[15] Chun-Chi Chen and Yi-Ming Chan. Redfold: accurate rna secondary structure prediction using residual encoder-decoder network. *BMC bioinformatics*, 24(1):1–13, 2023.

[16] Jörg KH Franke, Frederic Runge, and Frank Hutter. Scalable deep learning for rna secondary structure prediction. *arXiv preprint arXiv:2307.10073*, 2023.

[17] David W Staple and Samuel E Butcher. Pseudoknots: Rna structures with diverse functions. *PLoS biology*, 3(6):e213, 2005.

[18] Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262, 1989.

[19] Philip N Klein. Computing the edit-distance between unrooted ordered trees. In *European Symposium on Algorithms*, pages 91–102. Springer, 1998.

[20] Erik D Demaine, Shay Mozes, Benjamin Rossman, and Oren Weimann. An optimal decomposition algorithm for tree edit distance. *ACM Transactions on Algorithms (TALG)*, 6(1):1–19, 2009.

[21] 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.

[22] Michela Quadrini, Luca Tesei, and Emanuela Merelli. An algebraic language for rna pseudoknots comparison. *BMC bioinformatics*, 20(4):1–18, 2019.

[23] Martin Gellert, Marie N Lipsett, and David R Davies. Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences*, 48(12):2013–2018, 1962.

[24] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[25] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[26] Philipp Wenter, Boris Fürtig, Alexandre Hainard, Harald Schwalbe, and Stefan Pitsch. A caged uridine for the selective preparation of an rna fold and determination of its refolding kinetics by real-time nmr. *ChemBioChem*, 7(3):417–420, 2006.

[27] David H Mathews. How to benchmark rna secondary structure prediction accuracy. *Methods*, 162:60–67, 2019.

[28] Sarah A Woodson and Donald M Crothers. Proton nuclear magnetic resonance studies on bulge-containing dna oligonucleotides from a mutational hot-spot sequence. *Biochemistry*, 26 (3):904–912, 1987.

[29] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

[30] Richard M Karp and Michael O Rabin. Efficient randomized pattern-matching algorithms. *IBM journal of research and development*, 31(2):249–260, 1987.

[31] Ivo Hofacker, Walter Fontana, Peter Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte fuer Chemie/Chemical Monthly*, 125:167–188, 02 1994.

[32] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[33] Jessica S Reuter and David H Mathews. Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):1–9, 2010.

[34] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. Ipknot: fast and accurate prediction of rna secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, 2011.

[35] Stefan Janssen and Robert Giegerich. The rna shapes studio. *Bioinformatics*, 31(3):423–425, 2015.

[36] Liang Huang, He Zhang, Dezhong Deng, Kai Zhao, Kaibo Liu, David A Hendrix, and David H Mathews. Linearfold: linear-time approximate rna folding by 5'-to-3'dynamic programming and beam search. *Bioinformatics*, 35(14):i295–i304, 2019.

[37] Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528, 2019.

[38] RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, 49(D1):D212–D220, 10 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa921.

[39] S Roy, U Delling, C-HRCA Chen, CA Rosen, and N Sonenberg. A bulge structure in hiv-1 tar rna is required for tat binding and tat-mediated trans-activation. *Genes & development*, 4(8): 1365–1373, 1990.

[40] Tadeusz Kulinski, Mikolaj Olejniczak, Hendrik Huthoff, Lukasz Bielecki, Katarzyna Pachulska-Wieczorek, Atze T Das, Ben Berkhout, and Ryszard W Adamiak. The apical loop of the hiv-1 tar rna hairpin is stabilized by a cross-loop base pair. *Journal of Biological Chemistry*, 278(40): 38892–38901, 2003.

[41] Lechuang Chen, Zhimin Feng, Hong Yue, Douglas Bazdar, Uri Mbonye, Chad Zender, Clifford V Harding, Leslie Bruggeman, Jonathan Karn, Scott F Sieg, et al. Exosomes derived from hiv-1-infected cells promote growth and progression of cancer via hiv tar rna. *Nature communications*, 9(1):4585, 2018.

[42] Clemens Damke, Vitalik Melnikov, and Eyke Hüllermeier. A novel higher-order weisfeiler-lehman graph convolution. In *Asian Conference on Machine Learning*, pages 49–64. PMLR, 2020.

[43] Frederic Runge, Jörg KH Franke, and Frank Hutter. Towards automated design of riboswitches. *arXiv preprint arXiv:2307.08801*, 2023.

[44] Frederic Runge, Danny Stoll, Stefan Falkner, and Frank Hutter. Learning to design RNA. In *International Conference on Learning Representations*, 2019.

[45] Rohan V Koodli, Boris Rudolfs, Hannah K Wayment-Steele, Eterna Structure Designers, and Rhiju Das. Redesigning the eterna100 for the vienna 2 folding engine. *bioRxiv*, pages 2021–08, 2021.

