universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Computational analysis and design of RNA based gene regulators

verfasst von / submitted by

Birgit Tschiatschek, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 910
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Computational Science UG2002
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

DNA deoxyribonucleic acid
RNA ribonucleic acid
mRNA messenger RNA
tRNA transfer RNA
rRNA ribosomal RNA
sRNA small RNA
ncRNA noncoding RNA
miRNA microRNA
RNAi RNA interference
siRNA small interefering RNA
snRNA small nuclear RNA
snoRNA small nucleolar RNA
gRNA guide RNA
CRISPR clustered regularly interspaced short palindromic repeats
Cas CRISPR associated proteins
RBS ribosome binding site
taRNA trans-activating RNA
crRNA cis-repressed mRNA
sgRNA small guide RNA
STAR small transcription activating RNA
GFP Green fluorescence protein
UTR untranslated region
MFE minimum free energy
IPTG isopropyl-beta-D-thiogalactopyranoside
RMSE Root Mean Square Error
lm linear regression
pls partial least squares
enet elastic net
svmRadial support vector machine with radial basis function kernel
treebag bagged tree
rf random forest
gbm stochastic gradient boosting

1 INTRODUCTION

The following chapter provides general background information about RNA and its diverse functions, synthetic engineering of RNA and computational design of RNA-based gene regulators as well as a short overview about machine learning.

1.1 FROM RNA TO RIBOREGULATORS

Nucleic acids, which include deoxyribonucleic acid (DNA) and RNA, are essential for all known forms of life. DNA carries the genetic information, while RNA was primarily seen as an intermediate information carrier from DNA to protein, since the *central dogma of molecular biology* was proposed in 1958 by Francis Crick [1].

The information from DNA is transcribed into a so called messenger RNA (mRNA) and serves as a template for the protein synthesis in the ribosome. transfer RNA (tRNA) molecules transport amino acids in the ribosome to the growing amino acid chain until the entire mRNA sequence is translated into a protein. Ribosomal RNA (rRNA) is responsible for the formation of the peptide bond [2] and it takes part in the decoding of the mRNA. Together, rRNA and protein constitute the ribosome. Over 40 years ago the idea was introduced that RNA was the primordial molecule of life and RNA-based evolution existed before protein synthesis [3–5]. The *RNA world* [6] hypothesis was supported by the discovery of ribozymes, RNAs with catalytic functions [7, 8]. Over the last decades it turned out that RNA serves various functions besides their transitory role in protein synthesis. In prokaryotes these RNAs are prevalently termed small RNAs (sRNAs), while they are commonly referred to as noncoding RNAs (ncRNAs) in eukaryotes [9]. The functions of these RNAs - which do not code for proteins - range from genome organization to gene expression and catalytic functions [10].

REGULATORY FUNCTIONS OF RNA In eukaryotes, it has been shown that RNA-based catalysis takes place at the spliceosome. These eukaryotic ncRNAs belong to the subclass of small nuclear RNAs (snRNAs) [11]. Other ncRNAs, the small nucleolar RNAs (snoRNAs) are involved in chemical modifications of rRNAs, tRNAs and sRNAs that turned out to be needed for ribosomal and cellular function [10]. These snoRNAs are found in the nucleolus and guide methylation and pseudouridylation [10, 12]. microRNAs (miRNAs) are RNAs (~ 22 nucleotides) that inhibit translation and thereby repress gene expression. miRNAs were

first identified in *Caenorhabditis elegans*, but in the meanwhile they have been identified in many species [10]. Small interefering RNAs (siRNAs) are of similar size as miRNAs and also inhibit gene expression. They are involved in a process called RNA interference (RNAi), a gene silencing mechanism [13].

In bacteria many small regulatory RNAs have been found, which execute diverse adaptive responses [10, 14]). Some sRNAs show intrinsic activity such as RNase P, or support a ribonucleoprotein. Further, protein-binding sRNAs exist that control proteins by imitation of the interacting protein's nucleic acid structure. Thereby, the activity of a protein is regulated [14]. But the majority of identified sRNAs controls gene expression by pairing with complementary regions in mRNAs and usually they tune a mRNA's translation and stability. These sRNAs are either encoded in *cis* or in *trans* [14]. Cis-encoded RNAs are encoded on the same DNA sequence as their target RNA (but on the opposite strand), thus having extended complementary regions to the target. The two RNAs react in trans. Trans-encoded RNAs are placed in a different region on the chromosome and have limited complementary to their target. The majority of the cis-encoded antisense sRNAs expressed from plasmids, bacteriophage, and transposons serves to control the number of mobile genetic elements. [14–16] Other groups function as antitoxins or regulate the gene expression in an operon.

Most of the trans-encoded base pairing sRNAs inhibit translation, lead to degradation of mRNA, or both, thereby effectively repressing protein levels. Some of the trans-encoded base pairing sRNAs have dual function(s) - besides the base pairing with a target mRNA they encode proteins. Another class of RNA regulators are riboswitches, which are part of the mRNA they regulate, they act in cis. Upon the binding of small molecules by the so-called aptamer region a structural change is induced in the RNA, resulting in altered transcription or translation [14]. The latest discovery of regulatory RNAs are the guide RNAs (gRNAs) from the clustered regularly interspaced short palindromic repeats (CRISPR)/ CRISPR associated proteins (Cas) system that detects viral DNA or RNA and degrades it [10].

Frequently the function of RNA-based regulation of gene expression in bacteria relies on the formation of RNA secondary structures (see Section 1.2.1). Intrinsic terminator hairpins can control transcription elongation, for instance, or secondary structures that mask the ribosome binding site (RBS) can hinder translation initiation (see Figure 1). Those structures operate on gene expression in cis. Interactions with trans-acting sRNAs can as well control the formation of these cis-acting structures. Thereby, switches that flip at the RNA level can be created [17].

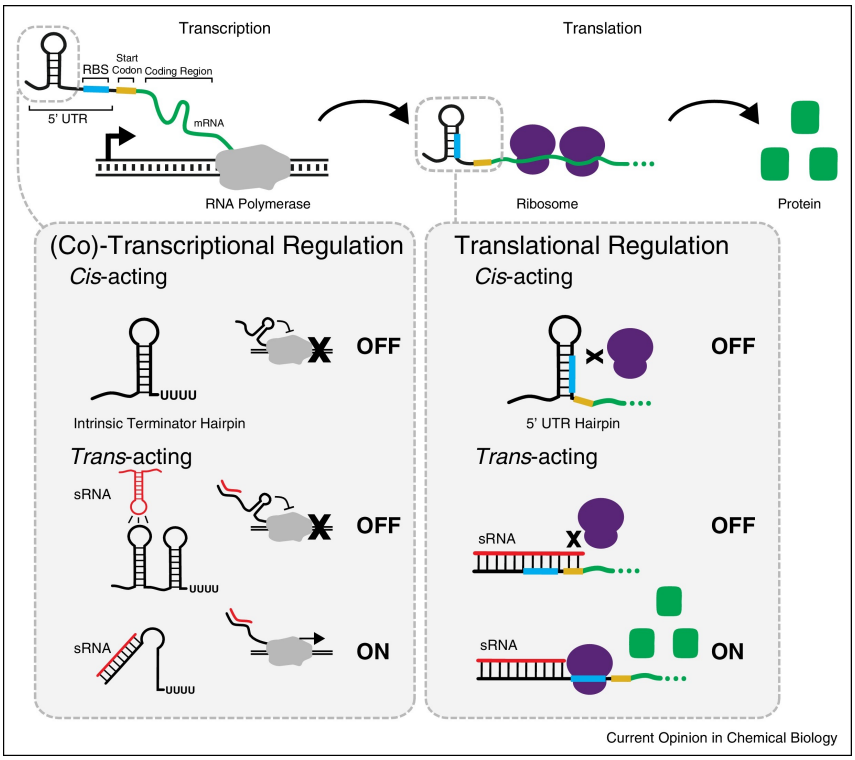


Figure 1: RNA-based gene regulation often involves RNA secondary structures. The dashed boxes show examples of RNA-based transcription and translation regulation mechanisms that operate in cis and in trans in bacteria. Premature transcription elongation can be achieved through the formation of co-transcriptional intrinsic transcription terminator hairpins in the 5′ untranslated regions (UTRs) of mRNAs (gene expression OFF; shown in the left dashed box). Alternatively, hairpins can occlude the RBS, and thus block the ribosome binding and the translation initiation (gene expression OFF; shown in the right dashed box). The formation of these cis-acting RNA structures can be controlled through interactions with trans-acting sRNAs. Therefore, RNA switches can be employed to control gene expression, based on sRNAs that flip RNA switches [17]. Graphic taken from [17].

