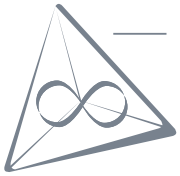
UNIVERSITÄT
LEIPZIG



Max Planck Institute for
Mathematics
in the Sciences

Faculty of Mathematics and Computer Science
Institute for Computer Science

From RNA folding to inverse folding: *an heuristic exploration*

PhD's Thesis

Submitted by

Nono Saha Cyrille Merleau

Aspired Academic Degree:
Doctor of Philosophy (Dr.rer.nat.) in Bioinformatics

Leipzig, Dezember 2021

Main supervisor:	Dr. Matteo Smerlak	MPI MiS
University supervisor:	Prof. Dr. Peter Stadler	University of Leipzig

Abstract

RNA molecules are ubiquitous in living organisms. Besides their role in translating genetic information into functional, often catalytically active proteins, various classes of RNA molecules are capable of catalysis themselves. Their capability to carry heritable information and perform certain functions makes them promising candidates for pre-cellular self-replicating systems exhibiting Darwinian evolution. One such molecule is the *Azoarcus* group I intron, which has been shown to catalyze self-assembly from smaller fragments. Finding an abundance of similar RNA molecules could strengthen the significance of RNA emerging prior to proteins in the origin of life, a view known as the RNA world hypothesis.

In this thesis, I employ computational methods based on thermodynamics in order to design RNA sequences that are structurally similar to the *Azoarcus* group I intron, motivated by the close relation of RNA structure and function.

In general, RNA structure is hierarchical, and efficient prediction of secondary structure from RNA sequence is possible, facilitating the reverse process of finding sequences given a secondary target structure. However, the structure of the *Azoarcus* group I intron contains a *pseudoknot* related to its function. This feature is usually considered to be part of the tertiary structure. Prediction of general *pseudoknots* has been shown to be NP-complete, and algorithms restricted to certain classes of *pseudoknots* are still more complex than pure secondary structure prediction methods.

The herein described design process accounts for this feature without explicitly modelling it, allowing to apply efficient secondary structure algorithms during RNA design while only requiring structure prediction methods with *pseudoknots* for verification post-design.

Acknowledgements

I am grateful to everyone who took part in my journey writing a master's thesis in the middle of a pandemic.

I would like to thank Peter Stadler for his straightforward and approachable supervision. I am particularly thankful for the guidance and opportunities to learn beyond the scope of my thesis that Matteo Smerlak provided.

Furthermore, I would like to express my appreciation for everyone in the workgroup *Structure of Evolution*, especially Nono Saha Cyrille Merleau and Vaitea Opuu for their valuable feedback and comments. I felt very welcome and valued. I would also like to extend my thanks to the staff at MPI MiS for competent organisational assistance throughout the pandemic. I wish to acknowledge Camille Lambert from *Laboratoire de Biochimie* at ESPCI Paris in France for working on *in vitro* experiments and generously providing preliminary results.

Finally, I cannot stress enough how thankful I am for the support from my family and friends — in particular from my parents for their trust, Franzi for her help in restructuring my first draft and Moritz for proofreading.

Contents

List of Figures **6**

List of Tables **6**

0.1 The RNA World Hypothesis . . . 7
0.2 Outline . . . 8

1 RNA Secondary Structures **10**
1.1 Secondary Structure Thermodynamics . . . 11
1.1.1 The Nearest-Neighbor Energy Model . . . 11
1.1.2 Secondary Structure Prediction . . . 13
1.2 Pseudoknots . . . 15

2 The *Azoarcus* Group I Intron **18**
2.1 Structural Features . . . 18
2.2 Self-Splicing . . . 19
2.3 Self-Assembly . . . 20
2.4 Native Sequence and Target Structure Data . . . 21

3 Methods **23**
3.1 Secondary Structure Prediction . . . 23
3.1.1 `pKiss` . . . 23
3.1.2 `RNAfold` . . . 24
3.1.3 `RNAPKplex` . . . 24
3.1.4 Energy Parameter Sets . . . 25
3.2 RNA Design . . . 25
3.2.1 Objective Functions . . . 27
3.2.2 Sequence Constraints . . . 29
3.2.3 Pipeline . . . 31
3.2.4 Quality Control and Selection of Designed Sequences . . . 34
3.3 Neutral Paths . . . 35

4 Results **37**
4.1 Numerical Results . . . 37
4.1.1 Structure Prediction of *Azoarcus* . . . 37
4.1.2 Sequence Designs . . . 42
4.1.3 Neutral Path Lengths . . . 48
4.2 Preliminary Experimental Results . . . 51

4.2.1 Mechanism of the Self-Splicing Assay 52
4.2.2 Assay Results 52

5 Discussion **54**
5.1 Design Challenges and Limitations 54
5.1.1 Revisiting Assumptions 56
5.2 Methodological Considerations 56
5.2.1 Outlook on Other Approaches 58
5.3 Conclusion 59

A Appendix **60**
A.1 Code and Data Availability 60
A.2 Superpositioned Nested Target Structures 60
A.3 MFE Predictions for the Reference Sequence 61
A.4 Sequence Design Examples 61
A.5 Initial Sequence Designs 62
A.6 Ensemble Defect Generalization 64

List of Figures

1 Structure Representation . . . 11
2 Loop Types . . . 12
3 Some Common Pseudoknot Types . . . 16
4 Secondary Structure of the *Azoarcus* Group I Intron . . . 19
5 Self-Splicing of the *Azoarcus* Group I Intron . . . 20
6 Self-Assembly of the *Azoarcus* Group I Intron . . . 21
7 Sequence and Structure Data . . . 22
8 Sequence Constraints . . . 30
9 Move Set in Space of Compatible Sequences . . . 31
10 Schematic of the Design Pipeline . . . 33
11 Constraining `RNAfold` Prediction . . . 38
12 Influence of the H-Type Pseudoknot Penalty on `pKiss` . . . 39
13 Free Energy of the P7 Helix . . . 40
14 Base Pair Probabilities of the `GISSD` Sequence . . . 40
15 *Azoarcus* Structure Prediction With Different Energy Parameter Sets . . . 41
16 Nucleotide Composition of Designs (Constrained Approach) . . . 43
17 Properties of Sequence Designs (Constrained Approach) . . . 44
18 Base Pair Probabilities of Selected Designs (Constrained Approach) . . . 45
19 Nucleotide Composition of Designs (Alternative Approach) . . . 47
20 Properties of Sequence Designs (Alternative Approach) . . . 48
21 Base Pair Probabilities of Selected Designs (Alternative Approach) . . . 49
22 Neutral Path Lengths . . . 50
23 Gel of a Splicing Assay . . . 53
A.1 Nucleotide Composition of Initial Constrained Designs . . . 62
A.2 Properties of Initial Constrained Designs . . . 63
A.3 Base Pair Probabilities of Initial Constrained Designs . . . 63

List of Tables

1 Energy Parameter Sets . . . 26
2 Constraint Sets . . . 30
3 Runtime and Accuracy of Different Prediction Methods . . . 37
4 Mean Neutral Path Lengths and Expected Random Hamming Distances . . . 50
A.1 Code Contribution to `ViennaRNA` . . . 60
A.2 Superpositioned Nested Target Structures . . . 60
A.3 MFE Predictions of the Reference Sequence . . . 61
A.4 Sequences Used in Dot Plots . . . 61

0.1 The RNA World Hypothesis

The fundamental question of the origin of life is, broadly speaking, concerned with the emergence of the first cellular organisms from a primordial soup of inanimate molecules. At a closer look, the search for an answer has to address multiple issues to bootstrap the first proto-cells.

In the most simplistic view of a cell hypothesized by **crick_protein_1958**, the genetic information stored in DNA is transcribed into RNA, which is translated into proteins, often carrying out catalytic activity [**crick_protein_1958**]. There is a dichotomy between nucleic acids carrying genetic information and proteins carrying out function, seemingly requiring both types of biopolymers interacting in the translation apparatus to facilitate Darwinian evolution of cells.

What did the first self-replicating molecules preceding cellular organisms look like? Such molecules would have been required to both store heritable information and carry out some catalytic function. **orgel_evolution_1968** argued that either nucleic acids or proteins could be such early self-replicating molecules [**orgel_evolution_1968**]. At this point, the existence of catalytically active nucleic acids was unknown, and **orgel_evolution_1968** was uncertain whether nucleic acid — DNA or RNA — chains with well-defined secondary structure would be capable of catalysis [**orgel_evolution_196**

The view of RNAs being the first or at least early self-replicating molecules is known as the *RNA world hypothesis*, a term phrased by **gilbert_origin_1986** after the first ribozymes catalyzing self-excision from surrounding RNA have been discovered in *Escherichia coli* and *Tetrahymena* [**gilbert_origin_1986**].

eigen_principle_1977 introduced a model of pre-cellular Darwinian systems, the *hypercycle*, building a complex translation apparatus from initial self-replicating units although they assumed nucleic acids to be limited in their information storage capacity so that multiple interacting replicators would be necessary [**eigen_principle_1977**]. Later, they proposed a concrete hypercycle starting from polynucleotides to bootstrap protein translation by tRNA-like self-replicating precursors and argued for RNA to precede DNA as the first genetic information-carrying molecule interacting with primitive catalytic proteins [**eigen_hypercycle_1978-1**, **eigen_origin_1981**].

Although RNA is often seen as fundamental to the origin of life, the hypothesized RNA world is not necessarily the beginning of it all. RNA-like worlds preceding RNA itself have been proposed as alternatives in which ribose would have been substituted by peptides, threose, or glycol [**robertson_origins_2012**].

Research efforts aiming to support the RNA world hypothesis or related concepts have taken various forms. Early attempts by Orgel, Zielinski, Kiedrowski and Sievers were based on DNA and DNA-analogue oligomers and yielded non-

enzymatic template-directed self-replicators [**kiedrowski_self-replicating_1986**, **zielinski_autocatalytic_1987**, **sievers_self-replication_1994**]. Approaches taken by scientists around Joyce have focused on finding replicase ribozymes, arguing that replicase functionality is within reach from existing polymerase ribozymes [**mcginness_search_2003**, **robertson_origins_2012**]. In fact, **breaker_emergence_1994** constructed a self-replicating RNA species which still required a DNA dependant RNA polymerase [**breaker_emergence_1994**].

However, template-based approaches were often limited as the product often diffuses slowly away from the template and inhibits further self-replication [**hayden_self-assembly_** The approach successfully pursued by **hayden_self-assembly_2006** built upon recombination of smaller fragments into a functional ribozyme, eliminating the need for a template and enabling both self-assembly and autocatalysis at the same time [**hayden_self-assembly_2006**].

Perhaps not surprisingly, a large part of research in the RNA world context built upon ribozymes related to the RNA molecules deemed promising by **gilbert_origin_1986** [**gilbert_origin_1986**]. These self-splicing ribozymes often found in tRNA precursors are called group I introns. The self-assembling autocatalyst Hayden used in his work belongs to this class of ribozymes and was used throughout this work as the design reference (see section 2).

0.2 Outline

This work aims to design sequences exhibiting secondary structures similar to a naturally occurring ribozyme *in silico* as a basis for artificial RNA catalysts. The motivation for this task is tied to the RNA world hypothesis as an abundance of such catalysts could strengthen the hypothesized role of RNA molecules in the origin of life. The ribozyme used as design reference is the *Azoarcus* group I intron (GII), which is introduced in section 2 with more details on its catalytic activity and distinctive structural features involved therein. In order to understand them, the previously introduced notion of secondary structures is concretized in section 1 and slightly extended to cover an important tertiary structural feature called a *pseudoknot* (section 1.2).

Central to the relationship between RNA sequences and secondary structure is the process of RNA *folding*, i.e. the formation of secondary structure from sequence and the prediction thereof. Historically, different approaches to structure prediction emerged, such as examining all possible structures or extracting phylogenetic information RNA sequences of a similar function [**zuker_rna_1984**]. The techniques used in this work all have an extensively developed thermodynamical foundation

in common whose principles and application to prediction have been outlined in section 1.1.

With RNA, assigning secondary structures to sequences can be interpreted as a mapping from genotype space to phenotype space. This genotype-phenotype map is not a one-to-one mapping; there are considerably more sequences than structures [**stadler_genotype-phenotype_2006**]. There are 4^N possible sequences of length N, composed from four different nucleotides, to be precise. In contrast, estimated numbers of structures range from 1.65^N to 2.35^N, depending on the exact definition of secondary structure [**stadler_genotype-phenotype_2006**, **haslinger_rna_1999**]. Consequently, given a target structure, there are potentially many sequences of that structure to find. The mapping process yielding a single such sequence is aptly named RNA *inverse folding* and builds, in practice, upon structure prediction.

Similarly, structure prediction is also the basis of the design pipeline implemented in this work, which is described in section 3 including the specific structure prediction methods used. The approaches taken for the design pipeline had to match the distinct characteristics of the target structure while simultaneously keeping computational complexity as manageable as possible, which is mirrored in the concrete design objectives and constraints specified in section 3.2.

The pipeline was assessed with a focus on sequence designs containing pseudoknots similar to the target structure and the diversity of the designed sequences (section 4). Finally, limitations and possible changes in methodology were discussed in section 5.

1 RNA Secondary Structures

In the literature, formal definitions of RNA secondary structure to be used in the context of combinatorial bioinformatics are usually quite similar. They may differ in details like allowed base pairs or the number of allowed pairings [**zuker_rna_1984**]. Here, a set of common rules is used to define legal pairings. Additionally, an RNA secondary structure is viewed here as a set of base pairs, although other mathematical objects could be used instead, for example vertex-labelled graphs [**hofacker_combinatorics_1998**].

An RNA sequence, i.e. the primary structure, can be viewed as as a word consisting of letters from the alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ representing the four RNA nucleotides. Therefore, given a RNA sequence r of length N, i.e. $r \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}^N$, a secondary structure of r is a set of ordered pairs (i, j) with $i < j$, where i and j denote paired positions satisfying the following rules [**hofacker_rna_2005**, **hofacker_rna_2006**]:

(i) for any two pairs (i, j), (k, l) with $i \leq k$: $i = k \Leftrightarrow j = l$

(ii) for any pair (i, j): $j - i > 3$

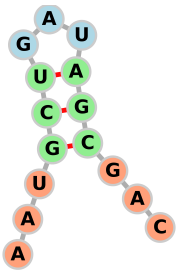
(iii) for any two pairs (i, j), (k, l) with $i \leq k$: $k < j \Rightarrow i < k < l < j$

(iv) for any pair (i, j): $r_i r_j \in \{\mathrm{AU}, \mathrm{UA}, \mathrm{CG}, \mathrm{GC}, \mathrm{GU}, \mathrm{UG}\}$

The advantage of formulating rules like this lies in the straightforward way to implement them in computational methods. Albeit seeming a bit convoluted, the intuition behind these rules is not too complex:

The first rule (i) specifies that each position participates in at most one base pair. Rule (ii) encompasses a simple steric constraint of RNA molecules; nucleobases too close to each other cannot form hydrogen bonds. Rule (iii) prohibits base pairs from crossing other base pairs in the same structure. Finally, (iv) provides some convenience, since restricting the allowed base pair identities to the most stable pairings reduces computational cost in stability-maximizing algorithms [**zuker_rna_1984**, **flamm_rna_1999**] aside from distinguishing from tertiary structures. A sequence and a structure are called compatible if they fulfill rule (iv). This notion is often not imposed in defitions of secondary structure [**hofacker_combinatorics_1998**], but it will be used later in section 3.2.3.

The terms *noncrossing* or *nested* structures are used synonymously for secondary structures characterized by these rules throughout this work. Figure 1a depicts a secondary structure as a vertex-labelled graph representing a coarse-grained view of the polymer.




```
...(((...)))...
AAUGCUGAUAGCGAC
```

(a) obtained using Forna [**kerpedjiev_forna_2015**]. **(b)** dot-bracket notation.

Figure 1: (a) A simple RNA secondary structure and **(b)** its dot-bracket notation aligned with the primary nucleotide sequence.

The so-called dot-bracket notation uses dots to denote unpaired positions and matching pairs of brackets or parentheses for base pair positions (Figure 1b).

1.1 Secondary Structure Thermodynamics

The underlying idea of using thermodynamics for RNA secondary structure prediction is that of phrasing a notion of structural stability in terms of the Gibbs free energy change ΔG relative to the unfolded state containing no base pairs at all [**tinoco_estimation_1971**]:

$$\Delta G = \Delta H - T\Delta S \leq 0 \tag{1}$$

where ΔH is the (pressure- and volume-dependent) enthalpy change, T the absolute temperature and ΔS the entropy change.

Furthermore, it is usually assumed that the free energy of a structure is the sum of its substructures' free energies [**dill_additivity_1997**]. With this assumption, the computation of RNA structure free energies may be reduced to smaller subproblems but requires a well-defined model of secondary structure decomposition and energy assignment such as the nearest-neighbor model introduced in section 1.1.1.

1.1.1 The Nearest-Neighbor Energy Model

Although defined by their base pairs, the hydrogen bonds between complementary nucleobases are not the major driving factor contributing to the stability of RNA secondary structures. In fact, the dominant stabilizing effect is attributed to consecutive (stacked) base pairs, whereas long unpaired regions enclosed between base pairs (loops) have destabilizing effects [**fresco_molecular_1960**, **hofacker_rna_2006**]. This is the foundation of the nearest-neighbor model, which is named after the loop-enclosing, *nearest neighboring* base pairs.

Following the additivity assumption, the stability of an RNA structure is computed by differentiating between stacks and different types of loops as seen in Figure 2 and summing over their individual stabilities.

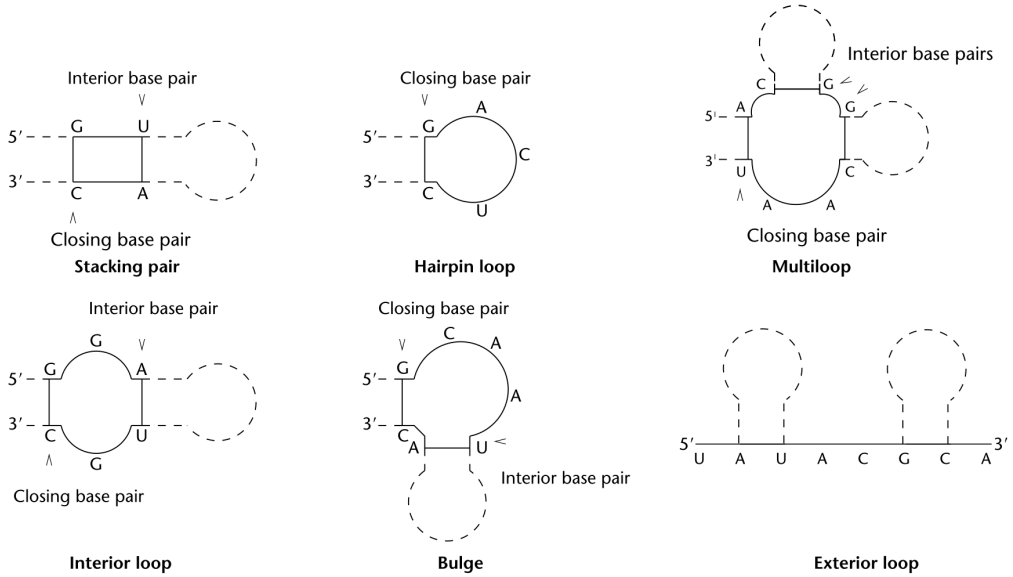


Figure 2: Types of loops considered in a loop decomposition of RNA secondary structures. Essentially, only stacking pairs have stabilizing effects. The destabilizing energy contribution of loops largely depends on their symmetry and enclosed unpaired nucleotides. Per convention, the exterior loop, or open chain, serves as a reference point relative to which free energy changes are defined. Modified after [**hofacker_rna_2005**].