[46] Jeff Anderson-Lee, Eli Fisker, Vineet Kosaraju, Michelle Wu, Justin Kong, Jeehyung Lee, Minjae Lee, Mathew Zada, Adrien Treuille, and Rhiju Das. Principles for predicting RNA secondary structure design difficulty. *Journal of molecular biology*, 428(5):748–757, 2016.

[47] Zakir Ali, Ambika Goyal, Ayush Jhunjhunwala, Abhijit Mitra, John F Trant, and Purshotam Sharma. Structural and energetic features of base–base stacking contacts in rna. *Journal of Chemical Information and Modeling*, 63(2):655–669, 2023.

[48] Elizabeth A Dethoff, Jeetender Chugh, Anthony M Mustoe, and Hashim M Al-Hashimi. Functional complexity and regulation through rna dynamics. *Nature*, 482(7385):322–330, 2012.

[49] Laura R Ganser, Megan L Kelly, Daniel Herschlag, and Hashim M Al-Hashimi. The roles of structural dynamics in the cellular functions of rnas. *Nature reviews Molecular cell biology*, 20 (8):474–489, 2019.

[50] Manja Wachsmuth, Sven Findeiß, Nadine Weissheimer, Peter F. Stadler, and Mario Mörl. De novo design of a synthetic riboswitch that regulates transcription termination . *Nucleic Acids Research*, 41(4):2541–2551, 12 2012. ISSN 0305-1048.

# A Additional Experiments

In this section, we provide additional results for the comparison of the performance measures, F1 score, MCC, F1-shift, and the proposed Weisfeiler-Lehman graph kernel (WL). In Section A.1, we show results for a simulated bulge migration experiment. Section A.2 describes an experiment to analyze the influence of mutations on the scores of the performance measures.

## A.1 Analysis of Properties

RNA structures are dynamic [48, 49]. One example of such a dynamic behavior is the phenomenon of bulge migration [28], where unpaired nucleotides on one strand of a stem move across the stem. For our analysis, we simulate such a migration event using a synthetic theophylline riboswitch construct, RS3, proposed by Wachsmuth et al. [50]. The structure of RS3 contains a single nucleotide bulge at position 60 (see Figure 4, top left) which we successively move by one position downstream in the stem. The results are shown in Figure 4. Since F1-shift was introduced to capture exactly such shifts, the F1-shift score constantly stays at 1.0 for all structures. However, the structures differ and not all bulges show a migration behaviour. Optimally, a performance measure, thus, should be capable of reflecting this in the scores. While the other measures all show reduced scores with increasing distance of the bulge from the original position, we observe that the WL shows the best balance since the decrease in the score is smoother and less drastically compared to the F1-score and the MCC.