1.2 RNA IN SYNTHETIC BIOLOGY

In general, synthetic biology can be characterized as the 'Engineering of biology' [18]. It focuses on generating new biological systems or altering the existing functionality of such systems [18, 19]. Therefore synthetic biology seeks to apply principles from engineering disciplines like mechanical and electrical engineering [20] and computer science [21]. Principles as standardization, modularity and modeling accelerate the development of biological modules and devices [21, 22]. As reviewed in [19] there have already been modules like oscillators, switches, logic formulas, and pulse generators. In contrast to electri-

cal or mechanical engineering, where elements can be electrically and spatially separated from one another, it is not possible to combine biological components in this manner. Within the cell biological parts can interact with each other and crosstalk between the components can influence the sought behaviour. If components are used, which are based on existing natural ones, it can come to interference, due to divergent environmental conditions than that in which they were described [19–22]. The composition of more complex circuits that function *in vivo* struggles with the limited number and orthogonality (non-crossreactivity) of available biological elements. Therefore, new regulatory biological parts have to show a broad and dynamic range and low system crosstalk, additionally modularity is required to allow flexible design of new biological devices [19, 22]. Those aspects have to be improved in order to make use of the potential of synthetic biology, since engineered biological systems possess a great capacity in developing applications to many (societal/global) challenges like environmental remediation, therapeutics and pharmaceuticals, and sustainability (e.g biofuels) [22, 23]. By using the cell's inherent capabilities, and to sense and react to changing environments. The underlying mechanism is a regulatory network of molecules that control the expression of certain genes as reply to environmental signals, therefore the competence to design (synthetic) biological systems is tightly connected to the ability to direct gene expression [17]. Most of the engineered systems worked on the transcriptional level using protein-based regulators. [19]. But many protein regulators show a peculiar behaviour thereby hindering their use in biological circuits. [24] Due to the various models of natural RNA regulators there is increasing interest in using RNA to set up synthetic biological systems. [23]

1.2.1 *Advantages of RNA in Synthetic Biology*

Additionally to the variety of mechanisms by which RNA can operate (see Section 1.1), RNA shows properties attractive for designing synthetic biology systems. RNA as well as DNA are biopolymers assembled from nucleotide monomers. The nucleotides consist of a 5-carbon (pentose) sugar, a phosphate group and a nitrogen base. In RNA the sugar is a ribose and in DNA it is a deoxyribose. DNA uses the four bases adenine (A), guanine (G), cytosine (C) and thymine (T), while RNA uses uracil (U) instead of T. The interactions of the four bases are well-determined through the effects of hydrogen bonds, base-stacking, and electrostatic interactions [23]. Watson-Crick base pairs connect two bases by hydrogen bonds, G-C and A-U. But also G-U base pairs - so-called wobble base pairs - can occur. G-C pairs form three hydrogen bonds, while A-T and wobble pairs form two. Therefore, the G-C pairs are stronger than the other base pairs. In contrast to DNA, RNA exists predominantly single stranded and folds back

onto itself. Segments of intramolecular base pairs form helical structures (also termed *stems*) with unpaired sections in between that form *loops*. Those arrangements of helices and loops are the *secondary structure* of the molecule. Further, non-covalent interactions can be caused through spatial proximity of secondary structure parts. In contrast to Watson-Crick base pairs the non-covalent interactions are frequently energetically less advantageous [25]. The secondary structure of RNA mostly prescribes its folding, whereas protein folding is in large parts based on tertiary structures [23]. Unlike for proteins RNA secondary structures seem to form previous to tertiary structures and to provide a scaffold for the formation of tertiary structures [26, 27]. Therefore it is reasonable to consider secondary structure prediction as a proxy for the actual tertiary structure. Furthermore, it has turned out that the secondary structures of RNA are sufficient to describe many aspects of RNA molecules for practical applications including thermodynamic(s), kinetic(s), and evolutionary properties/features [20].

Models that predict the secondary structure of a given RNA sequence have been established which depend on the optimization of energies caused by the Watson-Crick basepairs and the wobble pair [23]. The most applied approach is minimzation of free energy. NUPACK [28], RNAstructure [29], Vienna RNA Package [25], SFold [30], Foldalign/-FoldalignM [31], CentroidFold [32], RAF [33], RNAshapes [34] and MC-Fold [35] are programs for secondary structure prediction. For a non-exhaustive overview of the former mentioned RNA secondary structure prediction methods see [36].

Computational RNA design tools usually present a good starting point for rational design of RNAs [17]. RNA inverse folding tools predict RNA sequences that fold into a prescribed structure [37]. RNAinverse [25], NUPACK [28], RNAdesign [38, 39] and RNAblueprint [40] are such tools that can be used to design RNA sequences. For a comparison of RNA inverse folding tools see [37].

Figure 2 illustrates a future workflow of RNA design. Using fluorescent reporter assays and chemical probing the suggested RNA sequences are analysed for their cellular functionality and structure. The resulting probe reactivity maps can be integrated to improve the RNA folding algorithms by adjusting underlying models of RNA folding in the cell. Thereby the RNA structure-to-function-relationship can be analysed and further improvement of RNA design algorithms can be obtained by integrating the newly discovered design principles [17].

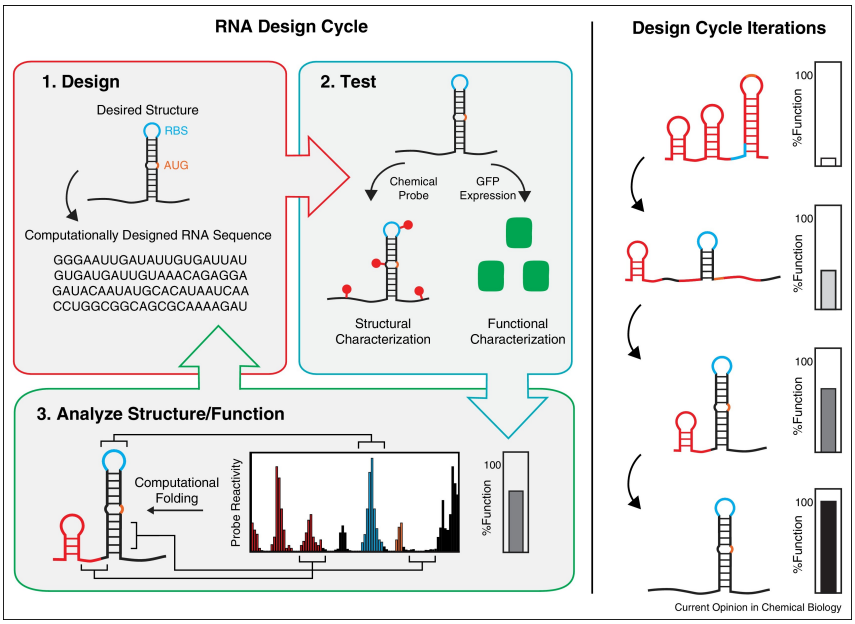


Figure 2: Workflow of (future) RNA design cycle; computational design algorithms can propose possible RNA sequences that fold into the desired RNA structure. In the next step the cellular functionality and structure of these sequences can be analysed using fluorescent reporter assays and chemical probing. To examine the structure-to-function relationship of RNA chemical probing can be used. The resulting probe reactivity maps may be used to improve the RNA folding algorithms by adapting the underlying models of RNA folding in the cell. Thereby RNA structure and RNA function can be linked and further improvement of RNA design algorithms can be achieved by integrating the newly discovered design principles [17]. Graphic taken from [17]

.

In comparison to protein-dependent regulators RNA regulators are less expensive for the cell. They do not need translation [14, 23, 41], therefore they are less costly in concerns of energetic and resource aspects [23]. Due to their length of only ~100 – 200 nucleotides RNA regulators enable a faster production. For example 1000 nucleotides are required for the average *Escherichia coli* protein of ~350 amino acids. Additionally, the effects of RNA-based regulators themselves can be fast. In the case of cis-acting riboswitches the direct connection of the

sensor domain to the mRNA makes it possible for the cell to react very fast and sensitive [14]. Furthermore, the scheduling of coupled control processes is of interest for synthetic biology, post-transcriptional systems based on RNA regulators will commonly operate faster than transcription based ones [23]. The transfer of such single-step RNA regulatory elements to other organisms is simpler than the transfer of transcription-based regulators. And it allows a modular combination with other elements [41].

Compared to protein-based regulators RNA-based regulators commonly exhibited a lower dynamic range and generated fewer orthogonal (non-corssreacting) regulators [17, 42]. But, new design ideas for RNA-based regulators have proved comparable functionality as reviewed in [17]. For examples of RNA-based regulators see Section 1.2.2.

1.2.2 *Examples of Engineered RNA-Based Gene Regulators*

Natural RNA-based elements exist that function at the transcriptional and post-transcriptional level [43, 44]. Based on these natural systems engineered RNA regulatory elements have been developed, e.g. riboregulators that regulate translation and transcription as reaction to RNA itself. But there are also de novo rational designed RNA elements.