As a simplified example, the destabilizing free energy contribution ΔG_{multi} of a multiloop as seen in Figure 2 may be modelled as

$$\Delta G_{\text{multi}} = \Delta G_{\text{init}} + b\Delta G_{\text{branch}} + u\Delta G_{\text{unpaired}} \tag{2}$$

where b is the number of all surrounding base pairs and u the number of base pairs [**dirks_partition_2003**].

Over time, energy values for stacked base pairs and the different loop types have been measured experimentally and used to estimate secondary structures [**tinoco_estimation_1**
tinoco_improved_1973]. Continuous improvements on such energy estimations for loops and unusual substructures eventually lead to comprehensive sets of energy parameters widely used in thermodynamical structure prediction [**turner_nndb_2010**].

Moreover, decomposition of a secondary structure into loops may be accomplished uniquely by identifying each loop with its closing base pair (see Figure 2) [**hofacker_combinatorics_1998**]. These closing base pairs are defined based on the 5'→3' direction of the sequence allowing to enumerate base pairs (and loops) in a precise order. Many structure prediction algorithms like the ones introduced in section 1.1.2 build on this loop decomposition.

1.1.2 Secondary Structure Prediction

The nearest-neighbor model enables realistically estimating the free energy of a given secondary structure or substructures, which is essential to thermodynamical structure prediction. Here, two common approaches to prediction used in this work are introduced.

Minimum Free Energy. With the notion of structure stability introduced earlier, predicting the most stable structure of an RNA sequence corresponds to finding the conformation with minimum (Gibbs) free energy (MFE) at thermodynamical equilibrium [**tinoco_estimation_1971**, **zuker_rna_1984**].

A simple approach to the problem would require exhaustively enumerating all possible structures and computing their free energies. However, the number of possible structures grows exponentially with sequence length [**schuster_sequences_1994**]. Nevertheless, efficient MFE prediction is feasible; the key to algorithms of polynomial complexity is the loop decomposition underlying the nearest-neighbor model and the additivity assumed for thermodynamical stability, both allowing to re-use substructures.

One of the most influential algorithms to achieve efficient MFE RNA secondary

structure prediction was developed by Zuker [**zuker_optimal_1981**]. The Zuker algorithm essentially implements a recursion for the loop decomposition of the nearest-neighbor model, weighting possible loops by their energy contribution.

This alone does not solve the problem of finding the structure with minimum free energy. The key to achieving a polynomial time complexity of $O(N^3)$ for sequences of length N lies in storing and using already computed MFEs of substructures shared by superstructures, a technique known as dynamic programming. The corresponding MFE structure is then obtained via backtracking on the stored intermediate computations.

Still, MFE structure prediction has some limitations; thermodynamical equilibrium is assumed, but that might not be the actual case. In long sequences, the additivity assumption for free energies breaks down with increasing occurrence of long-range base pairs restricting conformational options of enclosed subsequences [**dill_additivity_1997**]. MFE prediction of such long-range pairs tends to be inaccurate, possibly due to many alternative conformations with wide-spanning pairs [**amman_trouble_2013**]. More generally, the MFE structure of an RNA sequence is not necessarily unique. Since experimentally determined energy values are likely inaccurate to a certain degree, considering other possible, similar structures seems sensible.

Partition Function. Given the free energy change $\Delta G(s)$ of a structure s, the Boltzmann distribution describes the structure's probability at constant temperature T among all other possible structures of a sequence (equation 3).

$$p(s) = \frac{1}{Q} e^{-\Delta G(s)/RT} \tag{3}$$

Here, R is the ideal gas constant, and the precise probability depends on a factor $1/Q$. Q is the *partition function* and is defined on the Boltzmann ensemble Ω of all possible structures of a given sequence (equation 4).

$$Q = \sum_{s \in \Omega} e^{-\Delta G(s)/RT} \tag{4}$$

Assuming one is not only interested in the minimum free energy structure but also in other probable structures, the partition function becomes very useful to know. In particular, the probability $P_{i,j}$ of a specific base pair in the ensemble of structures of an RNA sequence is given by summation of all structures containing that base

pair:

$$P_{i,j} = \frac{1}{Q} \sum_{\substack{s \in \Omega: \\ (i,j) \in s}} e^{-\Delta G(s)/RT} \tag{5}$$

The sum in equation 5 closely resembles equation 4 and can thus be seen as a constrained partition function. Although individual base pair probabilities seem especially useful, this may also be applied to ensembles with arbitrary shared structural features. Effectively, the partition function computation for the ensemble of all secondary structures of an RNA sequence comes down to considering disjoint subensembles characterised by specific structural features. Indeed, McCaskill developed an algorithm computing the partition function (and base pair probabilities) with the same polynomial time complexity as the Zuker algorithm [**mccaskill_equilibrium_1990**].

This is made possible due to the functional equation $f(x+y) = f(x)f(y)$ satisfied by the exponential function in equation 4, allowing to decompose subensembles according to the constituents of their defining structural features after a change of algebra. A similar problem was already solved efficiently by Zuker's algorithm. However, in contrast, the uniqueness of the loop decomposition is paramount for McCaskill's partition function computation to ensure disjointness of the subensembles.

1.2 Pseudoknots

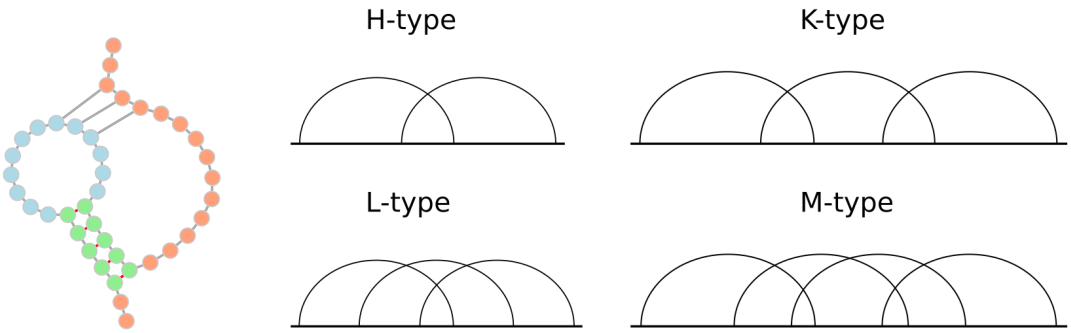
Pseudoknots are structural elements usually considered to be part of the tertiary structure [**hofacker_rna_2006**]. This view is embedded in the definition of secondary structure used in this thesis since pseudoknots are characterized by crossing base pairs and therefore violating rule (iii) (see section 1). In contrast to other tertiary interactions, they form canonical base pairs analogous to nested structures.

However, allowing base pairs to cross each other dramatically increases the space of possible structures for an RNA sequence because base pairs may cross as many other pairs as are available. In contrast to nested structures, the prediction of arbitrary pseudoknot structures is NP-complete [**lyngso_rna_2000**]. For that reason, many frameworks for pseudoknot algorithms that are available usually consider different, sometimes overlapping subclasses of all possible pseudoknots [**fallmann_recent_2017**]. Naturally, this also entails different, potentially overlapping terminologies for pseudoknots [**ponty_combinatorial_2011**, **mohl_lifting_2010**, **reidys_combinatorial_2011**].

Despite the daunting complexity, it is still important to account for pseudoknots as they do occur in natural RNA and are relevant for the function of RNAs [**janssen_investigating_2011**, **haslinger_rna_1999**]. This is also the case for

the object of interest in this work, as described later in section 2. Fortunately, most naturally occurring structures with pseudoknots tend to be relatively simple in the sense that their *crossings* are not interlaced and may be viewed as superpositions of two nested secondary structures (bi-secondary structures) [**haslinger_rna_1999**].

Indeed, for the purpose of this work, the simple class of H-type (hairpin) pseudoknots is sufficient to consider. Those pseudoknots are readily constructed as a superposition of two hairpins which is easier to visualize (Figure 3a) than to explain in words. Another bi-secondary pseudoknot is the K-type pseudoknot consisting of two *kissing* hairpins (Figure 3b). In contrast, L- and M-type pseudoknots as seen in Figure 3b are not bi-secondary.



(a) A simple H-type (hairpin) pseudoknot.

(b) More common pseudoknot types. The simplified H-type arc diagram can be visualized as *stretching* **(a)** into a straight line.

Figure 3: **(a)** An H-type pseudoknot. The dot-bracket notation `..(((((.......[[[...)))))..........]]]..` requires additional types of brackets in order to remain unambiguous. **(b)** More types of pseudoknots, modified after [**kucharik_pseudoknots_2016**]. Both the H-type and K-type (kissing hairpin) pseudoknot are bi-secondary in the terminology of [**haslinger_rna_1999**].

Aside from the complexity of arbitrary pseudoknots, integrating pseudoknots into the nearest-neighbor model is intricate since their energy contribution tends to be influenced by tertiary interactions [**liu_fluorescence_2010**].

Nevertheless, the free energy of pseudoknots is often modelled similar to nested loop types by considering stacks of base pairs to be stabilizing and the enclosed loops as destabilizing [**gultyaev_approximation_1999**]. While the former is straightforward, the latter is not exactly covered by the nearest-neighbor model.

Still, the approach for multiloops in nested structures (see equation 2) inspires a common approximation for pseudoknots. The destabilizing effect of a pseudoknot is approximated as the sum of a constant penalty (or initiation cost), a penalty depending on the number of enclosed unpaired nucleotides, and a penalty depending on the adjacent base pairs of the pseudoknot, i.e. the number of the participating crossing base pair stacks [**dirks_partition_2003**]. For an H-type pseudoknot

where only two stacks of base pairs are involved, computing this approximation is relatively simple. Yet, even for K-type pseudoknots, some form of recursion is necessary. A partition function algorithm of complexity $O(N^5)$ has been devised based on this approximation and is available in `NUPACK` [**dirks_partition_2003**, **zadeh_nupack_2011**].

Then again, for structures containing few and simple pseudoknots, the McCaskill algorithm for nested structures (see section 1.1.2) could be sufficient. Base pair probabilities obtained from it indicate possible pseudoknots, as was observed by Gaspin and Westhof and applied to *Tetrahymena* by Mathews [**gaspin_interactive_1995**, **mathews_using_2004**].

In the context of this work, secondary structures with pseudoknots are referred to as *crossing* structures to emphasize the distinction to nested structures or disambiguate where necessary.

2 The *Azoarcus* Group I Intron

A large class of ribozymes are the so-called group I introns. These catalytically active RNA molecules are common in a large variety of genomes, being found in bacteria, eukaryotes, organelles, even in archaea and viruses [**zhou_gissd_2008**, **nawrocki_group_2018**]. Although being very variable in primary sequence composition and length, defining characteristics of these ribozymes are high conservation of their secondary structure and a shared primary catalytic function; self-splicing from the precursor RNAs they are embedded in [**zhou_gissd_2008**].

The bacterium *Azoarcus sp. BH72* contains a group I intron located in the precursor of its Isoleucine codon-binding tRNA (pre-tRNA$^{\mathrm{Ile}}$), which is compelling for multiple reasons. With a sequence length of around 200 nucleotides, it is a particularly small group I intron. It has been shown to retain its catalytic activity at high temperatures up to 80 °C [**tanner_activity_1996**]. Furthermore, its outstanding capability of self-assembly (see section 2.3 for details) from smaller fragments piques interest in the context of the RNA world hypothesis.

2.1 Structural Features

Since RNA function is in large parts influenced by structure, taking a closer look at its structure beforehand is valuable in understanding the function of the *Azoarcus* group I intron. Not less importantly, some details of the ribozymes structure should be considered to increase the likelihood of successful RNA sequence design.

The conserved catalytic core of group I introns consists of two structural domains, P4-P6 and P3-P9, of which the latter only contains the helices P3, P7, P8 and P9 [**tanner_joining_1997**]. These domains are displayed as a secondary structure in Figure 4.

The presence of the (scaffold) domain P4-P6 is not strictly required to retain the catalytic function of the intron, albeit removal reduces the efficiency of the catalytic domain [**hayden_intramolecular_2015**]. Yet, crucial components of the *Azoarcus* group I intron structure are, in fact, part of its tertiary structure; the helices P3 and P7 form an H-type pseudoknot. This pseudoknot is of importance as P7 contains the guanosine binding site required for the introns self-splicing activity (see section 2.2) [**kuo_characterization_1999**]. This binding site is in direct proximity of the 1 nt bulge in P7 as is often the case for recognition and binding motifs [**turner_bulges_1992**, **hermann_rna_2000**]. Additionally, the nucleotide identities in P7 are highly conserved among group I introns [**zhou_gissd_2008**]. In other group I introns, it has been shown that tertiary interactions contribute

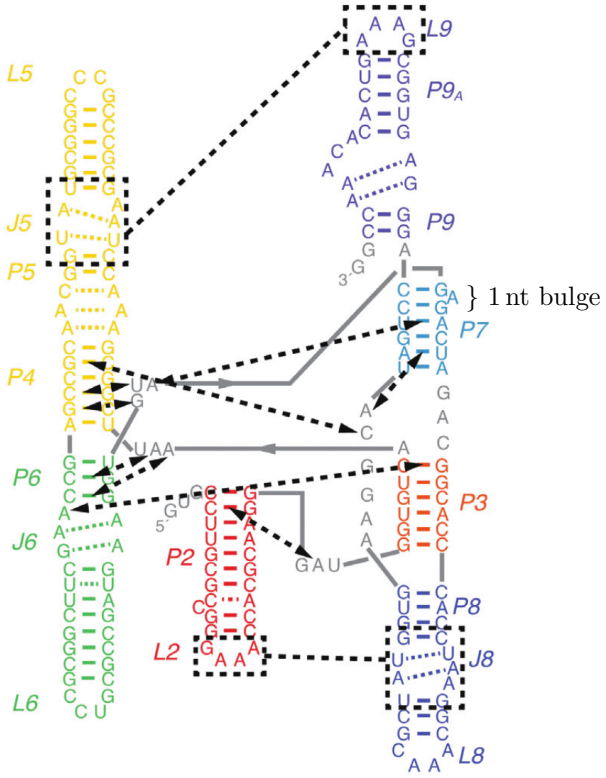


Figure 4: Secondary structure of the *Azoarcus* Group I Intron. Dashed black lines indicate long-range tertiary interactions, and colored dashed lines indicate noncanonical base pairs. Note that P3 and P7 form a pseudoknot and that P7 contains a single-nucleotide bulge. This secondary structure corresponds to `PDB 1U6B` [**adams_crystal_2004**]. Modified after [**mustoe_secondary_2016**].

to the formation of the catalytic core [**tanner_joining_1997**]. The tertiary interactions in `Azoarcus` are relatively non-specific with respect to the sequence but largely determined by its native secondary structure [**mustoe_secondary_2016**] (Figure 4).

2.2 Self-Splicing

Self-splicing in *Azoarcus* follows the same two-step mechanism as seen in other group I introns [**gleitsman_kinetic_2014**]. The first step is initiated by exogenous guanosine monophosphate (αG) binding at the active site of the intron (*pre-1S*) followed by an attack of the αG 3'-hydroxy group on the phosphodiester bond at the 5'-exon splice site (*post-1S*, see Figure 5) [**adams_crystal_2004-1**, **gleitsman_kinetic_2014**]. It is now phosphorylated at the 5' end of the intron.

Following a conformational change of the intron, the second step results in splicing both exons; the 3'terminal G (ΩG) of the intron binds to the binding site (*pre-*

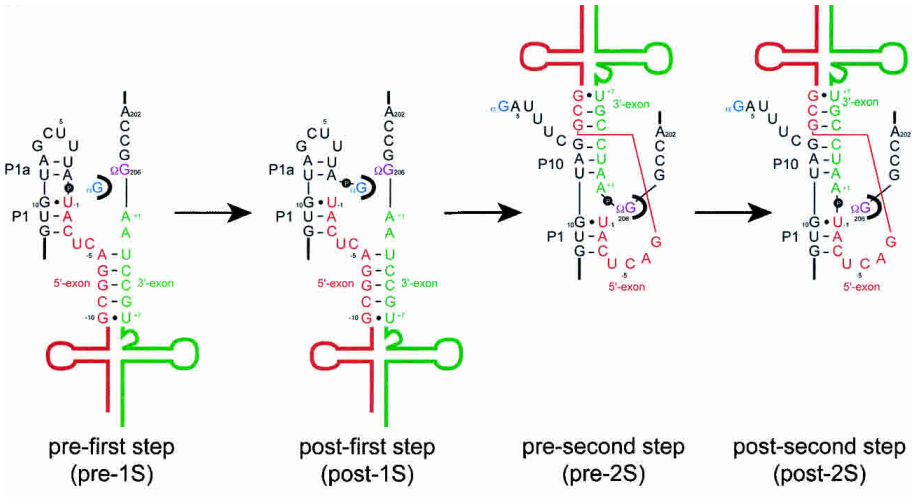


Figure 5: Overview of the self-splicing reaction. The guanosine binding site is stylized as a black pocket. Splice sites are marked as bold dots on backbone bonds. The change of orientation of the exons in *pre-2S* and *post-2S* indicates the conformational change of the ribozyme during self-splicing. Taken from [**adams__crystal__2004-1**].

2S), which allows the hydroxy group now attached at the 5'-exon to attack the phosphodiester bond. This enables a transfer of the hydroxy group to ΩG and immediate ligation of both exons (*post-2S*) [**adams__crystal__2004-1**]. This two-step transesterification requires proximity between both splicing sites and the guanosine binding site, which is maintained after the first step by binding the 5'-exon to the internal guide sequence (IGS).

This mechanism is not limited to self-splicing of the intron; spliced-out ribozymes catalyze cleavage and ligation of oligonucleotides resembling the 5'-exon [**kuo__characterization_**
gleitsman__kinetic__2014]. A related reaction scheme is later described in section 4.2.1.

2.3 Self-Assembly

Building on the primary self-splicing activity characteristic for group I introns, the *Azoarcus* group I intron has been shown to catalyze self-assembly from smaller inactive fragments by **hayden__self-assembly__2006** [**hayden__self-assembly__2006**], making it a promising candidate to support the RNA world hypothesis.

Similar to its self-splicing reaction mechanism via a two-step transesterification, the *Azoarcus* group I intron can catalyze the recombination and ligation of RNA sequences where a trinucleotide `CAU` complementary to the IGS is accessible as the recognition motif [**hayden__rna-directed__2005**].

Leveraging this mechanism, **hayden__self-assembly__2006** placed `CAU` trinucleotides into the L5, L6 and L8 hairpin loop regions of the intron, which then was

split into four inactive fragments [**hayden_self-assembly_2006**]. Incubation of the inactive fragments leads to (non-covalent) *trans*-assembly of a functional RNA complex catalyzing ligation of the incubated fragments by recombination (Figure 6) [**hayden_self-assembly_2006**, **hayden_systems_2008**].