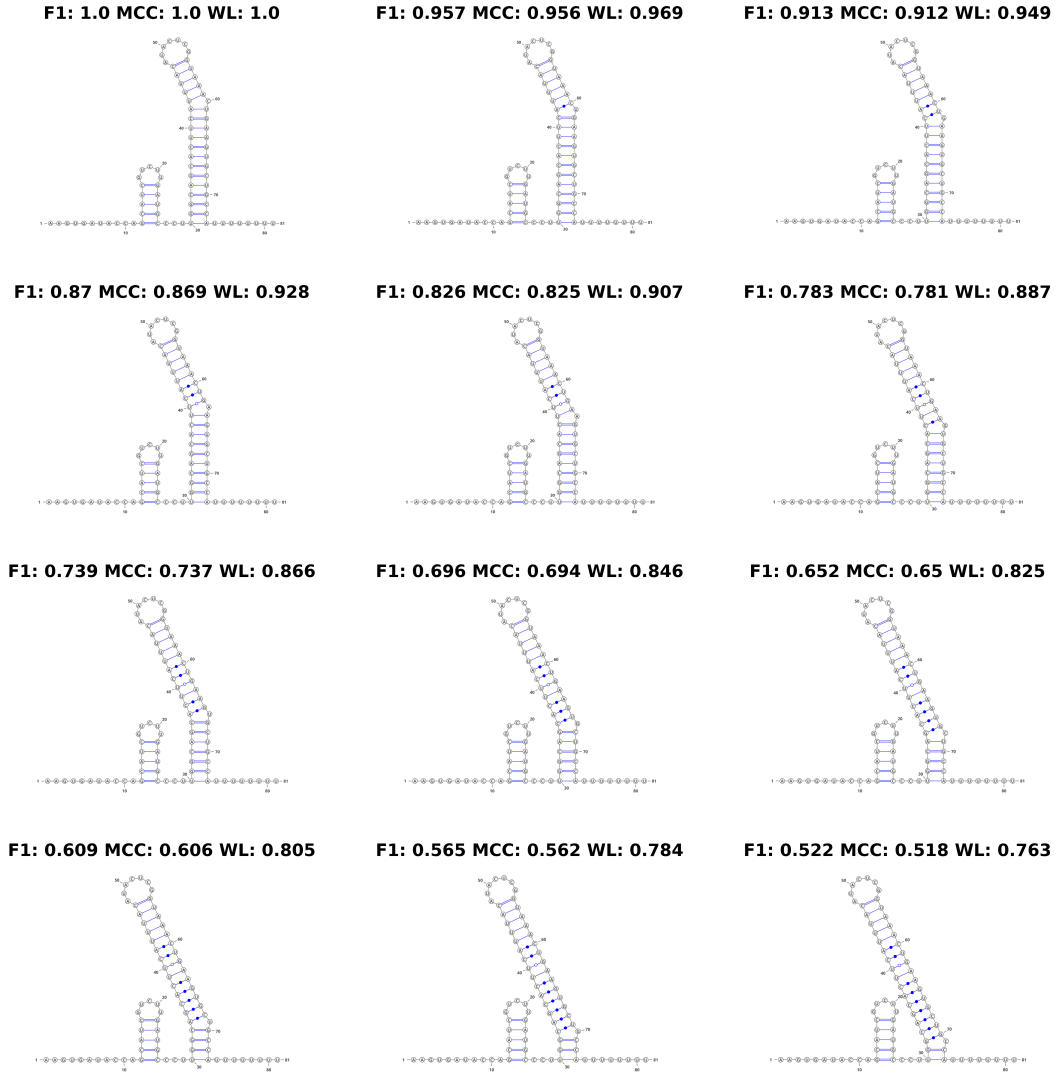


Figure 4: **Bulge migration example.** We show an example of a simulated bulge migration process on a synthetic theophylline riboswitch construct RS3 proposed by Wachsmuth et al. [50]. Top left shows the original construct. With each step, the bulge in the right stem is moving by one position.

### A.2 Application to Analysis

In this section, we show the benefits of the property of WL to consider changes in the sequence additional to the structure for scoring the distance. We again use the theophylline riboswitch construct, RS3, and introduce mutations of base pairs. Specifically, we either randomly change one, two, four, or eight base pairs, all base pairs of the first stem, all base pairs of the second stem, or replace the entire sequence with A's only. Figure 5 shows the structures and the respective scores for the performance measures. Except for WL, all performance measures are not capable of capturing changes in the sequence for their scores and the scores stay at $1.0$ for these measures since the structures remain unchanged. In contrast, the WL captures the mutation changes very well with decreasing scores depending on the number of mutations that we introduce. This property is a strong advantage of WL and could allow to use the WL kernel in evolutionary studies in the future. Further, besides capturing topological differences in RNA structures, the WL is generally capable of capturing changes in the base pair composition, which could allow for more fine-grained evaluation of structure prediction.

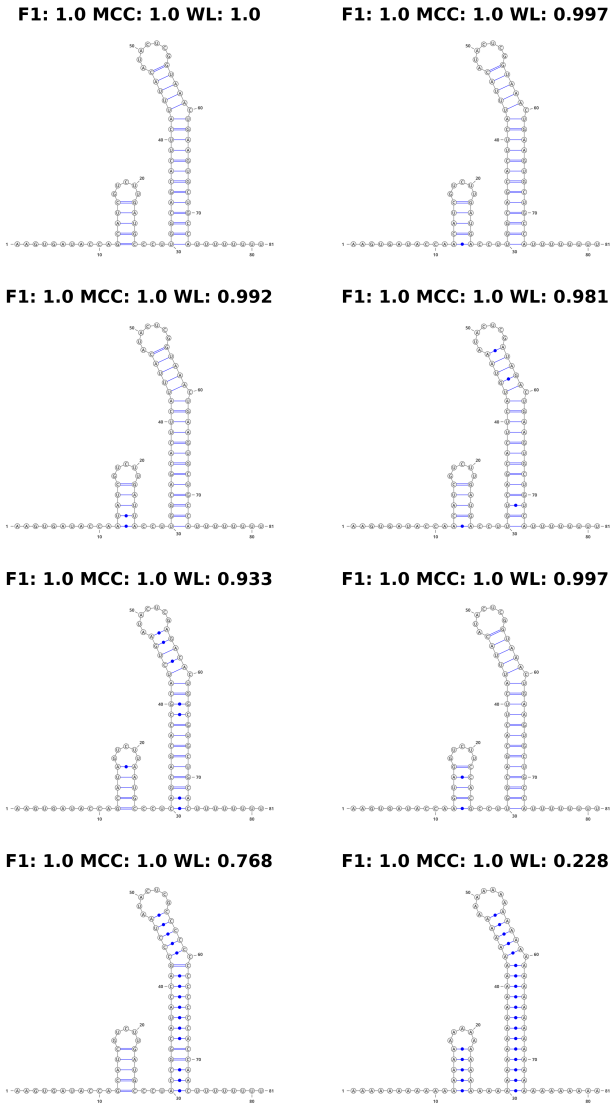


Figure 5: **Mutation Example.** We show an example of a simulated mutation process on a synthetic theophylline riboswitch construct RS3 proposed by Wachsmuth et al. [50]. Top left shows the original construct. With each step (left to right, top to bottom), we introduce the following mutations: 1 base pair (bp) mutated, 2 bp, 4 bp, 8 bp, entire first stem, entire second stem, entire sequence to 'A'.