Isaacs et al. developed riboregulators that allow post-transcriptional gene regulation. This regulation system depends on an RNA secondary structure that occludes the RBS. Therefore they inserted a complementary sequence to the RBS in cis directly upstream of the RBS. This results in the formation of a hairpin structure including the RBS at the 5′-UTR of the mRNA (which they termed cis-repressed mRNA (crRNA)). In this manner the RBS can not be recognized by the ribosome and thereby translation is inhibited. Upon binding of the so called trans-activating RNA (taRNA) (which is transcribed from a second promoter) to the hairpin loop of the crRNA gene activation is induced. The riboreguators exhibited dynamic ranges up to 19-fold [45]. Callura et al. used riboregulators from Isaacs et al. [45] as basis for their genetic switchboard. Their metabolism switchboard integrates four synthetic riboregulators and enables independent control of different genes in parallel. The results of this metabolism switchboard over multiple biological levels (RNA, protein, and metabolome scales) demonstrate the capability of such a higher-order control system. For activation, riboregulators induced fold changes in fluorenscene up to 200-fold increase [46].

Lucks et al. engineered the natural transcriptional attenuator from *Staphylococcus aureus plasmid pT181* which is mediated by antisense-RNA. Through mutations orthogonal variants of the transcriptional attenuator were generated that control different targets in the same cell. Additionally, they showed that attenuators can be arranged in

tandem on the same transcript and react to multiple antisense RNA signals. Finally, they used an attenuator to control the transcription of anitisense RNAs which were further used in an RNA-mediated transcriptional cascade [47].

Mutalik et al. rationally designed orthogonal RNA-based translation regulators. The pairs of orthogonal regulators were built from the RNA-IN-RNA-OUT antisense RNA-mediated translation system. Therefore, they build a model from numerous in vivo reporter assays. Based on the model they predicted the performance of further regulator pairs and validated these predictions by experiments. Thereby, they engineered two RNA pairs exhibiting predictable performances. They observed changes in dynamic range up to ten-fold for their target repression [48].

Rodrigo et al. computationally de novo designed RNA riboregulators based on a physicochemical model. New sequences were generated by searching the possible sequence space for sequences that are consistent with specified structures. The resulting riboregulators did not show significant similarity to any familiar ncRNA sequence. Further, their devices operate separately and can be used to build combinatorial logic gates [49].

Chappel et al. developed small transcription activating RNAs (STARs) small transcription activating RNAs using a completely artificial mode of regulation. The STAR regulators interfere with the formation of a(n intrinsic) terminator hairpin, thereby enabling transcription. Employing this mechanism they constructed a set of four highly orthogonal STARs, one exhibiting 94-fold activation. Further analysis led to design rules for STAR function. These STARs demonstrated high orthogonality to each other and to a library of RNA transcriptional repressors, furthermore they are composable. Thereby, making it possible to build novel RNA-only transcriptional logic gates that were inaccessible before [50].

The so-called CRISPR interference system only needs two molecules, a Cas9 protein that is catalytically inactive and a small guide RNA (sgRNA). The dCas9:sgRNA complex works by RNA-guided DNA binding. The sgRNA functions as DNA recognition platform that binds to DNA targets by Watson-Crick base pairs. With this system up to 300-fold transcription repression was achieved [51]. In this manner repurposed CRISPR systems combine the simplicity of RNA design and the good dynamic ranges of protein regulators [17].

Overall, new designs of RNA-based regulators can compete with previous protein-based regulators and at the same time make use of RNAs' designability and versatility [17]. Protein-based regulators showed dynamic ranges over 350-fold for a commonly used inducible promoter [52] and up to 480-fold for sigma factors and cognate promoters [53]. New design ideas for RNA-based regulators have proved comparable functionality exceeding averge fold changes in fluores-

cence of 400 [54] and up to 300-fold transcription repression [51]. In this thesis Green's toehold switch designs are used for further analysis, for more details of the toehold switch designs see Section 1.4

1.3 COMPUTATIONAL DESIGN OF RNA-BASED GENE REGULATORS

RNA-based gene regulation frequently requires the formation of specific secondary structures. Therefore, algorithms for RNA secondary structure prediction can be applied for designing RNA-based gene regulators [55]. Commonly *RNA inverse folding* has been used to design RNAs [37]. Due to the simplicity of the RNA secondary structure model (see Section 1.2.1), accurate predictions can be computed efficiently [25].

1.3.1 *RNA Inverse Folding*

The *inverse folding problem* seeks to compute sequences which adopt a given target structure [20, 56], while the *folding problem* is to compute the secondary structure (or native structure) θ into which the sequence x (of same length) folds [20, 56]. Folding algorithms predict the secondary structure $\varphi(x)$ in ground state of a given RNA sequence x and the corresponding folding (free) energy $g(x) := \min_\theta g(\theta, x)$ [20]. The minimum free energy (MFE) structure is a secondary structure exhibiting the highest probability at equilibrium, but it is not necessarily only one structure [57]. In contrast, inverse folding algorithms predict for a given secondary structure ψ the set $S(\psi) := \{x | \varphi(x) = \psi\}$ of sequences which have ψ as ground state structure [20].

There are efficient and exact algorithms to solve the folding problem $x \mapsto \varphi(x)$, but none to solve the inverse folding problem. Thus, heuristic algorithms have to be applied [20, 56]. Usually the problem is transformed into an optimization problem. The goal is to find a sequence x with a ground state structure $\varphi(x)$ that minimizes the distance to the target structure ψ, $d(x) := d(\varphi(x), \psi) \to \min$, where the length of sequence x and target structure ψ are equal. Further, $S(\psi) = \{x | d(x) = 0\}$ [20]. Different measures have been formulated to determine distances between RNA secondary structures. The base pair distance $d(\theta, \psi) = |(\theta \setminus \psi) \cup (\psi \setminus \theta)|$ is the simplest and most applied measurement, it enumerates the number of base pairs existing in only one of the structures, but not in the other one. The base pairing rules give a simple test whether a sequence x can possibly adopt the secondary structure ψ, since all base pairs of the target structure ψ have to match to pairs of nucleotides in sequence x that are either Watson-Crick or wobble pairs [20]. If this is true sequence x is consistent with structure ψ ($x \in C[\psi]$). Obviously, solutions of/to the inverse folding problem must be searched within $C[\psi]$. Thus the simplest inverse folding algorithms - adaptive walk like heuristics -

initially select a random structure $x_0 \in C[\psi]$ and try to improve x_0, either by point mutations or by the exchange of base pairs. If the mutated sequence x' satisfies $d(x') < d(x_0)$ it is accepted, otherwise another modification of x_0 is tried [20, 58]. If no mutation can be found that reduces the cost function, the inverse folding algorithm stops and restarts with another initial sequence [58]. As a result of the characteristic sequence-to-structure map of RNA [59–61], often such simple heuristics are successful even for longer sequences [20].

The set of sequences with ψ as the ground-state termed neutral network of ψ is approximately uniformly (and also densely for naturally occurring structure densely) embedded in $C[\psi]$.

Due to the computational inefficiency of the adaptive walk other approaches for example employ a divide-and-conquer strategy like RNAinverse, RNA-SSD, INFO-RNA, DSS-OPT [20]. Starting from a seed sequences a local search is applied to mutate the former sequences and repeatedly solve the RNA folding problem by energy minimization [37]. Near to the seed sequences a sequence with the specified structure can be found. Evolutionary algorithms can be an improvement to this problem [62]. Further approaches are different sampling strategies like global sampling as used in RNA-ensign or weighted sampling in IncaRNAtion. CPdesign (which is part of the RNAiFold framework) makes use of exact constraint programming. That allows to compute all possible sequences that fold into a specific structure. But this is accompanied by high computational demand. Therefore, in LNSdesign the same authors applied a neighborhood search heuristic. In another approach, the input is a predefined shape not a complete predefined structure, as implemented in RNAexinv [20, 37].

But RNA sequences do not only adopt a single structure, they produce an ensemble Φ consisting of secondary structures. Every structure θ in this ensemble $\Phi(x)$ that is consistent with x, $x \in C[\theta]$, occurs with relative frequency $p(x,\theta) = exp(-g(x,\theta)/RT)/Z$, with the partition function $Z := \sum_\theta exp(-g(x,\theta))$. This knowledge enables refining of inverse folding heuristics: now there are two design principles, rather than a simple optimization. The positive design principle requires to enrich the desired structure ψ and the negative design principle to deplete all undesirable structures ξ from the ensemble. An interpretation of the negative design rule could be, that alternative structures ξ are discouraged more rigorously the larger the difference to ψ ($d(\xi,\psi)$) becomes. If the entire Boltzmann ensemble of sequence x is considered, the minimization of the ensemble defect $\nu(x,\psi) := \sum_\theta d(\theta,\psi)p(x,\theta)$ is common [20]. The ensemble defect computes the average number of incorrectly paired nucleotides relative to the target secondary structure at equilibrium. The evaluation is carried out over the Boltzmann-weighted ensemble of unpseudoknotted secondary structures [57]. Based on the base pairing prob-

ability matrix of x the ensemble defect can be calculated [20]. The computation of the ensemble defect is implemented in NUPACK [57]. NUPACK is a software suite with focus on analysis and design of (nucleic acid) secondary structures. The sequence design algorithm is defined as an optimization problem with the objective of decreasing the ensemble defect below a user-defined value.