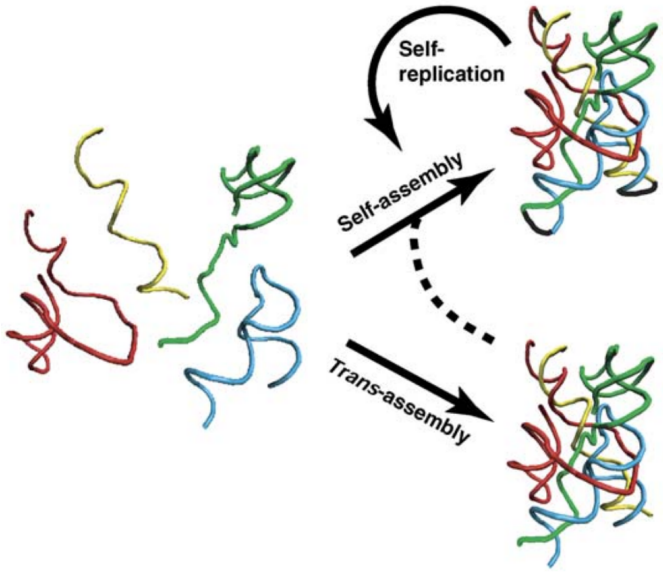


Figure 6: Formation of a *trans*-complex from inactive fragments catalyzes covalent assembly of the *Azoarcus* ribozyme which then in turn facilitates self-assembly. Taken from [**hayden_self-assembly_2006**].

Due to its autocatalytic replication capabilities, the *Azoarcus* group I intron is a vital model system supporting the RNA world hypothesis. Even more so, by modifying the IGS and the corresponding recognition motifs, cooperative replicator networks can be constructed. Specifically, **vaidya_spontaneous_2012** partitioned the *Azoarcus* ribozyme into two fragments instead of four by selecting a single site of the three loops used by **hayden_self-assembly_2006** [**vaidya_spontaneous_2012**]. Because of the three loops, this yielded three distinct fragment pairs assembling into the same ribozyme. The recognition motif of each fragment pair was modified to be unique with respect to the other two. Replacing the IGS of each fragmented ribozyme matching to the recognition motif of one of the others lead to a network of three catalysts, cooperatively assembling each other like in a biomolecular game of *rock-paper-scissors*.

2.4 Native Sequence and Target Structure Data

Since a comprehensive set of tertiary interactions in the *Azoarcus* group I intron was compiled in [**mustoe_secondary_2016**], the sequence and structure data used there served as the starting point for this work (Figure 7a).

The tertiary structure the simulations in [**mustoe_secondary_2016**] were based on originated from crystallography, which was facilitated by co-crystallization of an RNA binding protein and therefore required the insertion of a matching binding site into the intron sequence itself [**adams_crystal_2004**].

```
                gccguGUGCCUUGCGCCGGGAAACCACGCAAGGGAUGGUGUCAAAU
UCGGCGAAACCUAAGCGCCCGCCCGGGCGUAUGGCAACGCCGAGCCAAGCUUCGCAGCC
AUUGCACUCCGGCUGCGAUGAAGGUGUAGAGACUAGACGGCACCCACCUAAGGCAAACG
CUAUGGUGAAGGCAUAGUCCAGGGAGUGGCGAAAGUCACACAAACCGG
                ........((((((..((....)).)))))))...((((((....
((((((...((...((((((....))))))..))...))))))(((...(.(((((((((
..........)))))))).)..)))...[.[[[[[...))))))((((..((((....)
)).)))))......]]]]]].((..(((((....)))))).....))..
```

(a) Sequence and secondary structure as used in [**mustoe_secondary_2016**].

```
gcggacucauauuucgauGUGCCUUGCGCCGGGAAACCACGCAAGGGAUGGUGUCAAAU
UCGGCGAAACCUAAGCGCCCGCCCGGGCGUAUGGCAACGCCGAGCCAAGCUUCGGCGCC
           UGCGCCGAUGAAGGUGUAGAGACUAGACGGCACCCACCUAAGGCAAACG
CUAUGGUGAAGGCAUAGUCCAGGGAGUGGCGAAAGUCACACAAACCGGaauccguugg
.......(((......)))..((((((..((....)).)))))))...((((((....
((((((...((...((((((....))))))..))...))))))(((...(.((((((..
          ..))))))).)..)))...[.[[[[[...))))))((((...(((....)
))..))))......]]]]]]..((.(((((....)))))).....)).....
```

(b) Sequence and secondary structure taken from GISSD [**zhou_gissd_2008**].

Figure 7: Sequence and structure data used for the design of *Azoarcus*-like sequences. While **(a)** was taken from [**mustoe_secondary_2016**], the data corresponds to the structure of PDB 1U6B obtained through crystallography [**adams_crystal_2004**]. **(b)** was taken from GISSD and lacks a region added to 7a for the crystallography (boxed) [**adams_crystal_2004**]. Positions in gray indicate truncations dictated by requirements for experimental assays. Lower-case letters indicate positions not part of the intron.

With experimental assays in mind at later stages, data from the Group I Intron Sequence and Structure Database (GISSD) was taken into account [**zhou_gissd_2008**]. The sequence data from GISSD differed from the sequence data used in the crystallography mentioned above, primarily in the omission of the binding site for the co-crystallized protein (Figure 7b).

Moreover, positions of the exons adjacent to the intron were disregarded at later stages in this work (Figure 7b).

3 Methods

The methods used in this work are based on the overall goal of sequence design (see section 0.2). Implementations of different thermodynamical structure prediction algorithms and energy parameter sets of the underlying model are introduced in section 3.1.

In an approximation of inverse folding, the design pipeline described in detail in section 3.2 follows a general feedback loop starting from a candidate sequence whose predicted structure is compared to the target structure. By mutating candidate sequences and repeated structure prediction, the general design goal consists of exploring the sequence space and increasing similarity to the target [**dirks_paradigms_2004**]. Often, the distinction of positive and negative RNA design is made. The former attempts to increase affinity, whereas the latter aims to amplify specificity for a target structure [**dirks_paradigms_2004**]. Whether a design pipeline implements positive or negative design depends on the choice of concrete objective functions used to evaluate predicted structures (section 3.2.1).

Information about the exact versions of computational tools used in this work and the availability of self-written code was appended in section A.1.

3.1 Secondary Structure Prediction

A small variety of tools based on the thermodynamical model of secondary structures, as well as different sets of energy parameters, were assessed for their ability to recover large parts of the target structure, given the native sequence with a focus on its pseudoknot (see section 2.4).

3.1.1 `pKiss`

`pKiss` implements an algorithm for the prediction of minimum free energy (MFE) structures with different heuristics to predict a few kinds of simple pseudoknots [**janssen_rna_2015**]. Here, due to the simple H-type nature of the pseudoknot in the target structure, the legacy strategy only considering simple recursive pseudoknots from its predecessor `pknotsRG` was chosen [**reeder_design_2004**, **reeder_pknotsrg_2007**].

However, the class of simple recursive pseudoknots does not include helices with 1 nt-bulges as seen in region P7 of the *Azoarcus* group I intron. This was considered a relatively minor inaccuracy in this work and could be bypassed. The bulge was temporarily disregarded to estimate the energy of the native structure with its sequence (see section 4.1.1).

Experimental data for the energy contribution of pseudoknots tends to be scarce, and energy approximations have some limitations [**gultyaev_approximation_1999**, **fallmann_recent_2017**].

In `pKiss`, the H-type pseudoknot initiation cost (see section 1.2) is assumed constant with a default energy contribution of $9\,^{\text{kcal}}/_{\text{mol}}$. In the literature, initiation cost approximations range from $7\,^{\text{kcal}}/_{\text{mol}}$ [**rivas_dynamic_1999**] to roughly $10\,^{\text{kcal}}/_{\text{mol}}$ ($9.6\,^{\text{kcal}}/_{\text{mol}}$ in [**dirks_partition_2003**]), which is why values in this range were tested to possibly improve the prediction of the pseudoknot in the native sequence.

3.1.2 `RNAfold`

`RNAfold` from the `ViennaRNA` package [**lorenz_viennarna_2011**] was chosen for its speed and versatility despite being restricted to nested secondary structure prediction.

Structural constraints prohibiting positions in P7 from forming base pairs were applied during the prediction of minimum free energy structures to circumvent this limitation. This approach entailed constraints on candidate sequences at those positions in the design process (see section 3.2). The predicted free energies were then adjusted using the energy contribution of the base pairs in P7 (see Figure 13a). A penalty for implicitly adding the pseudoknot via prediction constraints was disregarded since this would not change the predicted structures.

Additionally, the `ViennaRNA` package provides an implementation of the McCaskill algorithm to compute the partition function and base pair probabilities for nested secondary structures. Using this feature enabled an alternative approach to the aforementioned constrained structure prediction; by examining dot plots of the base pair probabilities of an RNA sequence, potential pseudoknots can be extracted (see section 1.2).

A partition function algorithm implementation for structures with pseudoknots from `NUPACK` was briefly used for comparison [**dirks_partition_2003**, **dirks_paradigms_2004**

3.1.3 `RNAPKplex`

`RNAPKplex` provides minimum free prediction of structures containing at most one pseudoknot [**lorenz_viennarna_2011**, **beyer_rna_2010**]. The heuristic of `RNAPKplex` searches for intervals in an MFE structure that can be made accessible by removing base pairs to form a pseudoknot if the latter compensates the loss of stability by making intervals accessible. This algorithm was chosen for comparison since the energy calculation for its predictions is relatively similar to the constrained approach

using `RNAfold` albeit more general. For this reason, no pseudoknot initiation cost was used with `RNAPKplex` after initial tests.

During work on this thesis, basic functionality of `RNAPKplex` was restored and merged into `ViennaRNA` (see section A.1). As of version `2.4.18`, the overall implementation was extensively rewritten by Ronny Lorenz, meaning that the results in this work are not indicative of the latest changes to `RNAPKplex`.

3.1.4 Energy Parameter Sets

Although ultimately the default parameter set `turner2004` [**turner_nndb_2010**] was chosen for the structure prediction tools used here, multiple energy parameter sets (Table 1) for the nearest-neighbor model were tested for structure prediction with the reference data (see section 4.1.1).

The partition function algorithm of `NUPACK` was an exception to this because the energy parameter format is not compatible to the sets from above. This implementation was only used with its default parameter set.

3.2 RNA Design

The hereinafter described design pipeline was constructed with the *Azoarcus* GII used as the reference in mind. The primary design goal was to design sequences with similar structures as the reference data (see section 2). A strong focus was put on retaining structural features related to the catalytic activity of the *Azoarcus* GII (section 2.2). Specifically, the conserved P7 region containing a 1 nt bulge as the guanosine binding site is also involved in a pseudoknot with P3 (section 2.1). Additionally, some sequence constraints were imposed with tertiary interactions and catalytic activity in mind (see section 3.2.2 for details).

Two metrics were used to compare designs to the reference data. The *base pair distance* of two secondary structures measures the number of dissimilar base pairs. The (normalized) *ensemble defect* compares a structure ensemble of a sequence with a single structure and measures the average number (fraction) of incorrectly paired or unpaired nucleotides in the ensemble [**dirks_paradigms_2004**]. A detailed description of these measures used as objective functions is postponed until section 3.2.1 because the design approaches pursued in this work determined how they were used.

Design Approaches. Due to the pseudoknot in the target structure related to catalytic activity, using structure prediction methods capable of predicting pseudo-

Table 1: Energy parameter sets that were tested to improve structure prediction.

Energy Parameter Set	Reference	Notes
`turner1999`	[**turner_nndb_2010**]	based on experimental data
`turner2004`	[**turner_nndb_2010**]	based on experimental data
`andronescu2007`	[**andronescu_efficient_2007**]	computationally trained on structural and thermodynamical data
`CG*`	[**andronescu_computational_2010**]	extension of `andronescu2007`
`BL*`	[**andronescu_computational_2010**]	bayesian-learning based method
`langdon2018`	[**langdon_evolving_2018**]	evolutionary algorithm updating `turner2004` parameters

knots would have been a straightforward choice. Yet, for the design, computational methods explicitly modelling solely nested structures were used for multiple reasons. First and foremost, the nucleotide identities in the catalytically relevant P7 are conserved among group I introns, which was utilized with the first *constrained* approach described below. An *alternative* approach leveraged the observation of potential pseudoknots recovered by the McCaskill algorithm (cf. section 1.2).

Moreover, the prediction of structures with pseudoknots is more complex than without pseudoknots. It requires heuristics restricting types or amount of pseudoknots as done in `pKiss` and `RNAPKplex` respectively. These tools were only used to evaluate designed sequences as described in section 3.2.4.

Constrained Approach. For this approach, nucleotide identities at positions corresponding to P7 of the target structure were fixed to be the same as in the reference sequence to retain the conserved native guanosine binding site in sequence designs. With the assumption of hierarchical folding of RNA, i.e. tertiary interactions such as pseudoknots forming after secondary structure, the design process was simplified by explicitly prohibiting base pairs in P7 during structure prediction using `RNAfold` (see section 3.1.2). Consequently, base pairs of P7 were disregarded during the computation of objective functions relative to the target structure.

Alternative Approach. As described in section 1.2, even a partition function algorithm only considering nested structures as implemented in `ViennaRNA` may recover base pair probabilities for potentially pseudoknotted positions. This enabled a design approach without structural and sequence constraints on P7 positions during structure prediction while still not explicitly modelling pseudoknots.

With base pairs of P3 and P7 being involved in the pseudoknot, the target structure was split into two pseudoknot-free structures, each containing the base pairs of either P3 *or* P7 in addition to all base pairs not involved in the pseudoknot (cf. Table A.2). Designing for two target structures entails an important subtlety necessary to address in the design pipeline; there is an inherent conflict in designing sequences with a predicted structure as close as possible to both target structures.

3.2.1 Objective Functions

In this work, two metrics were used for comparison between secondary structures but also between Boltzmann ensembles of a sequence with a single structure. Although their actual use in the implemented pipeline may be more apparent in section 3.2.3, the technical details are introduced below.

Base Pair Distance. The base pair distance of two secondary structures is a widely used and computationally cheap metric in the mathematical sense. Interpreting secondary structures as sets containing their base pairs as elements, the base pair distance is defined as the cardinality of their symmetric difference [**lorenz_rna_2014**]:

$$\mathrm{d}_{\mathrm{bp}}(s_1, s_2) = |s_1 \triangle s_2| = |(s_1 \cup s_2) \setminus (s_1 \cap s_2)| \tag{6}$$

While `ViennaRNA` provides a base pair distance implementation only for nested secondary structures, this measure is not restricted to them [**haslinger_rna_1999**].

Lifting the implementation to secondary structures with pseudoknots requires only the addition of base pairs that may cross. For this thesis, this was done by generating a low-level wrapper around the `ViennaRNA` library written in Rust and translating the implementation directly while replacing the pair tables used as structure representations. This implementation can be easily imported into Python as a library (see section A.1 for availability).

In general, the base pair distance is helpful in conjunction with MFE algorithms which produce a single predicted structure. Using this metric as an objective function to be minimized may be seen as negative RNA design.

Ensemble Defect. The ensemble defect of an arbitrary target structure s^*, given an RNA sequence r, is defined as the average number (over all structures in the Boltzmann ensemble Ω) of incorrectly paired or unpaired positions with respect to s^*. This can be expressed as

$$\begin{aligned} \mathrm{ed}_r(s^*) &= N - \sum_{s\in\Omega}\left[p(s)\sum_{1\le i,j\le N} S_{i,j}S^*_{i,j}\right] \\ &= N - \sum_{1\le i,j\le N}\left[\sum_{s\in\Omega} p(s)S_{i,j}\right]S^*_{i,j} \end{aligned} \tag{7}$$

where $p(s)$ is the probability of a structure in the Boltzmann ensemble, as defined using the partition function (equation 3), and $S, S^* \in \{0,1\}^{N\times N}$ are base pairing matrices, i.e. having entries set to 1 if their (unequal) indices form a base pair in the structures s, s^* and 0 otherwise [**dirks_paradigms_2004**]. For convenience, non-zero diagonal entries correspond to unpaired positions. Therefore, S and S^* are symmetric and doubly stochastic, i.e. their rows and columns sum up to one, respectively. Note that **dirks_paradigms_2004** use an additional column for unpaired positions [**dirks_paradigms_2004**]. The right-hand side expression in equation 5

$$P_{i,j} = \sum_{s\in\Omega} p(s)S_{i,j} \tag{8}$$

sums over all structures of the ensemble containing the base pair (i,j) for $1 \le j \le N$, or containing an unpaired position at i for $i = j$ respectively. A matrix P with entries from equation 8 is called a base pair probability matrix. Per definition, P is symmetric and doubly stochastic, just as S and S^*. Since P consists of N rows (columns), each summing up to one, the *normalized* ensemble defect is well defined:

$$\mathrm{ned}_P(s^*) = 1 - \frac{1}{N}\sum_{1\le i,j\le N} P_{i,j}S^*_{i,j} \tag{9}$$

The notation of $\mathrm{ned}_P(s^*)$ is a little overloaded. Like $\mathrm{ed}_r(s^*)$, the normalized ensemble defect implicitly depends on the RNA sequence r via P.

In practice, the probability of a position i being unpaired is often not directly stored in $P_{i,i}$, but computed as $1 - \sum_{j,j\neq i} P_{i,j}$. Base pair probability matrices visualized as *dot plots* follow this convention throughout this work.

Since there are algorithms of polynomial complexity available to compute the partition function and base pair probabilities (see section 1.1.2), the normalized ensemble defect can be used to measure the whole Boltzmann ensemble of a sequence given a target structure. Using ensemble defect based objective functions, simultaneous positive and negative RNA design can be implemented [**dirks_paradigms_2004**]. Note that the dependence on the sequence is implicitly given via the base pair probability matrix in the equations above.

In order to implement the alternative design approach given two target structures, the maximum normalized ensemble defect of two structures was used:

$$\widehat{\mathrm{ned}}_P(s_1^*, s_2^*) = \max\{\mathrm{ned}_P(s_1^*), \mathrm{ned}_P(s_2^*)\} \tag{10}$$

It is essential to mention that this objective function cannot be zero. Its minimal possible value depends on the two target structures used (see section 5.2 for a discussion of alternatives).

3.2.2 Sequence Constraints

Certain positions of design candidate sequences were fixed to the same nucleotide identities as in the (truncated) reference sequence (see section 2.4). These *sequence constraints* were employed to support the design for the desired function of the reference ribozyme. Due to the assumptions made for the constrained approach—namely, the pseudoknot folding after secondary structure and conservation of P7 —, nucleotides in P7 had to be constrained (see section 3.2). As a side effect, the 1 nt bulge acting as the guanosine binding site in the native ribozyme was not explicitly modelled. This is potentially advantageous as the stability of such bulges does primarily depend on non-nearest neighbor interactions [**blose_non-nearest-neighbor_2007**].

Additionally, the terminal position was constrained to guanine (`G`) because this nucleotide binds to the guanosine binding site in the native ribozyme at the *pre-S2* state during self-splicing (see Figure 5). Analogously, the first three nucleotides were left identical to the internal guide sequence (IGS) of the native ribozyme because the 5'-exon binds to it.

Further potential sequence constraints were defined by nucleotides involved in

tertiary interactions in the structure of the native ribozyme due to their stabilizing effects (see Figure 4). The tetraloop formed by P2 (L2) was considered in particular as it was shown to have a substantial effect on the thermostability of the *Azoarcus* GII [**tanner_activity_1996**]. In conjunction with the previously introduced sequence constraints, these constraints were considered `minimal` to use designed sequences *in vitro* for experiments. Other positions of tertiary interactions were included for a subset of all sequence designs to provide a set of `complete` sequence constraints.

Due to the similarity of the sequences obtained from [**mustoe_secondary_2016**] and GISSD, the tertiary interaction sites could be easily transferred to the latter sequence. The exact positions and nucleotide identities used as constraints are depicted in Figure 8.

GUGCCUUGCGCCGGGAAACCACGCAAGGGAUGGUGUCAAAUUCGGCGAAACCUAAGCGC
CCGCCCGGGCGUAUGGCAACGCCGAGCCAAGCUUCGGCGCCUGCGCCGAUGAAGGUGUA
GAGACUAGACGGCACCCACCUAAGGCAAACGCUAUGGUGAAGGCAUAGUCCAGGGAGUG
GCGAAAGUCACACAAACCGG

Figure 8: The sequence constraints used in the design procedure. The differently shaded positions were fixed during sequence sampling: IGS, terminal G, L2, tertiary interactions and P7. Note that `GAAA` is actually part of the tertiary interactions. The displayed sequence is the truncated version of Figure 7b.

Table 2 summarizes the defined constraint sets based on the individual constraints as depicted in Figure 8. Two additional constraint sets suffixed `-alt` for use in conjunction with an alternative objective function (see section 3.2.1) were defined by omitting the constrained positions of P7.

Table 2: The different combinations of constraints used. + marks the inclusion of a feature as introduced in Figure 8 into the set. Note that `proto` was only used with the non-truncated sequence in Figure 7a. The `*-alt` constraint sets were only used for designs applying the alternative objective function defined in equation 10. `#`: total number of fixed positions in this set.

Set	#	IGS	term. G	L2	tert. Interactions	P7
`minimal`	21	+	+	+	−	+
`complete`	67	+	+	+	+	+
`minimal-alt`	8	+	+	−	−	−
`complete-alt`	54	+	+	+	+	−
`proto`	63	−	−	+	+	+

Sequences designed using the constraints of the `proto` set differed from the other sequence designs; in this case, the untruncated target structure from [**mustoe_secondary_2016**]

was used (see Figure 7a). The reason for this deviation is that these sequences were the first designs for which preliminary experimental results are available.

3.2.3 Pipeline

With design approaches, objective functions and sequence constraints described, most parts of the general design feedback loop outlined in section 3 are in place for pipeline construction.

Exploring Sequence Space. Perhaps the simplest way to obtain a new RNA sequence of the same length from a given one consists of point mutations. In doing so repeatedly, the sequence space may be explored. Usually, the sequence space is viewed as a graph with sequences of equal length as vertices and edges between sequences differing in exact one nucleotide. Such a graph is called a Hamming graph, and the (Hamming) distance between any two vertices corresponds to the minimal number of point mutations between two RNA sequences [**reidys_generic_1997**].

For this work, some sequences are inherently less interesting than others; RNA sequences not *compatible* to the target structure should be disregarded (cf. rule (iv), section 1) [**gruner_analysis_1996**]. Ideally, a walk on the sequence space suitable for RNA sequence design would allow base pair mutations to move only between compatible sequences. The necessary moves enabling such walks are depicted in Figure 9.

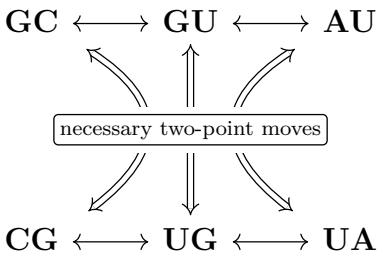


Figure 9: To move in the space of sequences compatible to a given structure, one-point mutations only cover unpaired compatible positions. Additionally, moves changing (up to) two points are necessary for positions that are involved in base pairs. The latter moves are shown above. One-point mutations are visualized as single double-ended arrows. Double arrows denote moves between base pairs that necessarily involve two-point mutations. For example, changing a base pair with nucleotides **CG** to **AU** requires changing both nucleotides in a single move. Otherwise, the mutated sequence would be incompatible to the given structure [**haslinger_rna_1999**].

In the implementation of the design pipeline, `RNAblueprint` [**hammer_rnablueprint_2017**] was used to enable exploration of the sequence space primarily for practical reasons. Most importantly, the library already existed, so no custom code had to be written.

Although sampling uniformly from sequence spaces compatible to a single (bi-)secondary structure is relatively simple to achieve by considering unpaired and paired positions separately, the graph-coloring approach of `RNAblueprint` guarantees uniform sampling with multiple structural constraints, enabling more elaborate RNA designs. As a bonus, the library provides easy integration of sequence constraints.

Pipeline Overview. The design pipeline resembles an adaptive walk on the space of sequences compatible to the target structure, minimizing the objective functions depending on the design approach.

Following the schematic in Figure 10, the design process is initialized by uniformly sampling a random RNA sequence using `RNAblueprint`. Instead of directly computing objective function scores for the initial candidate, conservative choices were made for the initial scores; N, the sequence length, as an initial score for base pair distance derived objective functions and 1 for normalized ensemble defect based objectives. At the next step, a point or base pair of the candidate sequence is mutated as a first move on sequence space. These first two steps of the pipeline are subject to sequence constraints and must be compatible to the target structure as previously described. Afterwards, MFE structure prediction and base pair probability computation is performed. This step differs between the design approaches; structural constraints are imposed in the constrained approach.

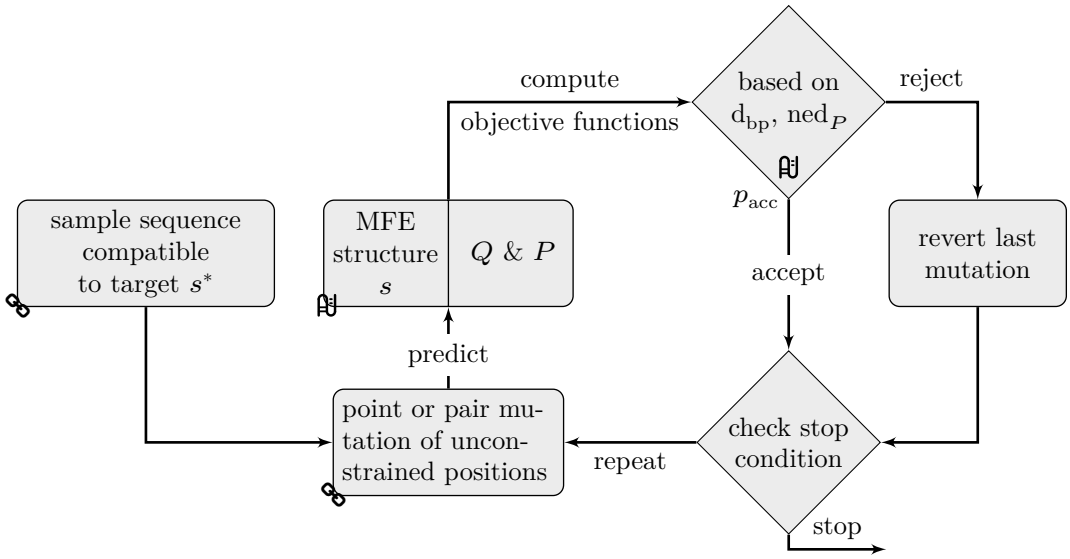


Figure 10: Schematic overview of the design pipeline resembling an adaptive walk on the space of sequences compatible to the target structure. Q and P denote the partition function and base pair probability matrix respectively. The objective function scores were compared to the scores of the previous step. In the context of RNA design, lower base pair distances d_{bp} and normalized ensemble defect values ned_P are seen as better. The precise minimization procedure including stop conditions is described in the main text for both constrained and alternative design approach. Objective function evaluation was relaxed by accepting worse-scoring candidates with probability p_{acc}. 🔗 indicates application of sequence constraints. Steps marked with ₪ depend on the design approach taken.

In the constrained approach, the normalized ensemble defect defined in equation 9 serves as the primary objective with the same structural constraints applied during structure prediction. The base pair distance of the MFE prediction to the target structure is used as a secondary objective only required not to increase with every iteration. Acceptance of the current candidate sequence is decided relative to the scores of the previous candidate sequence; candidates are accepted if their scores decrease or with some probability $p_{acc} = 0.001$ as a relaxation parameter in order not to get trapped in local minima. If rejected, the mutation step is reverted. The adaptive walk is stopped if $ned_P(s^*) \leq 0.05$ or the MFE prediction differs from the target structure by at most two pairs.

The pipeline variant for the alternative approach works very similar with some changes necessitated by the alternative objective function $\widehat{ned}_P$ given two *superposed* target structures (cf. equation 10, Table A.2). With that approach, no secondary objective is evaluated. The relaxation parameter is set higher $p_{acc} = 0.01$ to counteract walking into local minima more often due to the construction of the objective function. Additionally, the minimization is stopped at a higher threshold of $\widehat{ned}_P(s_1^*, s_2^*) \leq 0.15$ or when the pseudoknot-free base pair distance of the `RNAfold`

MFE prediction without structural constraints to one of both target structures was at most two. The choice of the higher score threshold is made plausible as follows:

Given a sequence design with $\mathrm{ned}_P(s_1^*) = 0$, the Boltzmann ensemble of the design would contain a single structure s_1^* and the alternative objective function would reduce to $\widehat{\mathrm{ned}}_P(s_1^*, s_2^*) = \mathrm{ned}_P(s_2^*) = \frac{2\,\mathrm{d}_{\mathrm{bp}}(s_1^*, s_2^*)}{N} = \frac{24}{197} \approx 0.122$ with the specific target structures from Table A.2. The factor 2 arises from the fact that the ensemble defect sums over *all* incorrect nucleotides, both paired and unpaired (see equation 7). Therefore, further minimization would necessarily increase $\mathrm{ned}_P(s_1^*)$. This is, of course, the extreme case and halving this estimation yields the minimum possible value of $\widehat{\mathrm{ned}}_P$. Yet, the chosen threshold was considered sufficient for the purpose of this work.

3.2.4 Quality Control and Selection of Designed Sequences

There are two primary reasons why a notion of *good* sequence designs is required in this work. Most importantly, the design pipeline itself only relies on computational methods developed for comparison and prediction of nested secondary structures, but the reference ribozyme contains prominent tertiary structural features; the pseudoknot formed between P3 and P7 and a 1 nt bulge.

In addition, $n = 1000$ sequences were designed for each constraint set. It is practically infeasible to assess that amount of different RNA sequences *in vitro* with regard to their catalytic activity. In consequence, only subsets of those sequence designs were selected for later use in experimental assays. MFE structures predicted using `pKiss`, and `RNAPKplex` were used to assess recovery of the pseudoknot present in the reference data and similarity to the target structure in general.

For sequences designed with the constrained approach, `RNAfold` with constraints was used as well. Recovery of the 1 nt bulge was not used as an explicit criterion because `pKiss` does not model bulges in pseudoknots, and sequence constraints were applied on P7 nucleotides in the first place. With the alternative design approach, no constraints were applied at P7, so `RNAfold` was used without constraints.

With three MFE predictions per designed sequence, the quality of the designs was measured as a consensus by computing the mean and standard deviation of the base pair distance between predicted structures and target structure. For each constraint set, subsets containing the best sequences according to this metric were manually checked for the presence of the pseudoknot. Furthermore, base pair probability matrices computed using the McCaskill algorithm were compared to the target structure for individual sequence designs.

3.3 Neutral Paths

The map between sequences and secondary structures also entails questions about genotype subspaces of RNA sequences of the same phenotype. Analogous to neutral mutations, subspaces consisting of sequences folding into the same structure are called neutral networks [**gruner_analysis_1996**]. Neutral networks percolating RNA sequence space have been shown to exist depending on the nucleotide alphabet [**schuster_sequences_1994**, **gruner_analysis_1996**, **reidys_generic_1997**], even in the case of crossing structures [**haslinger_rna_1999**].

To give a lower bound on their extent, neutral paths can be used. A neutral path is a walk starting from a random RNA sequence and iterating over neutral neighbors such that the distance to the initial sequence increases with each step until this is no longer possible. For convenience, the Hamming distance may be used [**reidys_generic_1997**]. Analogous to section 3.2.3, it is sufficient to restrict the view on sequences compatible with a shared structure. In consequence, neutral neighbors may be reached via point or base pair mutations. Neutral paths starting from sequence designs were used to estimate the extent of the sequence designs' neutral networks. To put neutral path lengths into perspective, it may be noted that the expected Hamming distance of two random RNA sequences is $\mathbf{E}[\mathrm{d_H}(r_1, r_2)] = {}^3\!/_4 N$ where N is the sequence length and therefore the maximum Hamming distance. However, this estimation does not define a threshold for percolation. It is not directly apparent that this estimation also applies to the restriction on compatible sequence because sequence spaces compatible to a structure depend on the number of base pairs in the structure. With $N = u+2b$, where u is the number of unpaired positions and b is the number of base pairs in the structure, the expected Hamming distance between to RNA sequences r_1, r_2 compatible to a shared structure can be estimated as follows:

$$\mathbf{E}\left[\mathrm{d_H}(r_1, r_2)\right] = \frac{3}{4}u + \frac{5}{6}b\left(\frac{2}{6}\left(2\frac{3}{5} + \frac{2}{5}\right) + \frac{4}{6}\left(2\frac{4}{5} + \frac{1}{5}\right)\right) = \frac{3}{4}u + \frac{13}{18}2b \qquad (11)$$

In equation 11, the previous estimation remains valid for the u unpaired positions. For the b pairs, the metaphor of having an unknown RNA sequence and trying to guess the paired positions might be helpful as an intuition:

Because there are six allowed base pairs, one may expect to incorrectly guess ${}^5\!/_6$ of the b pairs. Recalling Figure 9 and assuming independence of base pairs, some cases have to be considered. If the *correct* pair were **GU** or **UG**, there would be three of five remaining pairs to get both positions involved in the base pair wrong and two of five remaining pairs with one correctly guessed nucleotide in the pair.

Analogously, with the *correct* base pair being any other than **GU** or **UG**, there would be four of five cases of incorrectly guessing both paired positions.

4 Results

The results obtained in this work are primarily of numerical kind. The following sections evaluate structure prediction methods and energy parameter sets with respect to the reference data, as well as designed sequences generated using two design approaches. For initial designs, preliminary experimental results are available (section 4.2).

4.1 Numerical Results

4.1.1 Structure Prediction of *Azoarcus*

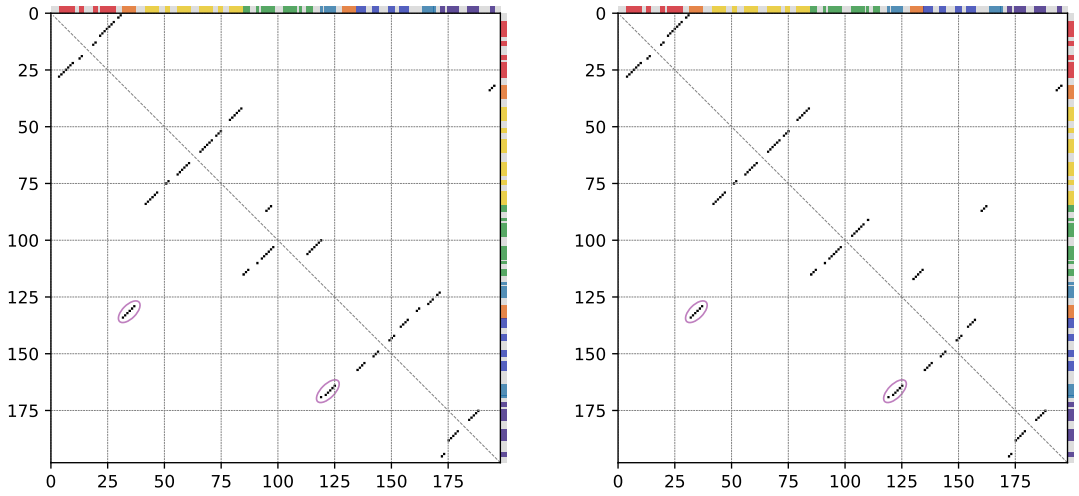
Minimum Free Energy Predictions. Using the default energy parameter set `turner2004`, the structure prediction tools introduced in section 3.1 were assessed based on their ability to recover the native structure of the *Azoarcus* group I intron. For this, the truncated sequence and structure from Figure 7b was used.

As summarized in Table 3, `RNAfold` had the fastest runtime. Also, by excluding the positions in P7 from forming base pairs using `RNAfold`, the base pair distance to the pseudoknot-free target structure was noticeably reduced. With a value of $-74.20\,\mathrm{kcal/mol}$, this prediction yielded the lowest predicted energy in this comparison. However, adding the computed energy contribution of the missing P7 base pairs from Figure 13a, an adjusted value of $-77.10\,\mathrm{kcal/mol}$ is somewhat more in line with the other MFE predictions.

Table 3: The different tools were tested on the *Azoarcus* Group I intron (see Figure 7b). ΔG: Gibb's free energy of the predicted structure, $\mathrm{d_{bp}}$: base pair distance to the assumed native structure (with pseudoknots except values marked with $*$). Here, the default `turner2004` parameters were used. See Table A.3 for the concrete predicted structures.

Tool	Runtime	ΔG	$\mathrm{d_{bp}}$	Notes
`RNAfold`	55 ms	-77.40 kcal/mol	45*	—
`RNAfold`	49 ms	-74.20 kcal/mol	29*	base pairs at P7 were forbidden
`RNAPKplex`	191 ms	-77.40 kcal/mol	51	default pseudoknot penalty
`RNAPKplex`	595 ms	-79.80 kcal/mol	53	no pseudoknot penalty
`pKiss`	683 ms	-83.90 kcal/mol	56	default H-type penalty
`pKiss`	674 ms	-81.60 kcal/mol	32	H-type penalty = 9.8 kcal/mol

Dot plots of the unconstrained and constrained `RNAfold` prediction confirm that applying this type of constraint improves the prediction of the rest of the structure, although some new false interactions appear (Figure 11).



(a) no constraints were applied during prediction.

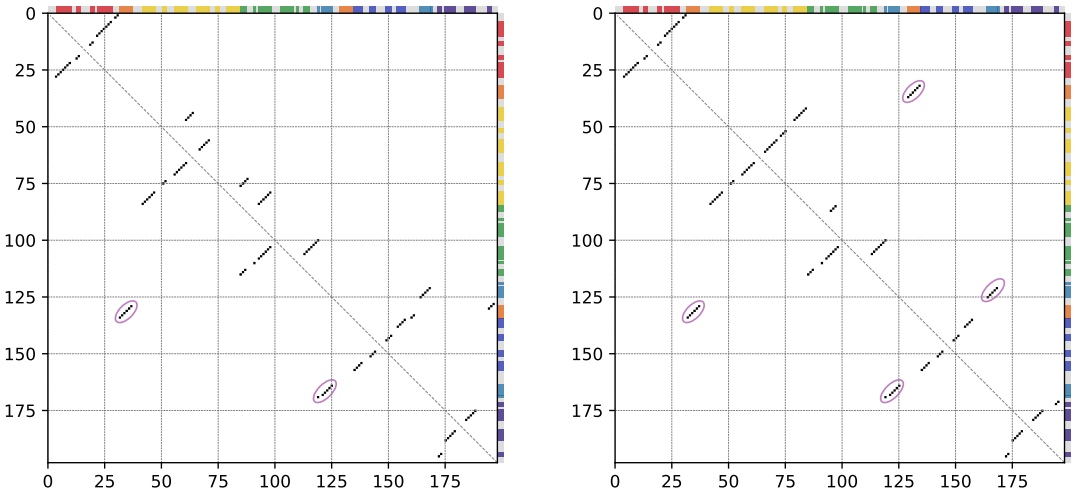
(b) positions in P7 were excluded from forming base pairs.