1.3.2 *Modeling External Triggers*

RNA can react to a variety of external triggers. From a computational perspective, a change in temperature may be the simplest external trigger to which an RNA may respond. For example, in the ViennaRNA package [25] enthalpies and entropies are tabularized separately, thus temperature is an explicit parameter. Therefore it is possible to use design functions like $d(\psi T'(x), \psi T') + d(\psi T''(x), \psi T'')$ to determine sequences x showing a switching behaviour from a ground state structure $\psi T'$ to another ground state $\psi T''$ as a result to a temperature change from T' to T'' [20].

As stated in [20], de novo designed RNA thermometers have been shown to work in vivo at physiological temperatures. Further, computer-based rational design combined with basic in vivo screening led to RNA thermometers exhibiting performance comparable to natural existing ones. Also naturally components like an RNA thermometer and the hammerhead ribozyme were combined to construct self-cleaving RNA molecules with temperature sensitivity [63].

In theory pH sensors [64] and RNA thermometers are very alike. But, due to the absence of ready-to-use parameters that describe how RNA folding energies are influenced by pH, yet it is not possible to computational design pH sensors [20]. However, it is possible to computationally design switches that react to a nucleic acid trigger. Combining software for RNA design and target prediction enables designing such switches [20]. RNAcofold [65] and RNAup [66] from the ViennaRNA package [25] can be used for this task. The former mentioned software suite NUPACK [28] and RNAiFold [67] already allow multi-strand design of nucleic acids. MultiSrch/StochSrchMulti [68] applies a similar strategy. The great advantage of nucleic acid triggers is the opportunity to model the entire system in silico [20].

One example is the RAJ11 system designed by Rodrigo et al. [49]. In contrast, riboswitches are triggered by small molecules, that usually are metabolite ligands. At the moment ligand binding can not be directly integrated into RNA folding algorithms. A workaround is to imitate the ligand bound state of the so-called aptamer region by using structural constraints. RNA folding algorithms offer the possibility to specify a set α of unpaired and paired bases representing the ligands' structure, that will then be enforced during structure calculation.

$\varphi(x|\alpha)$ is the most stable structure under constraint a, and $g(x|\alpha)$ its associated folding energy. Both can be efficiently computed with `RNAfold -C`. ($\varphi(x)$) is the unconstrained ground state and $g(x)$ its energy. It is important to notice that the unconstrained ground state structure $\varphi(x)$ must not display the aptamer structure with ligand bound. In the ideal case, all aptamer base pairs are completely excluded from the unconstrained secondary structure. In this case, $\varphi(x) \setminus \alpha = \emptyset$ is set as negative design goal. By design, the unconstrained ground state energy $g(x) < g(x|\alpha)$. The ligand binding energy $e_L < 0$ stabilizes $\varphi(x|\alpha)$ in presence of the ligand. It needs to be sufficietly high, making the ligand-bound state the ground state. e_L can be explicitly calculated for nucleic acid ligands, for small molecule ligands (metabolites) one has to resort to values from literature or estimates for e_L. If a design shall comprise a structural change which is induced by ligand binding, the ligand-bound state has to be the ground state. This leads to the following inequalities $g(x|\alpha) + e_L \leqslant g(x) \leqslant g(x|\alpha)$. A general method to formulate this as a design objective is to set $g(x) - g(x|\alpha) \approx e_L/2$, i.e. $g(x)$ should be roughly halfway between $g(x|\alpha) + e_l$ and $g(x|\alpha)$. Currently available design tools do not incorporate this constraint folding frame-work in a very systematic way. Therefore, the energetic contributions of ligand binding have to be integrated manually, leaving room for discrepancies between model and experimental setup [20].

1.4 GREEN'S TOEHOLD SWITCHES

In 2014 Green et al. [54] published de novo designed riboregulators using new design principles. The design is shown in Figure 3, colored regions are pre-determined and therefore the same in all devices. These riboregulators make use of linear-linear interactions that are mediated by a toehold, and are termed toehold switches. These devices involve two single stranded RNAs, a switch and a trigger. At the 5′ end of the switch is a single-stranded toehold region (domain a), which enables initial binding of the trigger RNA. Immediately adjacent to the toehold region resides a hairpin-based processing unit, which contains a strong RBS and the start codon. The repressing domain b (the stem at the 5′-end) and domain a together form the binding site. The switch contains the coding sequence of the gene intended to be regulated. A common 21 nucleotide linker sequence and the gene of interest are located downstream of the hairpin. The linker sequence encodes low-molecular-weight amino acids. The trigger incorporates a stretched single-stranded region. Region a* in the trigger is complementary to region a in the switch, and b* in the trigger to b in the switch. After the trigger binds to the toehold, it unwinds the hairpin, which leads to exposition of the RBS and the start codon, thereby translation is initiated. In the absence of the trigger

RNA the hairpin processing module works as a translation repressor. The RBS sequence stays unpaired within the loop of the hairpin.

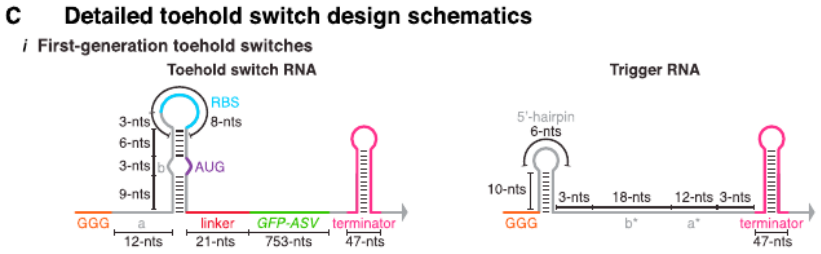


Figure 3: First-generation toehold switch design. The toehold switch RNA (shown left) contains a 12 nucleotide long toehold sequence (region a), followed by a hairpin that sequesters the hairpin and the start codon. Adjacent to the hairpin is a common linker sequence and the coding sequence of the gene of interest. Region a and b together form the binding site for the trigger. The trigger RNA (shown right) contains a complementary region a* to the toehold region in the switch, and also a complementary region b*. The regions illustrated in colour are pre-determined. Graphic taken from [54].

1.4.1 *Design Principles*

In general, the activation of gene expression by a riboregulator requires switching from a translation repressing structure to an active secondary structure. The formation of the translation promoting secondary structure is induced through interaction with a trans-acting RNA. Drawing conclusions from Kudla et al. [69] and [70] Green et al. assured that the toehold switches isolate the domain around the start codon to prevent translation, instead of binding to the RBS or the start codon [54]. According to [69] secondary structures around the start codon may be important for translation activation, besides the Shine-Dalgarno sequence. Additionally they referred to [70] who found that there is a trend towards reduced mRNA folding nearby the translation start, i.e. the start codon of mRNAs. Previous engineered riboregulators (Figure 4) of translation relied on base pairings to the RBS to inhibit ribosome binding, thus blocking translation. For examples see Section 1.2.2. Therefore, triggers designed to activate translation have to incorporate an RBS sequence that removes the repressing sequence. These sequence constraints minimize the potential sequence space. Further constraints arise from the use of U-turn loop structures - that occur in natural systems - that promote loop-loop and loop-linear interactions. Even though loop interactions are favoured evolutionary in nature, alternative approaches using linear-linear interactions are easily controlled through rational engineering, thereby providing advantageous reaction kinetics and ther-

modynamic behaviour [54]. Therefore Green's toehold switches use toehold-mediated linear-linear interactions to start RNA-RNA strand displacement. This toehold-mediated interactions were developed in vitro before [71]. Additionally, the adjacent bases to the start codon form base pairs of 6 and 9 base pairs, which build the stem of the hairpin, while the start codon stays unpaired. This results in a bulge (3 nucleotides) near the midpoint of the stem. Since the repressing domain b is not complementary to the start codon potential trigger sequences are not limited by the need to include start codon bases.

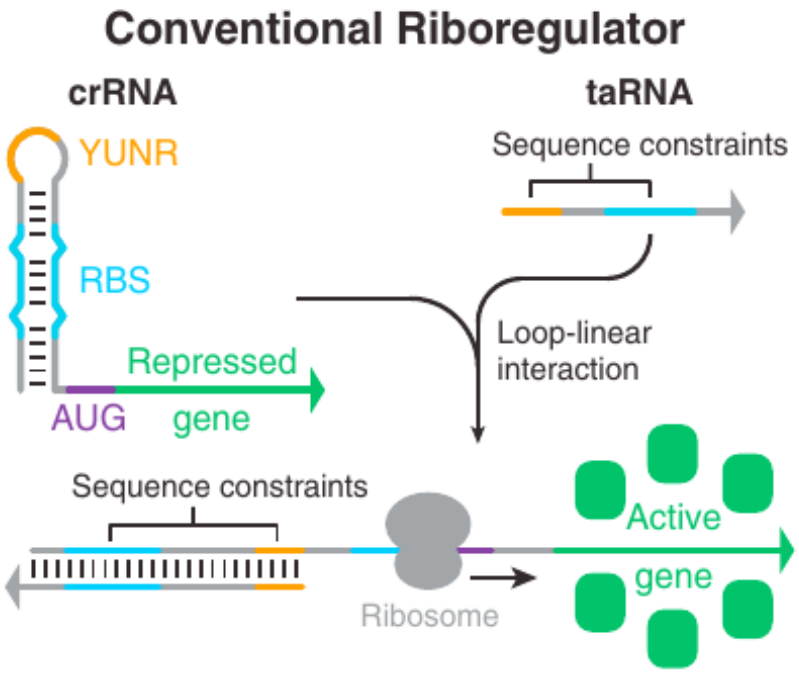


Figure 4: Conventional riboregulator: This design achieves repression by formation of an RNA hairpin structure that prevents translation by sequestering the RBS (cis repression). The corresponding mRNA is termed crRNA. A non-coding, trans-acting RNA, referred to as taRNA activates the gene expression through a linear-loop interaction with the crRNA. Previous designs obtained repression through binding/acting of certain molecules [45]. Graphic taken from [54].