Figure 11: The structure of the *Azoarcus* Group I intron as predicted by `RNAfold` with **(a)** no structural constraints and **(b)** structural constraints at P7 positions disallowing base pairs. The lower diagonal half of these figures shows the target structure, the upper diagonal half contains the prediction. Base pairs between two positions are indicated as black squares with position indices labelled at both axes. Helices of the native structure are color-coded on top and at the right according to Figure 4. The location of the original P3-P7 pseudoknot is highlighted in purple.

Changing the pseudoknot penalty of `RNAPKplex` did not produce dramatic changes in its structure prediction (Table 3). However, for the rest of this work, the pseudoknot penalty of `RNAPKplex` was set to zero to be approximately consistent with the constrained `RNAfold` prediction. Increasing the H-type initiation penalty of `pKiss` to $9.8\,{}^{\text{kcal}}/_{\text{mol}}$ notably improved the base pair distance of the prediction to the target structure (Table 3). In fact, the MFE structure predicted using the increased penalty recovered the pseudoknot formed by P3 and P7 almost completely (Figure 12), which is why this modified penalty was used with `pKiss` throughout this work.

Only a single base pair of the pseudoknot was missing in the predicted structure due to limitations of the heuristic model of the algorithm not including pseudoknots with bulges [**reeder_design_2004**] (Figure 12b). The most prominent inaccuracy of the prediction in Figure 12b corresponds to the P6 region of the native structure. However, P6 is part of the scaffold domain of the *Azoarcus* ribozyme and not directly required for catalytic activity. For this reason, the misprediction of this region was assumed negligible.

More general heuristics implemented in `pKiss` have the same limitation of not modelling bulges in pseudoknots and were disregarded for this reason and due to



(a) `--Hpenalty 9.0` **(b)** `--Hpenalty 9.8`

Figure 12: The structure of the *Azoarcus* Group I intron as predicted by `pKiss` with different penalties for H-type pseudoknots: **(a)** the default penalty of 9.0 kcal/mol does not recover the pseudoknot formed between P3 and P7. **(b)** a modified penalty of 9.8 kcal/mol recovers the pseudoknot of the target almost completely. Note the mispredicted structure of P6 in the center of the dot plot. The lower diagonal half of these figures shows the target structure, sthe upper diagonal half contains the prediction. Base pairs between two positions are indicated as black squares with position indices labelled at both axes. Helices of the native structure are color-coded on top and at the right according to Figure 4. The location of the original P3-P7 pseudoknot is highlighted in purple.

their time and memory complexity. Although this limitation also means that `pKiss` cannot calculate the free energy of the target structure given the native sequence, this is a relatively minor problem.

By removing the bulge position of P7 as displayed in Figure 13b, the free energy of the modified target structure can be calculated using `pKiss` and yields a value of $\Delta G = -80.90\,\mathrm{kcal/mol}$. Adjusting this result by the difference of including the bulge via the values in Figure 13, the native structure has a free energy of approximately $\Delta G = -77.10\,\mathrm{kcal/mol}$ (or $\Delta G = -76.30\,\mathrm{kcal/mol}$ with the modified H-type pseudoknot penalty).

Partition Function Predictions. Following the observation of potential pseudoknots using partition function algorithms (see section 1.2), base pair probability matrices of the *Azoarcus* group I intron were computed via the McCaskill algorithm implementation of `ViennaRNA` and the partition function algorithm of `NUPACK` (Figure 14).

Indeed, non-zero probabilities were predicted for positions close to the pseudo-

```
AGAGACUAG&AUAGUCCA
.(.((((((.&.)))))).
```

$\Delta G = -2.90\,\mathrm{kcal/mol}$

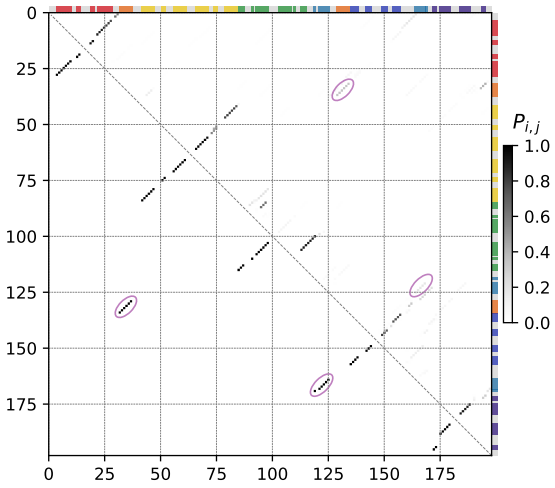
(a) Base pairs of P7 with a 1 nt bulge in dot-bracket notation and the free energy relative to the open chain.

```
AGGACUAG&AUAGUCCA
.((((((.&.)))))).
```

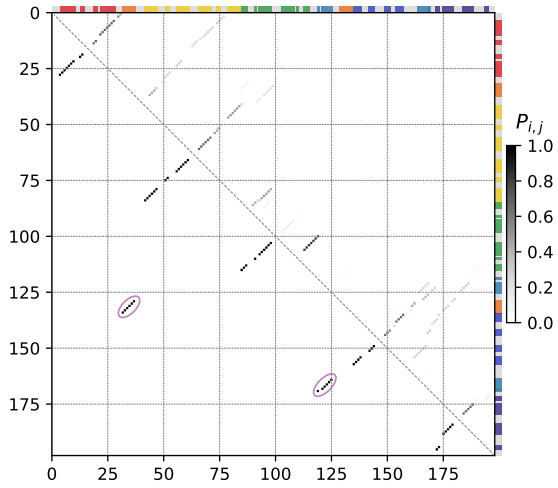
$\Delta G = -6.70\,\mathrm{kcal/mol}$

(b) Removing the bulged position from P7 stabilizes the stack.

Figure 13: Free energy of the base pairs in the region P7 of the *Azoarcus* group I intron **(a)** with the 1 nt bulge present and **(b)** without the 1 nt bulge present. Computed with `RNAeval -d 0` and the default energy parameters.



(a) Base pair probabilities of the sequence as computed by the implementation of the McCaskill algorithm implemented in `ViennaRNA`.



(b) Base pair probabilities of the sequence as computed by the partition function algorithm implemented in `NUPACK`.

Figure 14: Dot plots displaying base pair probabilities for the *Azoarcus* group I intron computed by **(a)** a partition function not explicitly modelling pseudoknots **(b)** a partition function accounting for pseudoknotted structures in the ensemble. The lower diagonal half of these figures shows the target structure. Position indices of $P_{i,j}$ are labelled at both axes. Helices of the native structure are color-coded on top and at the right according to Figure 4. The location of the original P3-P7 pseudoknot is highlighted in purple.

knot present in the native structure, despite using a partition function of a Boltzmann ensemble of nested structures (Figure 14a).

The base pair probabilities predicted by `NUPACK` did not match the native pseudoknot despite modelling pseudoknots in their algorithm. (Figure 14b).

Energy Parameter Sets. So far, only the default energy parameters `turner2004` have been used to examine the minimum free energy structure prediction quality of the tools used for this specific ribozyme.

In Figure 15, both free energy and base pair distance to the target structure are shown for multiple sets of energy parameters that were introduced in Table 1. Most of the parameter sets used here were generated using computational methods. Only `turner1999` and `turner2004` contain parameters obtained from experimental measurements.

Although some of the computationally generated parameter sets showed promising improvements on `turner2004`, the decisions made here were conservative to prevent overfitting in the design pipeline.

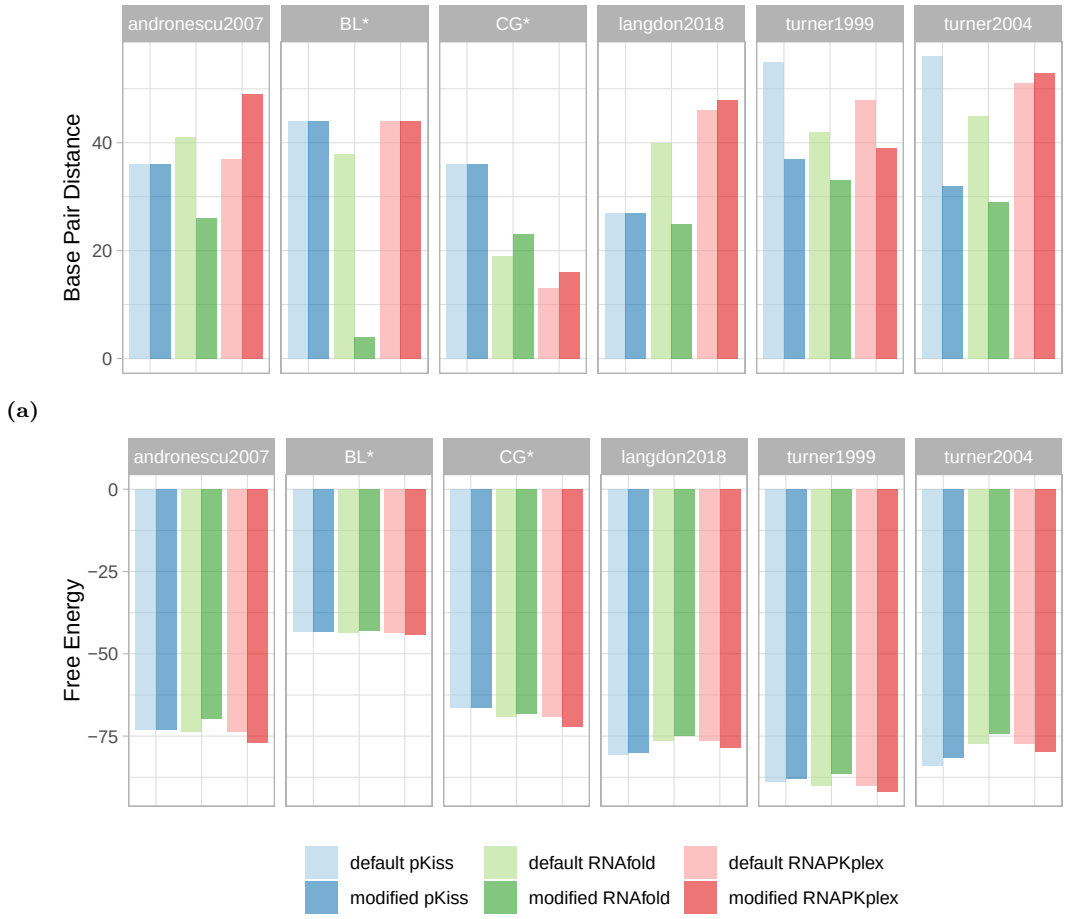


Figure 15: Minimum free structure predictions of the *Azoarcus* group I intron using the same tools and modifications as in Table 3 but with different energy parameter sets (see Table 1): **(a)** the predicted base pair distance to the target structure and **(b)** the free energy of the predicted structure. Note that the values for `turner2004` in the rightmost column correspond to Table 3.

As seen in Figure 15a, `BL*` greatly improved the base pair distance of the constrained `RNAfold` prediction (green) compared to every other parameter set. However, due to the substantial deviation of the predicted free energy from every other parameter set used (Figure 15b), the biological irrelevance of the prediction using

these parameters was a significant concern.

Both `CG*` and `andronescu2007` were ruled out because the pseudoknot formed by P3 and P7 was not recovered at all by `pKiss` with these parameters (not shown, cf. section A.1). Although the base pair distance of the `RNAPKplex` prediction improved noticeably using `CG*`, the pseudoknot was not recovered as well.

Generally, predictions made with `RNAPKplex` were somewhat imprecise. Still, `RNAPKplex` predicted suboptimal structures with pseudoknots similar to the native structure, and was only used in conjunction with `RNAfold` and `pKiss` to assess quality of designed sequences (section 3.2.4). Only `langdon2018` yielded potentially useful improvements; both `pKiss` variants recovered the pseudoknot as well as with `turner2004` and the modified pseudoknot penalty. Still, the improvement relative to the default parameter set consisted of only a few base pairs for the already adjusted methods (constrained `RNAfold` and `pKiss` with a penalty of $9.8\,{}^{\text{kcal}}/_{\text{mol}}$).

Since the goal was to use designed sequences in experiments eventually, the conservative choice of the default parameter set `turner2004` seemed reasonable and justified.

4.1.2 Sequence Designs

For each set of sequence constraints (see Table 2), $n = 1000$ RNA sequence were designed using the pipeline described in section 3.2.

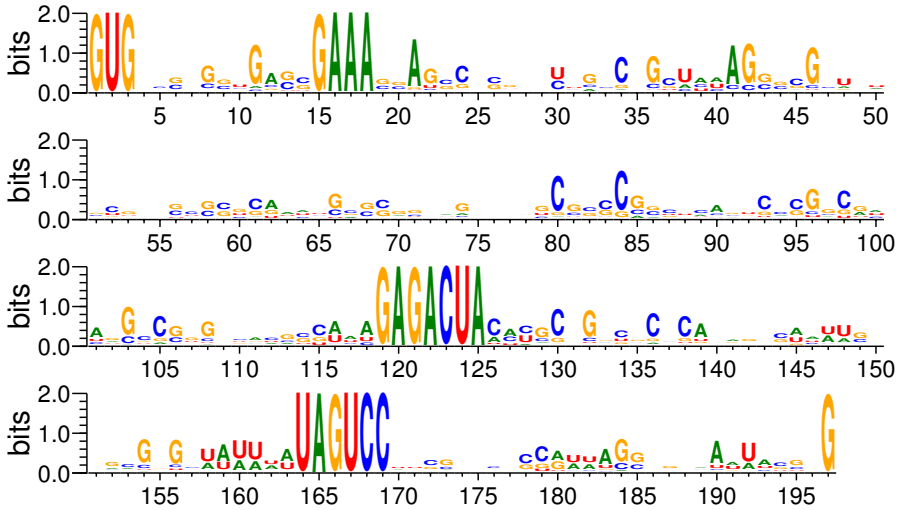
Designs with constrained Structure Prediction and Objective Function. The initially designed sequences of the constraint set `proto` followed the same design approach as the designs examined in this section. However, since the former differed in the reference sequence and target structure from the other constraint sets, they were considered separately in section A.5.

To get an overview of the nucleotide composition of all the designed sequences of a particular constraint set, sequence logo plots were generated (Figure 16), displaying the information content per position [**schneider_information_1986**]. Here, no prior underlying nucleotide distribution was assumed. Expectedly, positions with fixed nucleotides dominate the appearance in these plots.

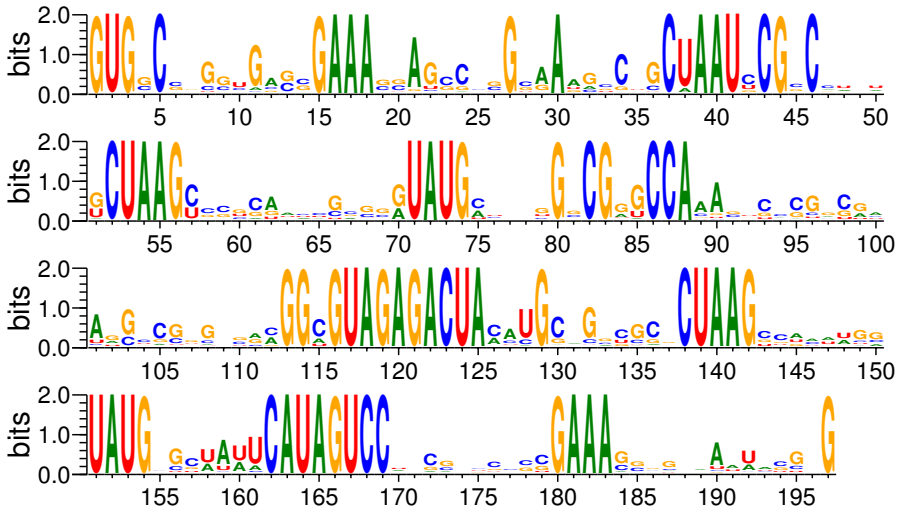
Positions unconstrained in both `minimal` and `complete` tended to be of quite diverse composition, with only a few isolated positions having a single dominating nucleotide identity. Overall, Figure 16a and 16b display largely similar nucleotide compositions where no nucleotide identities were enforced by sequence constraints. Positions that were constrained in `complete` but not in `minimal`, show no obvious overlap in their nucleotide composition in Figure 16, reassuring the orthogonality of

tertiary interactions and secondary structure.

These observations indicate that a diverse range of possible sequences was generated in the design process. The inclusion of sequence constraints to account for tertiary interactions did not create biases interfering with the design process, apart from the fixed positions.



(a) `minimal` constraint set.



(b) `complete` constraint set.

Figure 16: Sequence logos of **(a)** sequences designed with the `minimal` constraint set ($n = 1000$) **(b)** sequences designed with the `complete` constraint set ($n = 1000$). Nucleotides were colored according to their identity. Positions with an information content of 2 bit correspond to the sequence constraints.

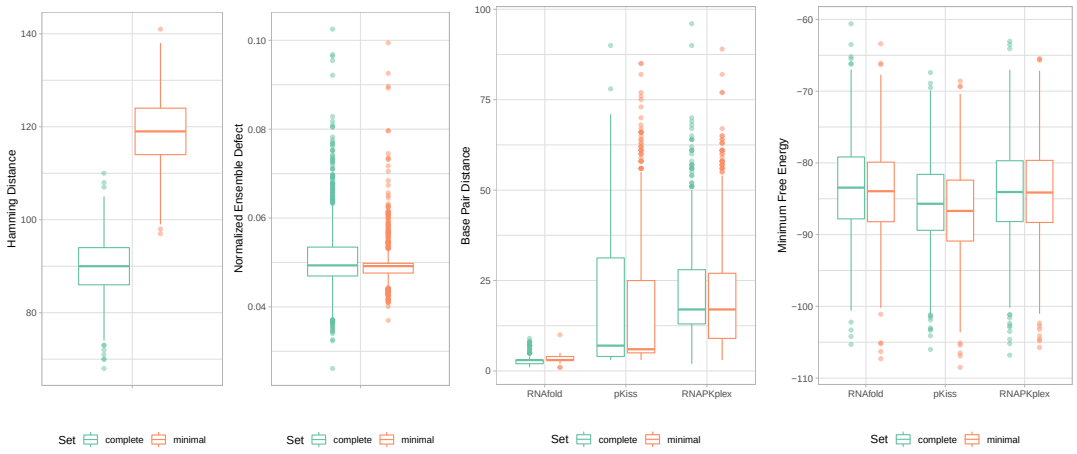
Indeed, the sequence constraints applied in the design process seem to be the primary cause for reduced diversity: The Hamming distance, i.e. the number of point mutations, relative to the native sequence is shown in Figure 17a for both sets of sequence designs.

Considering the total number of fixed positions for each constraint set (see Table 2) and assuming to correctly guess 25 % of nucleotide positions of the native reference sequence by choosing from equiprobable nucleotides, an average Hamming distance of $^3/_4(197 - 67) = 97.5$ for the `complete` set and $^3/_4(197 - 21) = 132$ for the `minimal` set respectively, could be expected.