1.4.2 *In Silico Design and In Vivo Validation of Toehold Switches*

At first the design specifications of the toehold switches were defined involving the secondary structure of switch and trigger, conserved sequences, and the size of interaction domains. The secondary structures are shown in Figure 3. The conserved sequences correspond to the colored domains, namely the RBS of the mRNA, the transcriptional terminator of T7 RNA polymerase, the linker sequence and the

coding sequence of the regulated gene, and a leader sequence (GGG) required by T7 RNA polymerase.

The software NUPACK was used to generate the RNA sequences according to the design requirements. In the beginning 24 toehold switches were designed to estimate the performance in vivo. Afterwards they expanded their library consisting of orthogonal elements chosen for low crosstalk with the remaining library: they designed 646 unique toehold switches and simulated all 417.316 pairwise interactions to assess crosstalk interactions. Using a Monte Carlo algorithm they identified a subset of 144 switches with the lowest crosstalk level. Crosstalk is the ratio of the Green fluorescence protein (GFP) fluorescence from a switch with a non-cognate trigger and the fluorescence of the switch with a cognate trigger (switch in triggered state). This resulted in a library of 168 toehold switches, encompassing the initial 24 toehold switches and the subset of 144. These toehold switches allow to regulate translation independently in vivo. Further, the switch sequences are sufficiently distinct, therefore sequence dependent effects can be discerned. These 168 toehold switches are referred to as first generation switches.

E. coli BL21 Star DE3 was used for testing. The switch sequences were expressed from medium copy plasmids, while the trigger sequences were expressed using separate high copy plasmids. Isopropyl-beta-D-thiogalactopyranoside (IPTG) was used to induce the expression of both strands through T7 RNA polymerase. The switch output performance was analysed via flow cytometry using GFP. GFP fluorescence was measured from cells in ON state (expressing both the switch and its cognate trigger) and from cells in OFF state (expressing the switch and a non-cognate trigger. The ON/OFF fluorescence ratios were calculated from the histograms showing the mode fluorescence intensity. The ON/OFF ratios then are used to describe the activity of the toehold switches [54]. For validation Green et al. compared their toehold switch designs with the engineered riboregulators 10 and 12 from Isaacs et al. [45] which showed ON/OFF ratios of 11 ±2 and 13 ±4. On average the 168 first-generation toehold switches showed an ON/OFF ratio of 43. 20 switches showed ON/OFF ratios exceeding 100, and 87 had ratios between 10-and 100, altogether almost 2/3 (0.64%) showed ON/OFF ratios greater than 10.

The in silico designed first-generation switches failed at the goal to create high-performance switches enabling faster generation of new genetic circuits, they exhibited a wide range of ON/OFF ratios. Therefore, Green et al. analysed the first-generation switches to determine design features that resulted in high-performance switches. The insights were incorporated into the second-generation toehold switches.

1.4.3 *Applications of Toehold Switch Designs*

Already in [54], they showed that toehold switches can be triggered by mRNAs and endogenous RNAs. Some modifications to the high-performance switch 1 from the first generation were made, such as extending the toehold region and adaption of a programmed RNA refolding mechanism. They designed sensors to detect *mCherry* mRNA and the endogenous *E. coli* sRNA RyhB. Further, they showed that toehold switches can be functional when inserted into the *E. coli* genome and can regulate translation of different endogenous genes such as *uidA*, *lacZ* and *cheY*. The strain *lacZ*::Switch C needs lactose/IPTG and trigger RNA C to start expression, resulting in a genetic AND circuit. They also documented that multiplexed regulation is possible with the toehold switches and demonstrated a functional 4-input AND circuit with integrated toehold switches [54]. The toehold switches were also successfully tested on a paper-based platform. Pardee et al. presented a method that allows embedding synthetic gene networks into paper (and also other porous materials). These paper-based toehold switches lead to induction greater than 20-fold up to 350-fold of synthetic gene networks [72]. Although a large library of toehold switches with high-performance existed, a scalable strategy to use the high-performance toehold switch parts in biological circuits was missing[73]. In [74], Green et al. demonstrate a method to construct RNA-only nanodevices for complex cellular logic computation, where all main circuit operations were converted into RNA-RNA interactions [73]. These 'ribocomputing' systems were built of de-novo designed parts and use RNAs as input and protein as output signal. A gate RNA is responsible for signal processing. The sensing domains of the gate RNA used in the ribocomputing devices are toehold switches. They showed that ribocomputing devices in *E. coli* are capable of evaluating multi-input computations such as four-input AND, six-input OR up to a 12-input expression ((A1 AND A2 AND NOT A1*) OR (B1 AND B2 AND NOT B2*) OR (C1 AND C2) OR (D1 AND D2) OR (E1 AND E2)) [74]. In [24], Kim et al. de-novo-designed two riboregulators for translational repression by employing toehold-mediated strand-displacement reactions. Their design is inspired by the toehold switches resulting in two types of repressors: the toehold repressor and the three-way junctions repressor. The repressors have sensing and logic capabilities and achieve repression up to 300-fold, which is comparable ro protein-based repressors. These selected examples already demonstrate the capabilities of the toehold switch design.

1.5 MACHINE LEARNING

The aim of this thesis is to define numerical features that describe individual RNA sequence designs in numerical manner and to evalu-

ate the predictive performance of the defined features using machine learning algorithms. Therefore, a short general overview about machine learning is given.

Machine learning comprises various techniques that enable making predictions about unseen data from observed one. *Learning* can be understood as inferring computational models from training data. Based on a training set the parameters of an adaptive model are tuned [75]. The models' complexity can vary from simple and inflexible classic statistical linear regression models to more complex and flexible such as neural networks [76]. Learning can be categorized into supervised learning, unsupervised learning and reinforcement learning. The methods differ in the type of learning signal or given feedback to the learning system [77]. In *supervised learning* the training data is provided with their output variables/target vector. Tasks, in which each input vector is assigned to one discrete category of a finite number are called classification. If the asked output is one or more continuous variables it is called regression [75]. In contrast, in *unsupervised learning* the training data does not include output values, only input values. In *reinforcement learning* (or *learning with a critic*) no output variable is provided. The output variables have to be inferred from a feedback whether the potential category is right or wrong. *Overfitting* describes the case when the model is too closely fit on the training data enabling perfect predictions for them, but performs weak on unknown data [78].

1.5.1 *Feature Selection*

Feature selection is of interest in research areas where data sets with many feature variables are available. Feature selection tries to find a subset of predictor variables that is as small as possible, and at the same time optimizes the prediction performance on new data. Possible benefits of feature selection may be an improvement of the prediction performance. That could be achieved by excluding irrelevant noisy features and/or by lessen the issue of overfitting by a dimensionality reduction, because irrelevant features may result in overfitting. Another point may be increased effectiveness, due to reduced memory usage, less learning time and processing time due to a smaller feature set. Additionally, an improved understanding of the data and the underlying process, to recognize relevant features to a target may be achieved [79, 80]. There are different methods to the feature selection problem.

1.5.2 *Cross-Validation*

In order to find the model with the best predictive power on new data the performance of the prediction should be tested on an inde-

pendent data set (also called validation set). If there is enough data then some of the available data can be used to train a single model or a set of models, while other data can be used as a test set to evaluate the predictive performance. The model with the best predictive performance can be chosen. Often only limited training and test data is available, but to build good models as much as possible of the data is needed for training. A small test set will only result in a noisy estimate of predictive performance. *Crossvalidation* is a solution to this dilemma. In S-fold cross-validation the data is split into S groups. For the training of the model(s) $S-1$ of the groups are considered, the evaluation then is performed on the remaining group. This procedure is applied for all S choices for the held-out group. Afterwards the performance measures from the S runs are averaged. In cases where the data is specially rare a useful approach may be to set $S = N$, with N as total number of data points. This is the so-called leave-one-out method [75].

1.6 FOCUS OF THIS THESIS

Functional non-coding RNAs play an important regulatory role in cells of all kingdoms of life. These regulatory functions can be mediated through various different mechanisms, e.g. translational gene regulation due to a small RNA input. Such a system was artificially designed in a publication by Green et al. [54]. This publication provides an extensive data set of de novo designed sequences including quantitative data of verification experiments, mainly fluorescence readout of gene reporter assays. The aim of this thesis is to define features that describe individual sequence designs in numerical manner. This should enable to predict the functionality of such sequence designs computationally in advance, before analysis in the laboratory. The features were determined by a knowledge-based approach and calculated using several RNA related tools. To evaluate the predictive performance of the defined features machine learning algorithms were applied.