Despite being a very rough simplification, this is surprisingly close to, albeit still overestimating, the median values of Figure 17a.

Since the minimization of the normalized ensemble defect was stopped at 0.05, the outcome of Figure 17b was expected. However, outliers arise due to the application of additional stop conditions (see section 3.2.3).

As seen in Figure 17c, the predicted minimum free energy structures were generally closer to the target structure than the same predictions for the native sequence (cf. Table 3). Here, the constrained `RNAfold` performed the best, which seems plausible given that its predictions were used in the design process itself. Nevertheless, the median base pair distance of predictions made by `pKiss` with modified pseudoknot penalty was still small and therefore seen as practical validation of the design approach. Structure predictions using `RNAPKplex` were in the same range as `pKiss`, although showing a higher median value.



(a) Hamming distance to the native sequence as a measure of sequence similarity.

(b) the normalized ensemble defect to the pseudoknot-free target structure.

(c) Base pair distances relative to the target structure. For predictions made by `RNAfold`, base pairs of P7 were assumed to be present.

(d) Free Energy of the predicted structures. `RNAfold` values were corrected by the energy of the assumed base pairs in P7 (Figure 13a).

Figure 17: Metrics computed for $n = 1000$ sequences designed using the constraint sets `minimal` and `complete` respectively. Note that computations done by `RNAfold` or `ViennaRNA` were subject to structural constraints as described in section 3.1.2.

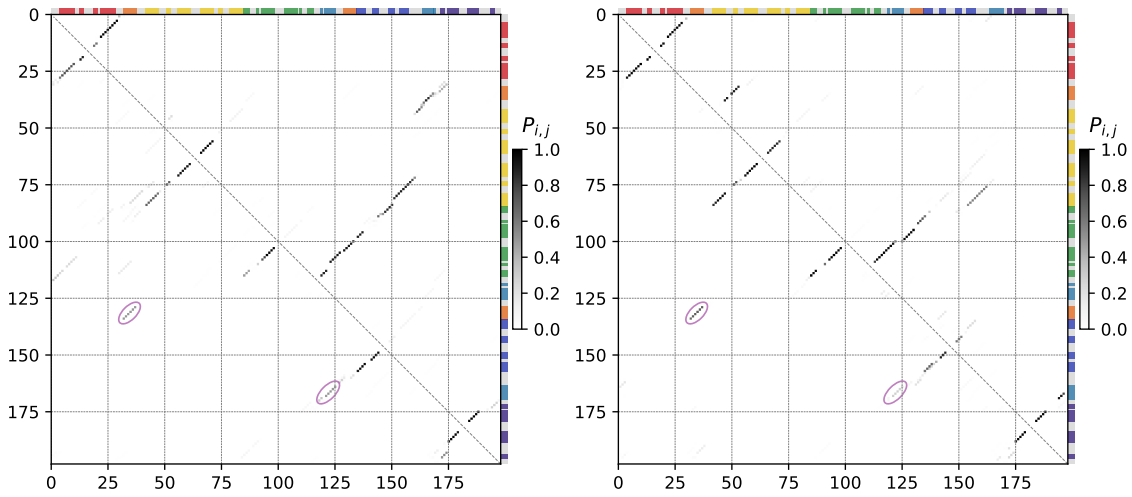
Interestingly, in Figure 17c, the difference between the `complete` and `minimal` set seemed relatively small, which may indicate that, even in the `complete` set, the total number of constraints was still small enough to explore the sequence space

somewhat sufficiently.

Moreover, the free energies of the predicted structures as depicted in Figure 17d were relatively consistent across the three different tools, as well as between both constraint sets. On its own, this seemed surprising, considering that different pseudoknot penalties were applied. Then again, it is consistent with the observation of similarly predicted free energy for the native *Azoarcus* sequence itself as already established in section 4.1.1 (cf. Table 3).

As written in section 3.2.4, the designed sequences were considered robust if the three MFE predictions were similar in the sense that their average base pair distance to the target was small.

This way, one of the best and one of the worst sequence designs were selected, and their base pair probability matrices were computed using the McCaskill algorithm implemented in `ViennaRNA` (Figure 18).



(a) `minimal` constraint set.

(b) `complete` constraint set.

Figure 18: Base pair probabilities of designed sequences, computed using `ViennaRNA`. Here, the normalized ensemble defect was used as the objective function. **(a)** sequence designed with the `minimal` constraint set **(b)** sequence designed with the `complete` constraint set. The lower diagonal half of these figures shows one of the best designs in the constraint sets according to the average base pair distance of the three MFE predictions used; the upper diagonal half shows one of the worst designs according to the same metric. Position indices of $P_{i,j}$ are labelled at both axes. Helices of the native structure are color-coded on top and at the right according to Figure 4. The location of the original P3-P7 pseudoknot is highlighted in purple. See Table A.4 for the exact RNA sequences used here.

As visible in the lower diagonal halves of both Figure 18a and 18b, sequences considered to be good did in fact show patterns in their base pair probability matrices very similar to the target structure (cf. Figure 14a) including positions of the

pseudoknot. More importantly, these figures illustrate an inherent limitation of this approach; the patterns seen in the upper diagonal halves of Figure 18a and 18b do not resemble the target structure in large parts due to the application of structural constraints in the design process. Specifically, the partition function computation was subject to the same constraints as the MFE structure prediction, effectively leaving P7 positions unpaired (see section 3.2). Conversely, the partition functions computed for Figure 18 were not constrained, effectively allowing nucleotides in P7 to interfere with otherwise carefully optimized positions. Hence, the constrained approach may produce sequences with structures similar to the target but does not discern unsound designs.

Designs with unconstrained Structure Prediction and alternative Objective Function. The designs of the sets `complete-alt` and `minimal-alt` differ to the previously analyzed sets in the removal of the sequence constraints at P7 and the use of a different objective function (see Table 2, section 3.2.1).

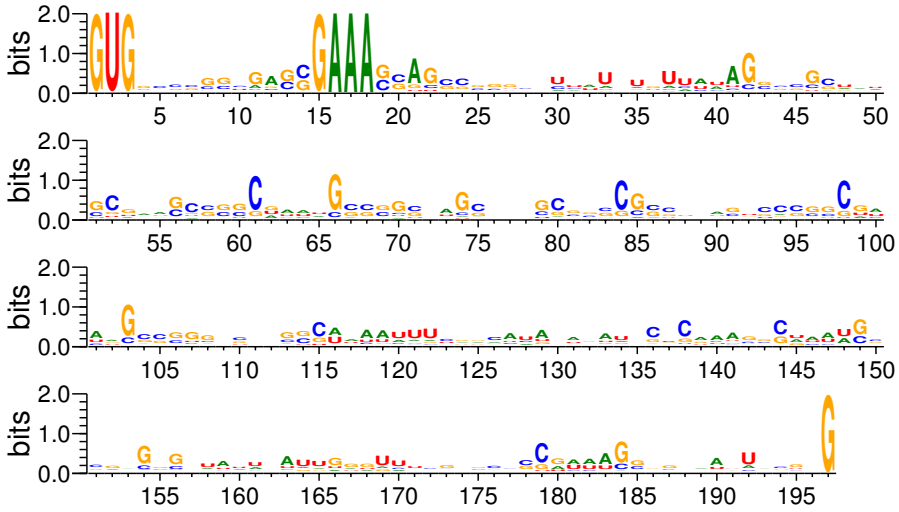
Nevertheless, some observations made with the `complete` and `minimal` set apply here as well. The sequence logos of the constraint sets `complete-alt` and `minimal-alt` show small conservation at unconstrained positions (Figure 19). Similar to Figure 16, both alternative sets of sequence designs show few positions with a single dominating nucleotide identity. Additionally, there is no evidence in Figure 19 that the design process favorably produced the nucleotide identities of P7 in the native reference sequence.

Figure 20a qualitatively resembles Figure 17a with differences of the median Hamming distance values attributable to the omission of the P7 sequence constraints.

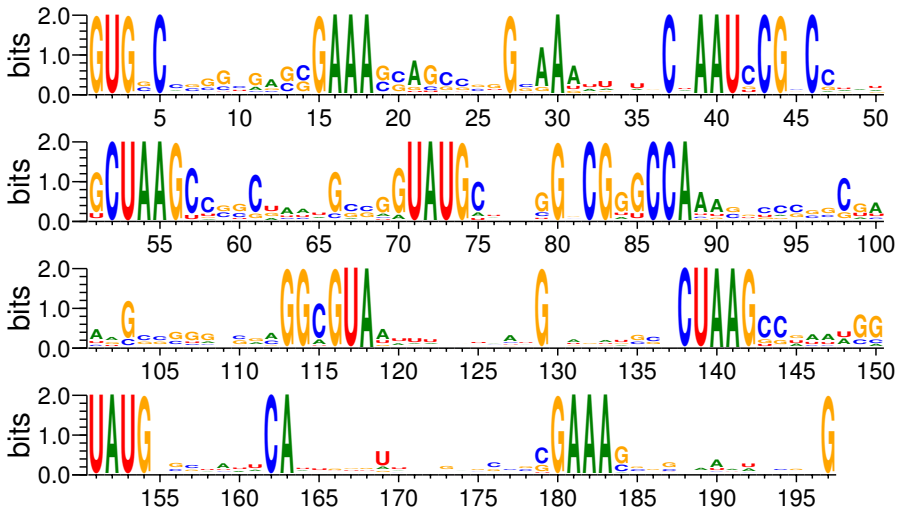
The minimization of the objective function was stopped at a threshold of 0.15 as described in section 3.2.3 (Figure 20b). The threshold was chosen higher than in Figure 17b to account for the inherent conflict of minimizing the normalized ensemble effect given two superposed target structures.

In order to assess the quality of the designed sequences with this approach, the base pair distances relative to the target structure were computed (Figure 20c). In contrast to Figure 17c, no structural constraints were applied here for the structure prediction with `RNAfold` since the relevant sequence constraints were omitted as well. The effect of this choice is immediately visible in 20c, yielding higher values than using the constrained prediction. However, this was not considered an issue since base pair interactions at positions of the targeted pseudoknot were expected.

Using `pKiss` and `RNAPKplex`, structures predicted for sequences using the alter-



(a) `minimal-alt` constraint set.



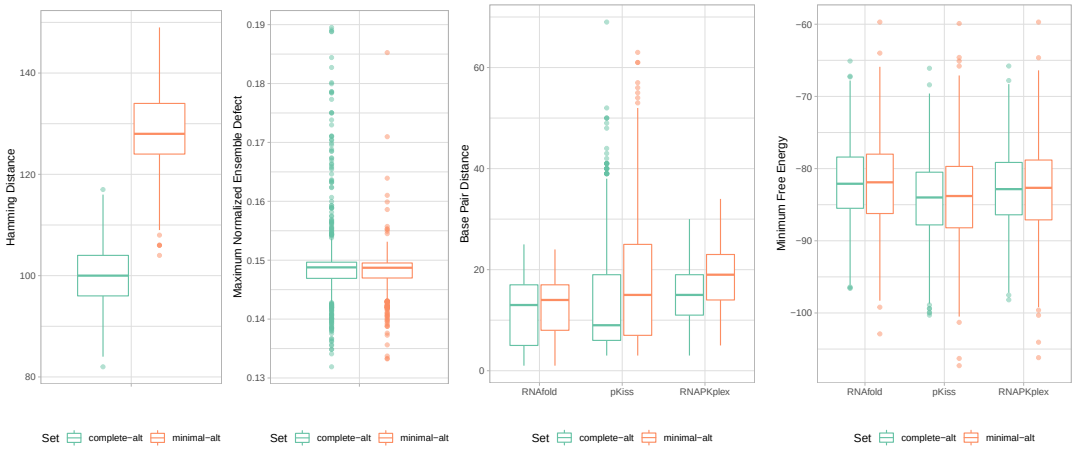
(b) `complete-alt` constraint set.

Figure 19: Sequence logos of **(a)** sequences designed with the `minimal-alt` constraint set ($n = 1000$) **(b)** sequences designed with the `complete-alt` constraint set ($n = 1000$). Nucleotides were colored according to their identity. Positions with an information content of 2 bit correspond to the sequence constraints.

native approach are characterized by a visibly shorter interquartile range than in Figure 17c. Furthermore, the median base pair distance to the target structure is higher for the `minimal-alt` constraint set, probably due to fewer outliers in comparison with Figure 17c.

Still, the free energies predicted by each of the tools (Figure 20d) were quite similar to each other and to the results using the constrained approach (see Figure 17d). Contrary to Figure 17d, the free energy values for `RNAfold` were not corrected in any way since not structural constraints were applied.

Just as before, example sequences were selected according to the average base



(a) Hamming distance to the native sequence as a measure of sequence similarity.

(b) the maximum normalized ensemble defect of two split target structures (see equation 10).

(c) Base pair distances relative to the target structure. For predictions made by `RNAfold`, base pairs of P7 in the target structure were disregarded.

(d) Free Energy of the predicted structures. `RNAfold` values were not subject to any correction (cf. Figure 17d).

Figure 20: Metrics computed for $n = 1000$ sequences designed using the constraint sets `minimal-alt` and `complete-alt` respectively. No structural constraints were applied for these sequence designs.

pair distance of the structures predicted using `RNAfold`, `pKiss` and `RNAPKplex`. Their base pair probability matrices were computed using `ViennaRNA` and are shown in Figure 21. A colored scale was chosen, and probabilities below a threshold of 0.1 were excluded to aid visibility.

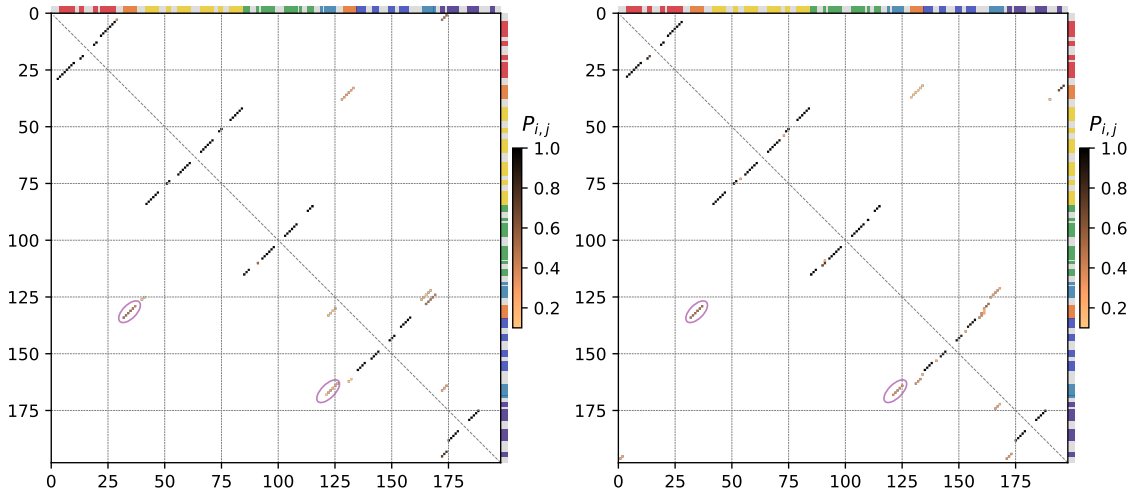
Similar to Figure 18, sequence designs of both `minimal-alt` and `complete-alt` considered good closely resemble the target structure including positions of the pseudoknot (see lower diagonal half of Figure 21).

The upper diagonal half of both Figure 21a and 21b still resemble the target structure in large parts and improve drastically upon the previously used constrained approach. Nevertheless, positions close to the pseudoknot of the target structure seem less well defined in their shape and show lower base pair probabilities than in the lower diagonal half. In contrast to one of the constrained designs (Figure 18a, lower diagonal half), the 1 nt bulge of P7 is not observed here.

4.1.3 Neutral Path Lengths

Since the structure predictions for the designed sequences made by `pKiss` generally produced pseudoknots similar to the reference structure, the following results regarding the lengths of neutral paths were based on those structure predictions.

Starting from each sequence design, a neutral path was computed on the space of sequences compatible to the predicted structure of the sequence design. `RNAblueprint`



(a) `minimal-alt` constraint set.

(b) `complete-alt` constraint set.

Figure 21: Base pair probabilities of designed sequences, computed using `ViennaRNA`. For these designs, $\widehat{\text{ned}}_P(s_1^*, s_2^*)$ was used as the objective function (see equation 10). **(a)** sequence designed with the `minimal-alt` constraint set **(b)** sequence designed with the `complete-alt` constraint set. The lower diagonal half of these figures shows one of the best designs in the constraint sets according to the average base pair distance of the three MFE predictions used; the upper diagonal half shows one of the worst designs according to the same metric. Position indices of $P_{i,j}$ are labelled at both axes. Helices of the native structure are color-coded on top and at the right according to Figure 4. The location of the original P3-P7 pseudoknot is highlighted in purple. The location of the original P3-P7 pseudoknot is highlighted in purple. Base pairings with probability < 0.1 were omitted and a colored scale was chosen to improve visibility. See Table A.4 for the exact RNA sequences used here.

was used to sample uniformly from the compatible neighbor sequences. However, the neutral paths were stopped early if no neutral neighbor with increased Hamming distance to the initial sequence was found after a constant number of trials. This was necessary, as the neighbors were not enumerated exhaustively (or chosen without repetition). The fact that neutral paths provide a lower bound for neutral network sizes was seen as a justification to stop early.

In addition, $n = 1000$ neutral paths starting from the reference sequence of *Azoarcus* were computed as well. The structure predicted by `pKiss` was used again to define the compatible sequence space.

The frequency distribution of neutral paths for the sequence designs of each constraint set and the paths starting from *Azoarcus* is displayed in Figure 22.

Overall, the distributions seem to be very similar. The distribution for `Azoarcus` appears to be shifted slightly to the left, which is probably an artifact resulting from stopping the neutral paths early. Still, the mean length of the computed neutral

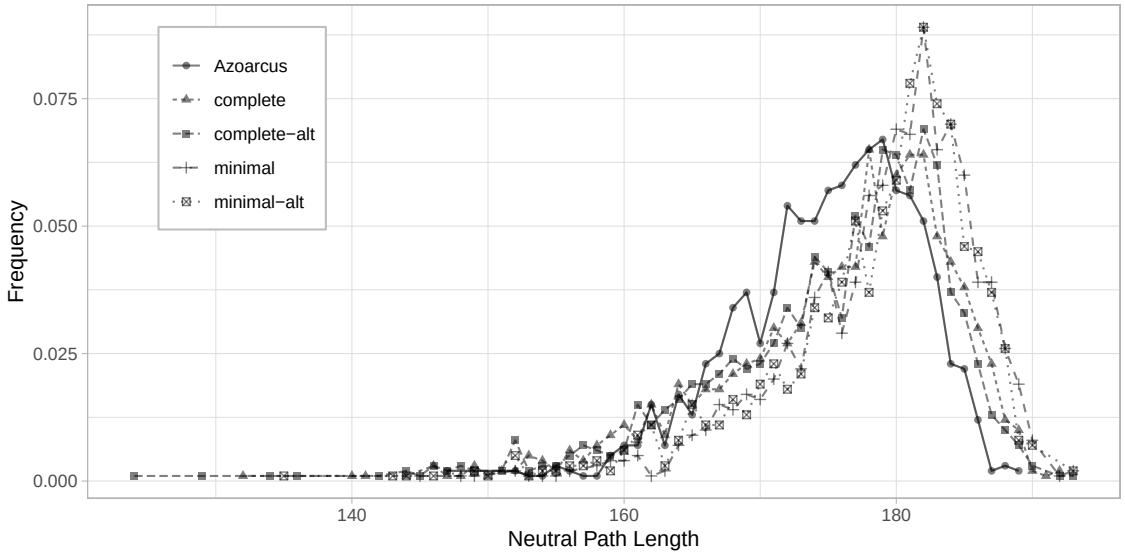


Figure 22: Lengths of neutral paths, each starting from a designed sequence. The paths were generated on the sequence spaces compatible to the `pKiss`-predicted structure of the initial sequence. Neutral paths of the same length were counted and normalized by the number of designs per constraint set. See Table 4 for the mean lengths of each constraint set. The generated paths starting from *Azoarcus* were plotted slightly darker.

paths starting from *Azoarcus* is very close to and within the standard deviation of the mean lengths computed for `complete` and `complete-alt` (Table 4). Although slightly higher, the same statement applies to the mean lengths of `complete` and `complete-alt`.