2 MATERIAL AND METHODS

In this chapter the analysed data and applied methods are described and summarized.

2.1 DATASET

RNA sequences designed by Green et al. [54] and corresponding performance measures, i.e. GFP fluorescence data, are the underlying data basis of this thesis. The fluorescence measurements combined with the RNA sequences were used for further analysis and validation of the informative value of the defined features (Section 3.1). The design process of the used RNA sequences is summarized in in Section 1.4. For further details see the original publication and supplemental information [54].
The used data consists of 168 first generation switch and trigger RNA sequences and the corresponding fluorescence data. GFP fluorescence was measured from cells in ON state (expressing both the switch and its cognate trigger) and from cells in OFF state (expressing the switch and a non-cognate trigger. The ON/OFF fluorescence ratios were calculated from the histograms showing the mode fluorescence intensity. The ON/OFF ratios then are used to measure the performance by [54], in this analysis the focus was set on the ON state fluorescence.

2.2 FEATURE SELECTION, GENERATION AND COMPUTATION

Some features from Green et al. [54] were adopted, as well as further numerical features were created in this thesis by a knowledge-based approach. Thermodynamic features were chosen that might be important for the desired folding and unfolding of the toehold switches and the accessibility of the RBS and the start codon. Most of the features were calculated using the python interface (v3.0 API) to the RNAlib (version 2.3.3) [25, 81]. Only the ensemble defect was calculated using a python interface to NUPACK 3.0 [82]. In order to calculate the selected features python scripts were written.

2.3 DATA ANALYSIS - DESCRIPTIVE STATISTICS AND VISUALIZATION

The computed features were analysed using R [83] (version 4.1.2) for descriptive statistics and visualizations of the distribution of the individual features. The dimensions of the dataset were reported us-

ing `dim(switch_designs`. `sapply(switch_designs,class)` was used to list the data types of each variable and check if there are any missing values. In order to compute the distribution of each variable `summary(switch_designs)` was used.

The plots shown in Section 3.2 were created with `hist(switch_designs[,i],main=names(switch_designs)[i])` for the histograms of the input features, `plot(density(switch_designs[,i]),main=names(switch_designs)[i])` for the density plots and `boxplot(switch_designs[,i],main=names(switch_designs)[i])` for the boxplots.

2.4 CORRELATIONS

The correlation coefficients R for the correlation of the individual features with the output variables (log(ON mode), log(OFF mode) and log(ON/OFF mode)) were computed with the `cor` function from the basic R Stats Package of R [83] version 4.1.2, `cor(switch_designs[,1:l],log_on_mode)`. The coefficient of determination R^2 was computed by squaring the correlation coefficient R. $R^2=(R)^2$.

2.5 FEATURE P_SW_ACC_RBS_LINKER - STRUCTURE AND ACCESSIBILITY

The single stranded switch RNA secondary structures were computed with the `RNAfold` command from the Vienna RNA package [25, 84], version 2.5.1., with the following command line call `RNAfold -p -d2 --noLP < sw1.fa > sw1.out`. The centroid structure drawings were made and annotated using the `RNAplot` command, rotated using the `rotate_ss.pl` script and colored using the `relplot.pl` script from the same package with the command line call `relplot.pl -p sw1_ss.ps sw1_dp.ps > sw1_ss_col.ps`. The plots are used for structure analysis.

`RNAcofold` from the same package was used to predict the secondary structures of the single stranded switch and trigger RNA upon duplex formation. The following command line call was used `RNAcofold -a -d2 --noLP < sw1_tr1.seq > sw1_tr1.out`. The secondary structures of the duplexes are used for structure analysis.

The sequences were used corresponding to Green et al. [54]. The switch RNA ranges from the GGG leader sequence up to 29 nucleotides of of the GFPmut3b and the trigger RNA from the GGG leader, including the 5′ hairpin, up to the transcriptional terminator.

2.6 MACHINE LEARNING ALGORITHMS

For the task of functionality/activity prediction supervised regression algorithms were used. Different algorithms were used to build predictive models based on thermodynamic features (described in Section 3.1). The models were built in R [83] (version 4.1.2) using the package `caret`. The package offers a unified interface to various modeling functions [85].

2.6.1 *Data preparation*

Due to their distribution some features were logarithmically transformed. For basic insight default parameters were used for the algorithms unless otherwise stated. During the first phase, features were used without highly correlated and linear combinations. Highly correlated features were removed using carets `findCorrelation`. This function takes a correlation matrix and returns the set with the smallest number of predictors that can be discard, in order that pairwise correlations are below a specified threshold [85]. 0.75 was used as cutoff. Some models, such as linear models and neural networks may produce unstable solutions or show poor performance if there is multicolinearity. While models, such as classification or regression tree might be unsusceptible to highly correlated features. But, the interpretability of the model may be complicated due to multicolinearity. A classification tree may show good performance despite highly correlated features, but the selection of the features to be in the model is random. If it is necessary to remove highly correlated features models that are unaffected to multicolinearity can be used (e.g. partial least squares) or principal component analysis can be used. Another option may be to determine high correlation predictors and remove them [85] as it was done in this work. In this way, problematic predictors were removed the same way for all models. Linear combinations were removed using carets `findLinearCombos`. Since the predictor that can be expressed through a linear combinations of other predictors not only is redundant, but also results in arbitrary estimated slopes in the regression model.

Since the features have different units of measure the data was centered and scaled via `train's` argument `preprocess(preProc=c("center,scale")` prior to model building. Due to non-normal distribution of some variables a log-transformation was performed on these features. Features computing probabilities were log-transformed. Data preparation was performed excluding the output variable `log(ON mode)`.

2.6.2 *Predictive Modeling*

Via the function `train` different tuning parameters can be specified that are used in model building, the model performance is calculated based on resampling [85]. The function `trainControl` enables to specify the resampling method, the number of folds and further arguments of the `train` function. Listing 1 shows an example of the methods `train` used with the function `trainControl`. Via `method` in `train` the different models can be specified. Seed was set to 100 for reproducable results. For further information on the package caret see [86].

Listing 1: Example of function trainControl and method train from caret used for predictive modeling

```
set.seed(100)
ctrl <- trainControl(method = "repeatedcv",
                     repeats = 5,
                     number = 10,

                     preProcOptions= list (cutoff=cor_cutoff),
                     returnResamp = "final",
                     savePredictions="all")

                     set.seed(100)
lm <- train(depVar~., data=data_switch_design,

             method = "lm",

             trControl = ctrl,

             preProc = c("center", "scale"))
```

From a statistical point of view the most efficient way to use the available data is to use all of it in model training and apply a resampling method like cross-validation or bootstrap to evaluate the model performance. Therefore all data was used in model training and repeated cross-validation was used as resampling with 5 repeats and 10 folds `trainControl(method="repeatedcv",repeats=5,number=10)`. Using a numeric output variable, data is split based on percentiles into groups sections [85]. Using k-fold crossvalidation it ensures that the test sets are independent since each sample is included only once in a fold [87].

In order to find out which algorithms could work well on the problem, a set of different methods was tried using the the package caret in R. The linear regression models linear regression (lm) and partial least squares (pls), further the penalized regression model elastic net (enet), and for non-linear regression support vector machine with radial basis function kernel (svmRadial) and some ensemble meth-

ods as bagged tree (treebag), random forest (rf) and stochastic gradient boosting (gbm) were tried.

To compare the performance of the different algorithms RMSE was computed for all algorithms. The different machine learning algorithms were evaluted by comparaing against the performance of a baseline model. Therefore, the mean value of the output variable log(ON mode) of each training set was calculated and assumed as predicted value. Then, the RMSE between this predicted value and the true values of the current test set was calculated, this procedure is performed k times. These values were used as baseline to evaluate the performance of the tested algorithms. Additionally R^2 was calculated. A good model hast to outperform the baseline. In order to determine whether the differences between two models are statistically significant a paired t-test was performed.

In order to access if there are problems with overfitting the training set error was computed. Therefore no resampling was performed (set `method='none'` in `trainControl`) and predictions were made on the training set, i.e. on the same set as the model was trained. Then, repeated cross-validation was performed with the basic feature set. In a next step a best subset regression was performed using `regsubsets` from the `leaps` package [88], the best subset by sequential replacement for a regression model was computed (when correlated and linear combinations are removed). This subset was used for model building. Finally, models were built with the single feature e_RBS_linker.

3

RESULTS AND DISCUSSION

This chapter summarizes the results and discussions of the analysis of first generation toehold switches designed by Green et al. [54]. The first generation library consists of 168 switches and triggers. The toehold switches are labelled according to their fluorescence ON/OFF ratio in descending order, meaning switch 1 - trigger 1 pair showed the highest fluorescence ON/OFF ratio and switch 168 - trigger 168 the lowest.

Prior to using the selected features (Section 3.1) in several machine learning algorithms (Section 3.5), the computed features were analysed regarding their descriptive statistics in Section 3.2. Additionally, the correlations between the features and the three possible dependent variables were further examined in Section 3.3. Because of its unexpected correlation properties the feature P_sw_acc_RBS_linker was examined in more detail in Section 3.4.

3.1 FEATURE SELECTION AND COMPUTATION

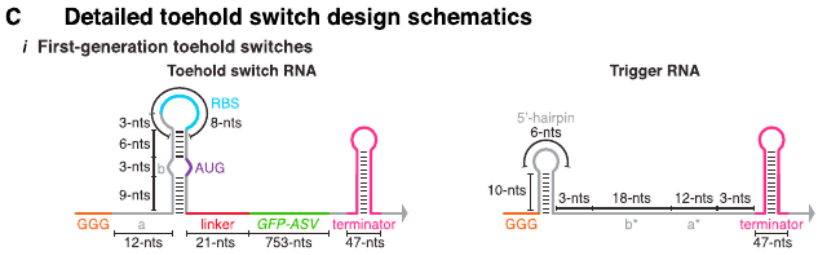
3.1.1 *Prior Work / Context*

Green et al. already performed an extensive analysis to determine thermodynamic parameters that are related to the ON/OFF ratio of the fluorescence output of the toehold switches. They used the ON/OFF ratio, since there was relatively little variation for OFF state fluorescence levels in comparison to the ON state levels. Thus, they argue, ON/OFF ratio essentially measures the ON state fluorescence (supplemental information of [54]). They used 48 thermodynamic parameters (list of parameters in Appendix A (Figure 25)) that were computed using NUPACK [28]. Since they found out that thermodynamic parameters calculated using Mathews et al. [89] experimentally determined energy parameters resulted in stronger correlations than using Serra and Turner [90] energy parameters they used Mathews et al. energy parameters for their analysis. They performed a linear regression analysis between $\log_{10}$ON/OFF ratios and each of the thermodynamic features. Significant correlations were only detected within subsets (that satisfied certain sequence criteria) of the first-generation switches (supplemental information of [54]). Only a single parameter, $\Delta G_{RBS-linker}$, exhibited a correlation for the subset of 68 switches with a closing A-U base pair at top of the hairpin stem. $\Delta G_{RBS-linker}$ is the free energy of the RNA sequence ranging from the RBS up to the end of the linker. A linear fit of this sub-

set of $\Delta G_{RBS-linker}$ to the corresponding $\log_{10}$ON/OFF ratios resulted in a coefficient of determination $R^2 = 0.40$. Green et al. also forward-engineered their switches, a linear fit of $\Delta G_{RBS-linker}$ of these switches to the ON/OFF ratios showed a much stronger correlation, $R^2 = 0.79$. They determined that $\Delta G_{RBS-linker}$ could best explain the switch performance [54].

3.1.2 *Adopted and Newly Created Features*

In this study, some features were adopted from Green et al. and additional were created. Table 1 shows an overview of all these features and their computation. In the following sections (Section 3.1.3, Section 3.1.4 and Section 3.1.5), their computation is described in detail.



Reproduction of Figure 3 from Section 1.4 to recall the definition of the toehold switch regions used for the feature computation in this section: Toehold switch design by Green et al. [54].

Unless otherwise noted, the switch RNA ranges from the GGG leader sequence up to 29 nucleotides of the GFPmut3b and the trigger RNA from the GGG leader, 5′ hairpin, up to the transcriptional terminator (corresponding with Green et al. used sequences).

The following features are adopted from Green et al.: e_RBS_linker corresponds to Greens' $\Delta G_{RBS-linker}$, e_tr_sw_toe corresponds to ΔG toehold binding, e_tr_sw_bs corresponds to ΔG trigger binding and e_net_complex corresponds to ΔG complex. ON/OFF ratios of the fluorescence, ON state fluorescence and OFF state fluorescence were used as output variables.

Feature	Description	Switch a	Switch b	Switch RBS	Switch linker	Trigger a*	Trigger b*	
e_RBS_linker^{G}	Ensemble free energy of the RBS-linker region in switch RNA	.	.	x	x			
e_tr_sw_toe^{G}	Ensemble free energy of the toehold region (a-a*) in the duplex formed by switch and trigger RNA	x	.	.	.	x	.	
e_tr_sw_bs^{G}	Ensemble free energy of the binding site (a-a* b -b*) in the duplex formed by switch and trigger RNA	x	x	.	.	x	x	
e_net_complex^{G}	Ensemble free energy of the trigger-switch-RNA duplex minus energy of the single-stranded trigger RNA and energy of the single-stranded switch RNA.	x	x	x	x	x	x	
		$X_{specific\_structure}$						x
P_sw_acc_RBS_linker	Probability that the RBS-linker region is accessible (stays unpaired) in the switch RNA.	.	.	x	x			-
P_sw_acc_toe	Probability that the 12 nt toehold region (a) is accessible (stays unpaired) in the switch RNA.	x	.	.	.			-
P_sw_acc_RBS_linker_bs_b	Probability that the RBS-linker region is accessible (stays unpaired) when the binding site (a+b) is bound in the switch RNA.	x	.	x	x			a+b
P_sw_hairpin	Probability that the switch RNA forms the hairpin.	.	hairpin		.			-
P_tr_acc_bs	Probability that the binding site (a*-b*) is accessible (stays unpaired) in the trigger RNA.					x	x	-
P_tr_acc_toe	Probability that the reverse complement to the toehold region (a*) is accessible (stays unpaired) in the trigger RNA.					x	.	-
P_tr_sw_acc_RBS_linker_bs_b	Probability that the RBS-linker region is accessible (stays unpaired) in the trigger-switch-duplex, when/after the binding site in the switch is bound by the trigger.	(	(	x	x	)	)	-
defect_dup	Ensemble defect of trigger-switch duplex RNA sequence and the target trigger-switch secondary structure.	x	x	x	x	x	x	
defect_sw	Ensemble defect of switch RNA sequence and the target toehold switch RNA secondary structure.	x	x	x	x			
defect_tr	Ensemble defect of trigger RNA sequence and the target trigger RNA secondary structure.					x	x	
defect_avg	Average of ensemble defect of defect_dup, defect_sw and defect_tr.	x	x	x	x	x	x	
ON mode	ON state fluorescence.							
OFF mode	OFF state fluorescence.							
ON/OFF mode	OF/OFF ratio of fluorescence.							

Table 1: Overview of defined features: Description of features and annotation of used regions for computation. The column 'Feature' shows the names of the defined features. There are three categories of features: ensemble free energy features, secondary structure probability features and ensemble defect features. The column 'Description' gives a short description of each feature. The columns 'Switch' and 'Trigger' illustrate which part(s) of the toehold switches were used for compuation. The coloring of the regions is in accordance with the graphic of the toehold switch design (Figure 3 and the reproduction above). '^{G}' marks that the ensemble free energy features are adopted from Green et al. For the ensemble free energy features an 'x' marks the used regions for computation, '.' the region is not used for computation. For the secondary structure features '$X_{specific\_structure}$' is the first constraint for the compuation and 'x' the second constraint, which is mostly unconstrained and marked with '-'. Only the feature P_sw_acc_RBS_linker_bs_b uses the region 'a+b' of the switch as second constraint. For these features 'x' marks that the region is constrained to stay unpaired, while '(' and ')' mark constraints to enforce basepairs. For the ensmble defect features 'x' marks the used regions. For detailed formula and input listings see Section 3.1.3 for the ensemble free energy features, Section 3.1.4 for the secondary structure probability features and Section 3.1.5 for the ensemble defect features. The constraints for the probability features are listed in the Appendix A (Section A.1.2).

3.1.3 *Ensemble Free Energy Features*

As mentioned before Green et al. used NUPACK [28] with Mathews et al. energy parameters for their calculations, whereas in this analysis the RNAlib (version 2.3.3) was used which uses Turner 2004 energy parameters [91] as energy model [25]. More precisely, python bindings were used to access the functions of the RNAlib.

To compute the energy of single stranded molecules the function `pf()` from the RNAlib was used. This function computes the partition function Z for a given RNA sequence, therefore it returns the Gibbs free energy ($G = -RT * \log(Q)$) of the ensemble in kcal/mol. RNAlib python objects of type `fold_compound` have `pf()` attached as method. Thus, to compute the partition function of an RNA sequence using RNAlibs' python bindings works as described in Listing 2.

Listing 2: Example of partition function computation of an RNA sequence using RNAlibs' python bindings

```
import RNA

sequence = "CGCAGGGAUACCCGCG"

# create new fold_compound object
fc = RNA.fold_compound(sequence)

# compute partition function
fc = RNA.fold_compound(sequence)
(pp, pf) = fc.pf() # pp pairing propensity, pf partition function

# print result
print(pp, pf)
```

`pf()` was used for the calculation of the energy of the single stranded RBS-linker region e_RBS_linker. The switch RNA sequence ranges from the first base of the loop up to the last base of the 21 nucleotide linker sequence.