Table 4: Mean neutral path lengths of designed sequences and the expected Hamming distances between two random sequences compatible to the `pKiss`-predicted structure for each designed sequence per constraint set. The displayed values are mean values for each set including standard deviation. The expected Hamming distance of two random compatible sequences was estimated for each neutral structure as described in section 3.3 (see equation 11). For *Azoarcus*, the reference sequence was used as initial sequence and 1000 neutral paths were generated on the sequence space compatible to the structure predicted by `pKiss`.

Set ($n = 1000$)	Neutral Path Length	Expected Hamming Distance
`minimal`	178.90 ± 6.90	144.47 ± 0.13
`complete`	175.66 ± 8.67	144.50 ± 0.12
`minimal-alt`	178.10 ± 7.83	144.52 ± 0.17
`complete-alt`	175.13 ± 9.05	144.52 ± 0.15
Azoarcus	174.97 ± 6.55	144.36 ± 0.00

The mean expected Hamming distance of two random sequences compatible to the same structure is almost equal in the spaces of compatible sequences used for neutral path computation in each set (Table 4). This observation is due to the

overall design goal to produce structures similar to the *Azoarcus* group I intron and the dependence of the expected Hamming distance on the number of unpaired positions and base pairs (see equation 11). Naturally, the standard deviation of the expected Hamming distance is equal to zero given the same predicted structure in all neutral paths for *Azoarcus*. It may be notable that in the sequence space compatible to the native structure (see Figure 7b), the expected distance between two random sequences is approximately 144.47 ($u = 79$, $b = 59$).

Recalling section 3.3, it should be emphasized that the expected Hamming distance between to random compatible sequence is not a threshold for sequence space percolation but used to put the computed neutral length paths into perspective. Furthermore, since $^{13}/_{18} \approx {}^{3}/_{4}$, equation 11 is well approximated by the expected Hamming distance of two random sequences not necessarily compatible to a shared structure: ${}^{3}/_{4}\,N$ [**haslinger_rna_1999**].

In summary, neutral paths of designed sequences do not reach maximum length but are still very long, indicating extended neutral networks of similar structures. This is encouraging since neutral network size correlates with mutational robustness [**jorg_neutral_2008**] although this should be interpreted with care. The lower bound provided by neutral path lengths does not correspond to the number of sequences folding into a structure but rather how far-reaching the neutral network extends through sequence space. If a designed RNA sequence proved to be functional, this would be a desirable property similar to results in the *Azoargus* group I intron [**hayden_intramolecular_2015**].

It has been previously shown that percolating neutral networks exist in spaces mapping RNA sequences (of length 100 nt) to crossing secondary structures [**haslinger_rna_199**
The results described here are aligned very well with the data in [**haslinger_rna_1999**], even though a different mapping between sequences and structures was employed by using `pKiss`. However, the generated neutral paths are not neutral to the actual target structure (section 2.4).

4.2 Preliminary Experimental Results

To test the designed sequences, an assay based on the self-splicing mechanism of the *Azoarcus* group I intron (see section 2.2) was developed by the partner group in Paris, France. Currently, preliminary results are available only for a selection of designed sequences of the `proto` set. These sequences had to be post-processed to fit experimental requirements, i.e. they had to be truncated, and the first three 5'-nucleotides were replaced with the native IGS, and the 3'-terminal nucleotide was set to `G`.

The results of a negative control experiment with known non-catalytically active RNAs are not yet available. Therefore, the following paragraphs do not represent the latest experimental results.

4.2.1 Mechanism of the Self-Splicing Assay

The goal of the assay was to screen a small pool of designed synthetic introns for self-splicing and ligating activity at their 3' end.

Therefore, a synthetic 3'-exon was appended to the potential synthetic catalysts. A short RNA substrate with the trinucleotide `CAU` at its 3' end was added to bind to the introns IGS. This setup corresponds to the *pre-2S* state of the self-splicing mechanism in the native intron (see Figure 5). Catalytically active synthetic introns were expected to be able to splice the synthetic exon and ligate it to the substrate by the same mechanism.

A second RNA substrate with a 5'-oligonucleotide `GGCAU` containing the IGS recognition motif was used for the reverse reaction. Since the 3'-terminal nucleotide of the synthetic introns was `G`, this nucleotide would be expected to bind to the guanosine binding site and splice the substrate from its 5'-oligonucleotide, followed by ligation of the substrate to the 3' end in catalytically active introns (*post-2S* in Figure 5).

In total, three different states were expected for catalysts; the intron with the synthetic 3'-exon, the intron without any exons attached, and the intron with the second substrate substituting the 3'-exon. However, the presence of the third state would be sufficient to indicate the success of the assay.

4.2.2 Assay Results

Template DNA of the designed sequences with a synthetic exon added to the 3' end and a promoter sequence was ordered, amplified and purified by polymerase chain reaction (PCR) and transcribed into RNA by the partner group in Paris.

The designed RNA sequences were incubated in small pools of multiple sequences with the two substrates added to conduct the assay. Afterwards, a reverse transcription (RT-PCR) was initiated with a primer complementary to the 3' end of the second substrate. That way, only introns with this substrate would be reversely transcribed into DNA. So far, except for a control run using the native *Azoarcus* intron, no activity was detected. Therefore, the reversely transcribed DNA was amplified to increase a potential signal. After the amplification, bands corresponding to sequences roughly 200 nt long were observed for the designed synthetic introns as well as the native intron (Figure 23).

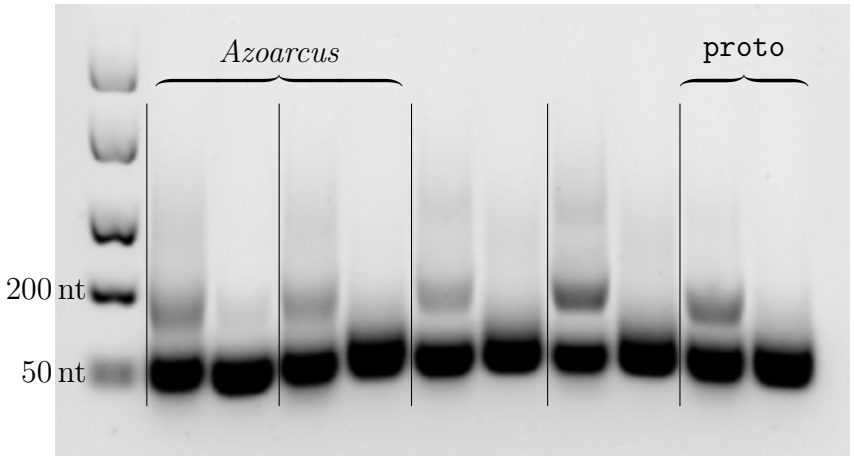


Figure 23: Gel obtained by gel-electrophoresis of reversely transcribed and amplified DNA from the assayed introns. Vertical lines separate pools of different design methods. Unlabelled pools originate from methods not part of this thesis. Lanes labelled *Azoarcus* contain DNA corresponding to the native intron as a control. The right lane of each pool was obtained from assay runs without any substrate added. Camille Lambert generously provided the gel (Laboratoire de Biochimie, ESPCI Paris, France).

Due to the selective RT-PCR, this band should correspond to catalytically active synthetic introns from at least one of the designed sequences in the pool.

The dominant bands at around 50 nt correspond to non-specific DNA since the RT-PCR was highly sensitive. These bands were also visible in control runs without any substrate added.

5 Discussion

Ultimately, the catalytic activity of the *Azoarcus* group I intron determined the focus of the overall design on the pseudoknot and specifically the P7 region containing the guanosine binding site. Therefore, two different approaches were pursued.

The first one employed structural and sequence constraints to keep the P7 region intact, motivated by the binding site relevant for catalytic activity that P7 contains, and the strong conservation of the region among group I introns.

The second, alternative approach started as a modification of the first one. The core idea was inspired by an observation made by Gaspin and Westhof and later applied to the *Tetrahymena* group I intron by Mathews; base pair probabilities obtained using the McCaskill algorithm indicate possible pseudoknots [**gaspin_interactive_1995**, **mathews_using_2004**]. Using the normalized ensemble defect as the primary objective function required computation of the partition function, so no computational cost was added apart from using computed base pair probabilities twice.

Notably, no explicit modelling of pseudoknots was necessary for the implementation of both design approaches. Naturally, it should be questioned how well these approaches suited the primary goal of generating RNA sequences with pseudoknots similar to the one present in the target structure.

In attempting to address this question from a technical perspective, the challenges and limitations of the design approaches implemented in this work are discussed in section 5.1. In that context, the desired catalytic activity of the designed RNA sequences in wet-lab experiments should not be ignored, and core assumptions are revisited. Finally, improvements to the design pipeline are proposed, and various other approaches are suggested (section 5.2).

5.1 Design Challenges and Limitations

A central problem in the design stems from the prediction of pseudoknots. Although both design approaches did not directly depend on explicit pseudoknot models, designed sequences still had to be evaluated with the desired pseudoknot in mind. Hence, the prediction accuracy of the used tooling with respect to the *Azoarcus* group I intron largely determined the reliability of the designs. Furthermore, the thermodynamical energy parameters used in MFE prediction were modified as little as possible to keep predicted free energies interpretable.

Although limited by these considerations, the heuristics of `pKiss` could successfully recover large parts of the target pseudoknot with two caveats; firstly, the 1 nt bulge in P7 was not modelled and secondly, region P6 was not well predicted.

Consequently, the default energy parameters were interpreted as generally reliable, and the missing bulge was addressed by applying sequence constraints with the assumption of it being a structural feature primarily determined by local interactions. Inaccuracies in the prediction of the P4-P6 domain were seen as non-critical because this scaffold domain is not strictly necessary for catalytic activity [**hayden_intramolecular_2015**].

Recognizing these decisions, all four constraint sets yielded good designs in terms of base pair distance to the target structure. However, the similarity to the target according to this metric varied considerably across sequences of different constraint sets, necessitating quality control of design candidates.

Between the two approaches pursued, some noteworthy differences should be mentioned. The constrained approach depends on sequence constraints matching the structural constraints. In the case of this work, this worked well because of the conserved P7 region. Consequently, there is manual adjustment needed in order to apply the design pipeline to other targets.

The alternative approach does not require structural constraints imposed on structure prediction, mitigating the need for manual adjustment. Still, with this approach, applying sequence constraints corresponding to P7 might be beneficial as the nucleotides of P7 are conserved among group I introns and the region itself is relevant to the catalytic activity of the native ribozyme. The decision to not impose such constraints using the alternative approach was made to demonstrate increased versatility.

Besides positions in P7, sequence constraints were also used to preserve nucleotide identities at positions involved in tertiary interactions, with the primary effect of reducing complexity during sequence design. Nucleotides involved in those interactions are paired less specifically than canonical base pairs. It is questionable that preserving nucleotide identities, as attempted herein, also preserves these interactions. Initially, this strategy was motivated by the influence of the secondary structure on tertiary interactions according to Mustoe [**mustoe_secondary_2016**].

Strictly speaking, modelling tertiary interactions apart from pseudoknots was beyond the scope of this thesis. Having said that, tertiary interactions are quintessential to RNA structure as secondary structure does not define three-dimensional conformations uniquely; more attention should be paid to tertiary interactions regarding their influence on catalytic activity. In group I introns, they are significantly involved in joining scaffold and catalytic domain [**tanner_joining_1997**]. On top of that, there is evidence of tertiary interactions stabilizing substrate binding in the *Azoarcus* group I intron and RNA folding in general [**gleitsman_kinetic_2014**,

chauhan_tertiary_2008].

5.1.1 Revisiting Assumptions

Prediction of a single most stable (MFE) structure is undoubtedly a very simplistic foundation for the design of catalytically active RNAs. Catalysts like the *Azoarcus* group I intron are not static molecules. In fact, self-splicing of this molecule involves changes of its structural conformation [**adams_crystal_2004-1**, **gleitsman_kinetic_2014**]. Therefore a suitable kinetic model of the desired catalytic function could be of value.

The insufficiency of a single most stable structure prediction was partly addressed by utilizing the partition function. However, in the pursued alternative approach, diversity of the structure ensemble was reduced as a side effect. Despite secondary structure guiding tertiary structure in naturally occurring RNAs, designing for secondary structure alone should not be assumed sufficient to achieve a function similar to the reference ribozyme.

A solution to the problem of finding a suitable design objective beyond secondary structure and relating this objective to an experimentally verifiable function is not directly apparent.

5.2 Methodological Considerations

Following the previous section, many possible changes could be applied to the design pipeline. For example, a structure predicted for the reference sequence could be used as a target structure and selected sequence constraints could be reconsidered. Moreover, other characteristics like the free energy of predicted structures or ensemble diversities obtained from the partition function could be incorporated as part of the objective function. These are rather general modifications that do not reach beyond the conceptual limits of the model.

Instead, let us consider specific changes motivated by four central building blocks of the design pipeline.

First, sequence constraints were used to naïvely account for tertiary interactions. Similar to how pseudoknots are a violation of rule (iii) (see section 1) allowing unconveniently many structures, tertiary interactions violate rules (i) and (iv). Allowing structures with nucleotides involved in multiple, possibly noncanonical pairs seems as unattractive as arbitrary kinds of pseudoknots. For that reason, it might be more feasible to impose some restrictions and model only some tertiary interactions, similar to the approach of `RNAwolf`, modelling nucleotides forming up to two potentially

noncanonical pairs [**zu_siederdissen_folding_2011**].

Secondly, neutral networks percolating the sequence space are not rare even for secondary structures with pseudoknots. The sequences designed in this work are not some rare examples with confined neutral networks. Inspired by [**reidys_generic_1997**], random drift could be introduced in the design pipeline by allowing moves between neutral neighbors. Therefore, the parameter p_{acc} could be repurposed or even omitted to allow neutral moves (see section 3.2.3).

Thirdly, sampling sequences compatible to multiple target structures was facilitated by `RNAblueprint`. Although used to handle a pseudoknot in this work, this library allows sampling sequences subject to more complex structural constraints. It has been successfully used in RNA design with multiple conformational states [**findeis_silico_2018**]. With the conformational change of the *Azoarcus* group I intron during self-splicing in mind, a similar approach should be explored (cf. section 2.2).

Finally, the efficient partition function and base pair probability computation via McCaskill's algorithm could replace MFE prediction directly. Nevertheless, computing a partition function for ensembles of nested structures to detect pseudoknots has limitations. Detecting more complex or multiple pseudoknots using dot plots of base pair probabilities could be problematic because base pair probabilities at corresponding positions might be indistinguishable from alternative nested conformations. In fact, this problem translates to the extension of the maximum normalized ensemble defect to multiple superpositioned target structures (see equation 10). Minimizing this objective function may be interpreted as concurrent minimization of normalized ensemble defects for multiple conflicting nested target structures. Fortunately, the ensemble defect relies indirectly on the ensemble of possible structures via the base pair probabilities obtained from the partition function. Using crossing structures as input for the ensemble defect is merely an implementation detail. Of course, this is usually questionable, but it makes sense in the case of this work. For this reason, I extended the ensemble defect implementation in `ViennaRNA`. Those changes were accepted and are available in the `ViennaRNA` package as of version `2.4.18` (see section A.1).

A more generalized version of the ensemble defect could be used in order to compare structure ensembles of sequences [**dirks_paradigms_2004**]:

$$\mathrm{d}(P, P') = N - \sum_{i,j} P_{i,j} P'_{i,j} \tag{12}$$

where P and P' are base pair probability matrices of two sequences. The distance

in equation 12 may be seen as an estimation of the metric induced by the Frobenius norm (see section A.6). Then again, this metric is computationally more expensive than the ensemble defect.

Early on in this work, using a partition function algorithm explicitly modelling some types of pseudoknots implemented in `NUPACK` was disregarded. However, there is a design approach implemented in `ENZYMER` similar to the alternative approach of this work; it primarily consists of an adaptive walk — with varying step sizes — minimizing the ensemble defect, which was computed using `NUPACK` [**zandi_adaptive_2016**].

5.2.1 Outlook on Other Approaches

The design approaches in this work followed a *de novo* pattern by starting from an arbitrary sequence and exploring the sequence space. A straightforward variation of the herein implemented design pipeline would utilize the reference (or hypothetical promising sequence designs) as the initial sequence. Similarly, the neutral network of the reference sequence could be explored. Generally, these variations would not address the limitations outlined in section 5.1.

Extending on multi-state RNA design, briefly hinted at in section 5.2, incorporating RNA folding kinetics in a design pipeline should be worthwhile. Not only does RNA fold hierarchically, i.e. from secondary to tertiary structure, but also sequentially via intermediate conformational (secondary structure) states [**tinoco_how_1999**]. Indeed, group I introns like the one in *Azoarcus* fold into intermediate structures along conserved pathways [**mitra_rna_2011**]. While there are computational models of RNA folding kinetics available [**kucharik_basin_2014**, **kucharik_pseudoknots_20**
modelling kinetics for every design candidate might not be feasible. Instead, intermediate conformations computed for the reference could be used to define multiple target structures and energies.

In order to enable more complex approaches, reducing the size of the reference molecule might be in order. In the case of the *Azoarcus* group I intron, removing the scaffold domain does not entirely prohibit the catalytic activity of the remaining molecule, enabling the construction of a smaller ribozyme. With that, some new complications might arise because the scaffold domain contributes significantly to the stability of the *Azoarcus* group I intron and increases mutational robustness [**hayden_intramolecular_2015**]. Still, the construction of a minimal synthetic intron is a way to reduce computational complexity but also requires experimental verification of catalytic activity beforehand. Similarly, designing only fragments with the goal of self-assembly might reduce complexity while simultaneously complicating

experimental verification.

Another direction could be taken by utilizing phylogenetic data, e.g. from GISSD [**zhou__gissd__2008**]. In fact, the unlabelled lanes in Figure 23 represent sequences designed by Vaitea Opuu based on *Azoarcus* group I intron homologs. One of his approaches utilized *Direct-Coupling Analysis* (DCA). The core idea of DCA is to model the probability of a sequence parametrized by coupled nucleotides and position-specific biases for specific nucleotides [**de__leonardis__direct-coupling__2015**]. With parameters derived from multiple sequence alignments of homologous sequences, DCA can be used for structure prediction, including tertiary interactions, and sequence design by maximizing probability according to the model.

Yet, all of these suggestions represent computational methods, and their validity strongly depends on experimental verification.

5.3 Conclusion

Within limits of the underlying model and pseudoknot heuristics, designing secondary structures including simple pseudoknots works surprisingly well by utilizing methods developed for minimum free energy prediction of nested structures in conjunction with sequence and structure constraints. Even further, by diverting McCaskill's partition function algorithm from its intended use with nested secondary structures, structural constraints may be omitted. However, the hope for success of the designed sequences in wet-lab experiments is questionable. The diversity of the designed sequences indicates that the solution space is still vast. Relating secondary structure objective functions to *in vitro* experiments testing catalysis is challenging. But there is potential in evolving sequences designed for secondary structure *in vitro* as indicated by RNA shape space covering, i.e. most common secondary structures being reached from random sequences by just a few mutations [**schuster__sequences__1994**, **reidys__generic__1997**], and experimental evidence showing improvements of catalytic activity by *in vitro* evolution [**ameta__next-generatio**

Shifting from a purely thermodynamical view on secondary structures to kinetic models and modelling tertiary interactions seems necessary to design functional ribozymes similar to the *Azoarcus* group I intron.

A Appendix

A.1 Code and Data Availability

Source Code. `pKiss` (`https://github.com/jlab/fold-grammars`) was used at version `2.2.12`. The `ViennaRNA` package and accompanying tools were used at version `2.4.17`, except `RNAPKplex`, for which the exact changes are available as shown in Table A.1. The extension of the ensemble defect implementation discussed in section 5.2 is also listed in Table A.1.

Table A.1: URLs to contributions made to `ViennaRNA` during work on this thesis.

`RNAPKplex`	`https://github.com/ViennaRNA/ViennaRNA/compare/77861405002d93a35cec3f615e2d1a5d210964d8...7a7e84ae8f6954dff43cc31d581b0bcc63b8a1e1`
Ensemble Defect	`https://github.com/ViennaRNA/ViennaRNA/compare/0f6e876ab4e80aaf5bf2c5f678ba9a7f4c840849...67445f01d690f661138e67e43d2139fdd26dfba0`

The low-level bindings for the `ViennaRNA` library written in Rust during work on this thesis are available at `https://github.com/fncnt/librna-sys`, including a base pair distance implementation extended to pseudoknots (cf. section 3.2.1). Code implementing the design pipeline, including auxiliary scripts, is both appended digitally (on CD) and made available at `https://github.com/fncnt/azodesign`.

Data generated for this work is appended digitally (on CD).

A.2 Superpositioned Nested Target Structures

Table A.2 contains the split target structures used for the alternative design approach in dot-bracket notation (see section 3.2).

Table A.2: Structures in dot-bracket notation obtained from the native structure by removing either one of the stacks P3 or P7 forming the pseudoknot (section 3.2). The first structure is the native structure for comparison (see Figure 7b).

```
>native structure (Figure 7b)
...(((((((..((....)).)))))))...((((((....((((((...((...((((((....))))))..))...)))))))(((
...(.((((((....))))))..)..)))...[.[[[[[..)))))))((((...(((....)))..))))......]]]]]]..((.
(((((....)))))).....))..
>native structure without P7 pairs
...(((((((..((....)).)))))))...((((((....((((((...((...((((((....))))))..))...)))))))(((
...(.((((((....))))))..)..))).............))))))((((...(((....)))..)))).............((.
(((((....)))))).....))..
>native structure without P3 pairs
...(((((((..((....)).)))))))..............((((((...((...((((((....))))))..))...)))))))(((
...(.((((((....))))))..)..)))...(.(((((.........((((...(((....)))..))))......))))))..((.
(((((....)))))).....))..
```

A.3 MFE Predictions for the Reference Sequence

Table A.3 contains structures in dot-bracket notation, predicted using different MFE algorithms (see Table 3).

Table A.3: Structures in dot-bracket notation as predicted by different MFE prediction tools as summarized in Table 3. The first structure is the native structure for comparison (see Figure 7b).

```
>native structure (Figure 7b)
...(((((((..((....)).)))))))...((((((....((((((...((...((((((....))))))..))...)))))))(((...(.(((((
(....))))))..)..)))...[.[[[[[...)))))))((((...(((....)))..))))......]]]]]]..((.(((((....)))))).....)
)..
> RNAfold
((.(((((((..((....)).)))))))..))(((.......((((((....(((.((((((....))))))..)))...))))))(((.......)))
..(((((((......)))))))...((.(((.((...((((...(((....)))..))))...))...)))..))...(((((....)))))....))
)..
> RNAfold -C # with P7 positions constrained to being unpaired
((.(((((((..((....)).)))))))..))(((.......((((((....(((.((((((....))))))..)))...))))))(((...(.(((((
(....))))))..)..(((((............)))))(((...(((....)))..))))..)))............(((((....)))))....))
)..
> RNAPKplex
((.(((((((..((....)).)))))))..))(((.......((((((....(((.((((((....))))))..)))...))))))(((.......)))
..(((((((......)))))))...((.(((.((...((((...(((....)))..))))...))...)))..))...(((((....)))))....))
)..
> RNAPKplex -e 0.0
((.(((((((..((....)).)))))))..))(((.......((((((....(((.((((((....))))))..)))...))))))(((..[[...)))
]](((((((......)))))))...((.(((.((...((((...(((....)))..))))...))...)))..))...(((((....)))))....))
)..
> pKiss --mode mfe --strategy P
((.(((((((..((....)).)))))))..))............[[[[........]]]]...[[[[..]]]].....(((((((......)))))))
.[[[[[....((((((...(((....)))..))))..))..]]]]]......(((((....)))))......
> pKiss --mode mfe --strategy P --Hpenalty 9.8
((.(((((((..((....)).)))))))..))[[[[[[....((((((....(((.((((((....))))))..)))...))))))(((.......)))
..(((((((......)))))))....]]]]]](((((...(((....)))..))))........((..(((((....)))))).....))..
```

A.4 Sequence Design Examples

The exact RNA sequences used to produce dot plots for examples of each constraint set can be found in Table A.4.

Table A.4: Sequences in FASTA format used in dot plots of base pair probability matrices. The locations of the corresponding figures are noted in the header lines.

```
>Figure 18a (lower half)
GUGUCUUCGCCCGGGAAACCAGUGAAGAACCGUCGGCGACAGGGCGGUAAGCAGACUGCCCAUCUGGGCAGGAGCCGGUCGCCCGCU
UGACUCCGGCCAAACGGCCGGCGAUGGCACAGAGACUAAAAGCCGACACUCGACGUGAAAUCACGAGAGUAUUUUAUAGUCCUUCGU
CCACUAAUAAGUGGUUUAGCGUG
>Figure 18a (upper half)
GUGGCCCCUUGCGUGAAAAUAGAGGGGCCCUCUCGGCUAAACAGCGCUACUCGAACUGCAGAUAUCUGCAGAAGACCCGCGCUGGAG
UCAUGCACGGCAUCGGCCGUGGAACCUCAGUGAGACUACGUGCCGAGGCCCAUAGGGAAUUCCCGCGGGUUUUUGUUAGUCCUGGGA
AGGGCGUAAGCCCUAAGUACCCG
>Figure 18b (lower half)
GUGACUCGGUGACGGAAACGAGCCGAGUAAAACCAGCUAAUGCGACAGAAUCUAAGCGUUCAUCAGAGCGUAUGACGGUGUCGCGCC
AAACUCGAGGCAGAGGCCUCGCGAAGGCGUAGAGACUAAAUGCUGGUGCCCUAAGCAUAUUUGUAUGGGUUAUCCAUAGUCCCCCGG
UCCACGAAAGUGGAGAGUUCGAG
>Figure 18b (upper half)
GUGGCAGGGUGAGCGAAAGUAGCCCUGCGAGAGCCGCUAAUACGUCAGGGGCUAAGCUCGCAGUAGCGAGUAUGCGGAUGGCGUUCC
ACAGCGAGCAGAUUUCUGCUCUCAAGGAGUAGAGACUAUCUGCGGCUGCGCUAAGCCAAUAGGUAUGCGCUAUUCAUAGUCCUUCCC
UCCCGGAAACGGGAUAUUUGGAG
```

```
>Figure 21a (lower half)
GUGCAGCGUGGAGCGAAAGCACACGCUGCUAGCCUUUCAUAGCCUCGGUUCCUUUGCACGCGCUAGCGUGUUUGGUUACGAGGCCAG
UCAGUCACUCUGUAGAGAGUGUCAUCUGUAUAUUUCCUAAAAAGGGCGCUCAUGGCGACUUCGCCCGAGCUUAUCUAGGAGUUUCGC
CGGAUUUAUAUCCGUAUUGCGUG
>Figure 21a (upper half)
GUGGGAUCAGAACGGAAACGGCUGAUCCUCUUUUUGCUCUAGGUCGUUGGCUGAGCCGCCCAAAUGGGCGGGAAGAAUACGACCGCC
UCUAUACGGUCGAAAGACCGUUUCCGGCUAAAUUUCACUCAGCAAAGCCUCAAACGCAAACGCGGAGAGGCGGAUAGUGGGUUACAC
GCGAGGGUACUCGCAAAAUUGUG
>Figure 21b (lower half)
GUGGCCGGGUGACGGAAAUGAGCCCGGCAAAACCAACAAAUCCGGCUGAAGCUAAGCCGCCUAAUGGCGGUAUGCGAAAGCCGGGCC
AAAGACCCGCCGAAAGGCGGGCCUAGGCGUAUACUCGAGACGUUGGUCCGCUAAGGCAAAUGCUAUGCGGUAUCCAUUGAGAAGCUC
GCCGUGAAAACGGUCUAAAAGCG
>Figure 21b (upper half)
GUGGCCGGGCGGCCGAAAGGAGCCCGGCAAAGCUUUCGAAUCCGUCCAUAGCUAAGCAGCCUAGUGGCUGUAUGCCCCGGACGGGCC
AAACGUACGGGAUACCCCGUGGGGCGGCGUAUAAAAUUAGUGAAGGUGCCCUAAGUCAAUUGAUAUGGGCUACUCAAAUUUAAAUUC
CCUCCGAAAGGAGGCCCUUAGCG
>Figure A.3 (lower half)
CAACUUAAGCAGGACGCGGGAAACCAGUCCUGCAAACCACCCAAAUCCGCCCAAGGCUAAGGCCUCGGUAGAGGCUAUGUAACGGGC
GGGCCAAAGGGCCCUCGCUGAGUUAGGUGCGAGGGUGCAAGGCGUAGAGACUAAGAGGGUGGGAGCUAAGGCAAGAGCUAUGCUCAA
UUCAUAGUCCCGAACACACCGAAAGGUGUCAACAUCGA
>Figure A.3 (upper half)
GAACGGGGGCAUCUUGCCGGAAACGAGAGAUGCGAAGGACCCAAAUCCGGCCAAAGCUAAGCCUCCGAAGGGAGGUAUGCGAAGGCC
GGGCCAUUGCGUGGGCGCUCCUUUUUUUGCGCCCACACUCGGCGUAGAGACUACUCGGGUCCGGCCUAAGGUGUAAACUAUGGCCAU
AACAUAGUCCAAUAAGUGGUGAAAGCCACAAUUAAUGA
```

A.5 Initial Sequence Designs

The figures in this section were included for completeness. The approach used and the overall performance corresponds to designed sequences of the `complete` set (cf. Figure 16b and 17), with the exception that the data from Figure 7a was used as reference instead.

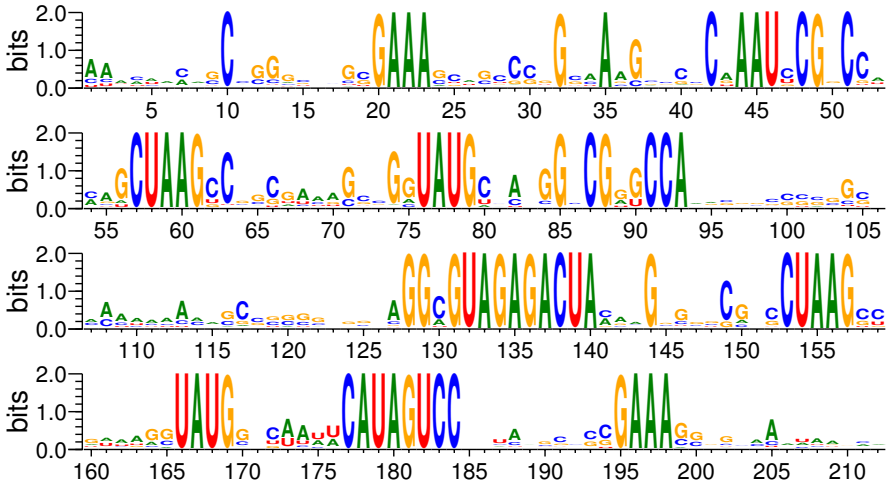
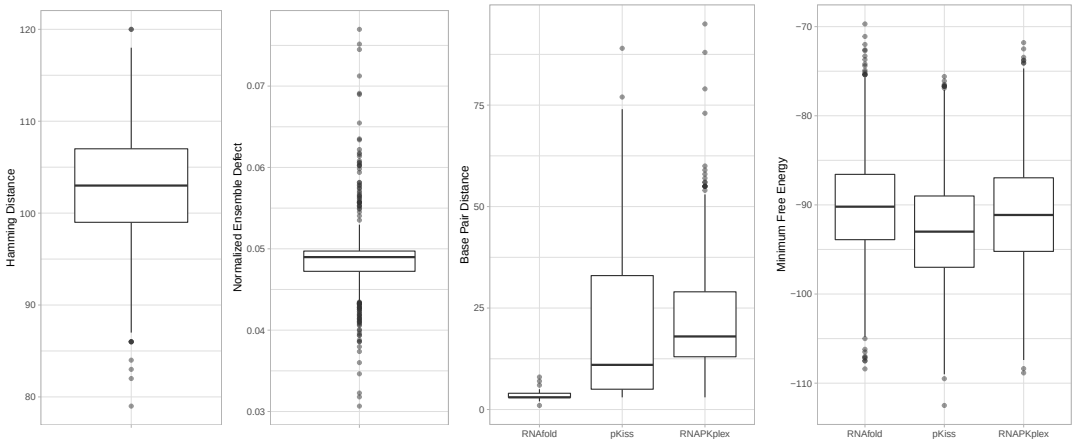


Figure A.1: Sequence logo of sequences designed with the `proto` constraint set ($n = 1000$) Nucleotides were colored according to their identity. Positions with an information content of 2 bit correspond to the sequence constraints.



(a) Hamming distance to the native sequence as a measure of sequence similarity.

(b) the normalized ensemble defect to the pseudoknot-free target structure.

(c) Base pair distances relative to the target structure. For predictions made by `RNAfold`, base pairs of P7 were assumed to be present.

(d) Free Energy of the predicted structures. `RNAfold` values were corrected by the energy of the assumed base pairs in P7 (Figure 13a).

Figure A.2: Metrics computed for $n = 1000$ sequences designed using the constraint sets `proto`. Note that computations done by `RNAfold` or `ViennaRNA` were subject to structural constraints as described in section 3.1.2. In contrast to Figure 17, the target structure and native reference sequence as seen in Figure 7a was used.

Additionally, these designs did produce base pair probability dot plots with positions corresponding to P7 missing (Figure A.3).

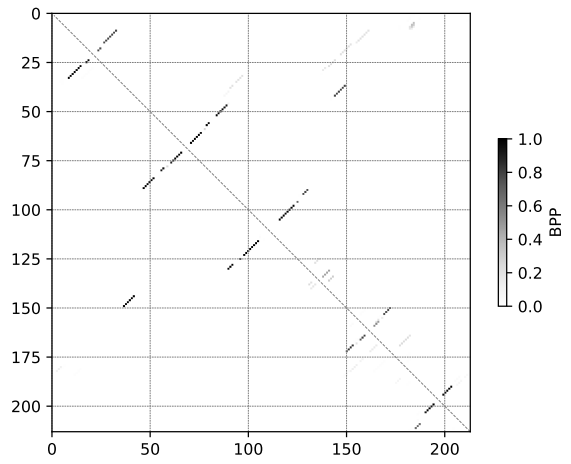


Figure A.3: Base pair probabilities of two designed sequences, computed using `ViennaRNA`. Here, the normalized ensemble defect was used as the objective function. The lower diagonal half of shows one of the best designs in the `proto` constraint set according to the average base pair distance of the three MFE predictions used; the upper diagonal half shows one of the worst designs according to the same metric. See Figure 18b for comparison and Table A.4 for the exact RNA sequences used here.

Based on the average base pair distance of the structures predicted using con-

strained `RNAfold`, `pKiss` and `RNAPKplex` to the target structure, 10 sequences were selected for preliminary experiments (see section 4.2). These candidates had to be postprocessed for the experimental assays though, which is why for later design runs, the reference data was truncated (Figure 7b) and new sequence constraints were added (Table 2).

A.6 Ensemble Defect Generalization

dirks_paradigms_2004 defined equation 12 on $N\times(N+1)$ matrices [**dirks_paradigms_2004**
In this thesis, I defined base pair probability matrices as square $N \times N$ matrices for their convenient properties. The following paragraphs should be seen as a sketch of the relation between equation 12 and the metric induced by the Frobenius matrix norm, restricted to symmetric, doubly stochastic matrices P, P' for simplicity. Every property used can be found in [**horn_matrix_2013**]. The Frobenius inner product (equation 13) is the sum of all pairwise multiplied entries of two (base pair probability) matrices:

$$\langle P, P' \rangle_F = \sum_{1 \leq i,j \leq N} P_{i,j} P'_{i,j} \tag{13}$$

The Frobenius norm $||P||_F = \sqrt{\langle P, P \rangle_F}$, induced by equation 13, can be used to define a metric $\mathrm{d}_F(P, P') = ||P - P'||_F$. With these definitions, $\mathrm{d}_F(P, P')^2$ can be expanded to:

$$\begin{aligned} \mathrm{d}_F(P, P')^2 = \langle P - P', P - P' \rangle_F &= \sum_{i,j} (P_{i,j} - P'_{i,j})^2 \\ &= \langle P, P \rangle_F + \langle P', P' \rangle_F - 2\langle P, P' \rangle_F \end{aligned} \tag{14}$$

Since P is (doubly) stochastic, $\langle P, P \rangle_F \leq \sum_i \sum_j P_{i,j} = \sum_i 1 = N$, applying for P' respectively. Therefore, for base pair probability matrices, and with equation 12:

$$\mathrm{d}_F(P, P')^2 \leq 2N - 2\langle P, P' \rangle_F = 2\,\mathrm{d}(P, P') \tag{15}$$

Naturally, this estimation also holds for non-square base pair probability matrices as defined by **dirks_paradigms_2004** [**dirks_paradigms_2004**]. The advantage of using symmetric, doubly stochastic matrices would be more apparent by expressing the Frobenius inner product as a trace $\langle P, P' \rangle_F = \mathrm{tr}(P^{\mathrm{T}} P')$, observing that $P^{\mathrm{T}} P'$ is still doubly stochastic and the largest eigenvalue of a doubly stochastic matrix is equal to 1 [**horn_matrix_2013**]. The *root-mean-square deviation* (RMSD) is a normalization of d_F that is being practically used to compare base pair probability matrices [**zhang_linearpartition_2020**].

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet.
Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.
Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, den 31. Mai 2021