The energy of double stranded molecules was computed with the partition function for dimers `pf_dimer()`, therefore the two sequences have to be linked with an ampersand (&). Otherwise the calculation follows the same procedure as for single stranded sequences. The features e_tr_sw_bs and e_tr_sw_toe were directly calculated using `pf_dimer()`. e_tr_sw_toe is the ensemble free energy of the 12 nucleotide long toehold region (region a) and its reverse complement (a*) in the duplex of switch and trigger RNA. e_tr_sw_bs is the ensemble free energy of the entire binding site, 30 nucleotides long (a - a*, b - b*) region, in the duplex.

The net complex energy was computed as the energy of the duplex of switch and trigger sequences minus the energy of the single stranded trigger and switch sequences (Net ΔG duplex = ΔG duplex - ΔG switch - ΔG trigger). The energy of the duplex was calcu-

lated with `pf_dimer()` and the energies of the single stranded RNAs each with `pf()` resulting in the energy of the feature e_net_complex. e_net_complex is the net free interaction energy between the switch and trigger RNA to build the activated switch complex.

3.1.4 *(Constrained) Secondary Structure Probability Features*

The next 7 features describe probabilities that certain secondary structures in the switch RNA, respectively the duplex of switch and trigger RNA are formed. The probability that a certain (sub)sequence folds into the desired structure ($P_{specific\_structure}$) is calculated by the fraction of the constrained ($Z_{X_{specific\_structure}}$) and the unconstrained partition function (Z_x) (or another constrained partition function).

$$P_{specific\_structure} = \frac{Z_{X_{specific\_structure}}}{Z_x} = e^{\frac{G_{X_{specific\_structure}} - G_X}{RT}}$$

where R is the gas constant with 1.98717kcal and T is the temperature in K. The calculations were performed using a temperature of 310.15K(37°C). The required structural constraints were set as hard constraint. (Therefore the required energies of the structures were calculated in advance.) The following features are based on the probabilities that certain structures form: P_sw_acc_RBS_linker, P_sw_acc_toe, P_sw_acc_linker_bs_b, P_sw_hairpin, P_tr_acc_bs, P_tr_acc_toe and P_tr_sw_acc_RBS_linker_bs_b. P_sw_acc_RBS_linker is the probability that the region starting from the RBS up to the end of the linker (RBS-linker region) stays unpaired in the switch RNA. P_sw_acc_toe is the probability that the 12 nucleotides long region a (the toehold) stays unpaired in the switch RNA. P_sw_acc_linker_bs_b is the probability that the RBS-linker region stays unpaired (and the binding site), when the binding site stays unpaired. P_sw_hairpin is the probability that the switch RNA forms the hairpin. P_tr_acc_bs is the probability that the binding site (a*-b*) stays unpaired in the trigger RNA. P_tr_acc_toe is the probability that the region a* stays unpaired in the trigger RNA. P_tr_sw_acc_RBS_linker_bs_b is the probability that the RBS-linker region stays unpaired in the trigger-switch-duplex, when the binding site in the switch and trigger basepairs. For further information on the used constraints see Appendix A (Section A.1.2).

3.1.5 *Ensemble Defect Features*

Green et al. used the ensemble defect as a stop condition for the optimization of the toehold switches with NUPACK. Therefore, also in this study the ensemble defect is used as a feature. It is calculated using a python interface to NUPACK 3.0 [82]. The ensemble defect is calculated for the duplex of switch and trigger (defect_dup), switch

and trigger alone (defect_sw and defect_tr) and the mean of the ensemble defect of the duplex, the switch and the trigger (defect_avg). Defect_dup is the ensemble defect calculated with each trigger-switch duplex RNA sequence and the target trigger-switch secondary structure as input. The ensemble defect of def_sw is calculated with the switch RNA sequence and the target toehold switch RNA secondary structure and def_tr with the trigger RNA sequence and the target trigger RNA secondary structure as input. defect_avg is the sum of def_dup, def_sw and def_tr dividey by three.

3.2 DATA ANALYSIS - DESCRIPTIVE STATISTICS AND VISUALIZATION

3.2.1 *Independent Variables (Features)*

The created features were analysed using descriptive statistics and visualization. Plots of the distribution or spread of features can help identify outliers or if a data transformation may be required. Features that might be redundant can be found by plots of the relationship between features and by computing the correlation coefficient R.

DATASET METRICS The dimensions of the dataset were accessed via `dim(switch_designs)`. For the toehold switches there are 168 instances with 18 attributes, 15 features and the 3 output variables (`ON mode, OFF mode, ON/OFF mode`). The data types of each variable can be listed with `sapply(switch_designs, class)`, all variables are numeric. The summary function also returns if there are missing values: in the dataset there are no missing values. Since there are no missing values no impuatation is necessary.

DATA DISTRIBUTION In order to access the distribution of each variable the `summary(switch_designs)` function was used. The numeric values are shown in the Appendix A (Table 10). There are listed the minimum value, value of 1st quartile, median value, mean value, value of 3rd quartile and maximum value. In this section, univariate visualizations of each feature variable were plotted to analyse and show the distribution, central tendency and spread. The distribution of the data is visualized in combined histograms and density plots of each feature (Figure 5). Additionally the mean and median are shown as lines for each feature the mean is red and median blue. The red line shows the mean value and the blue line the median.

The mean was greater than the median for: P_sw_acc_RBS_linker, P_sw_acc_toe, P_sw_acc_RBS_linker_bs_b, P_tr_sw_acc_RBS_linker_bs_b, and minimally higher for defect_sw, defect_tr and defect_avg which indicates a right skew of the distribution of the data (see Figure 5a-g). Similar mean and median indicate the data are symmet-

ric. A mean less than the median indicates that the distribution of the data is skewed to the left. The mean is less than the median for e_RBS_linker, e_tr_sw_bs, e_net_complex, P_sw_hairpin, P_tr_acc_bs, P_tr_acc_toe and defect_dup (see Figure 5h-n), but overall the differences between the mean and the median are not that high as for the right skewed data.
For e_tr_sw_bs, e_net_complex, P_sw_hairpin and P_tr_acc_bs and defect_dup the skew is very small, they can be assumed as unimodal normally distributed. The feature e_tr_sw_toe has a multimodal distribution (see Figure 5o).

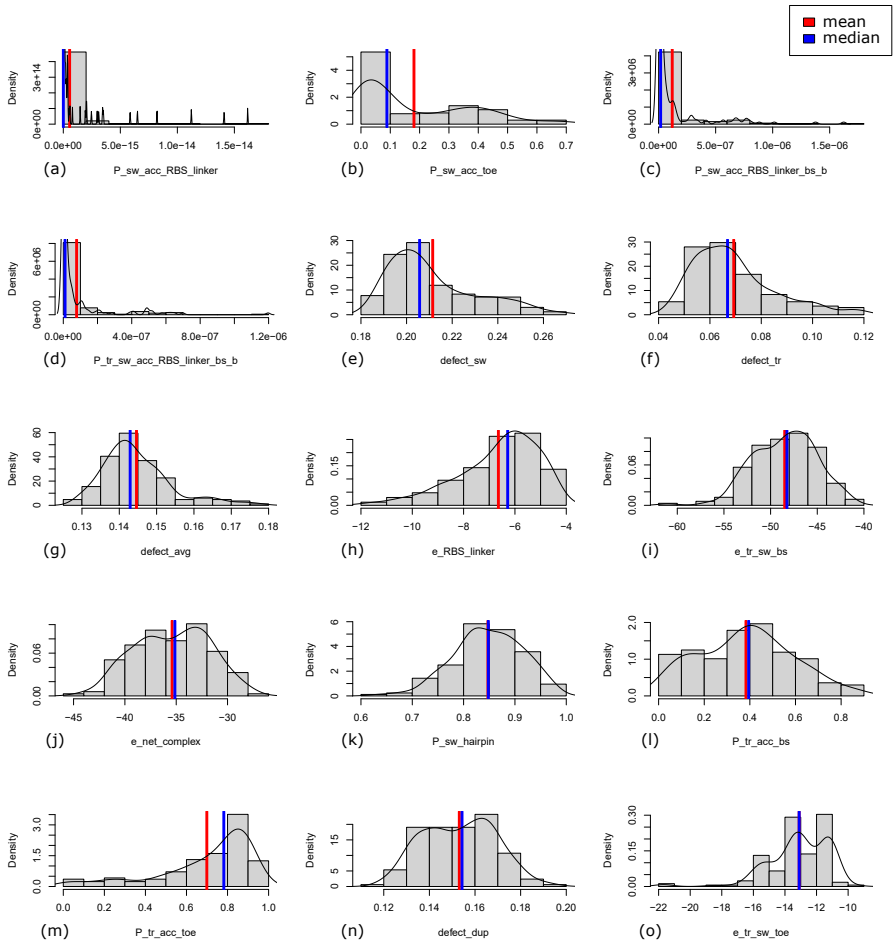


Figure 5: Univariate visualization of each feature: For each feature a combined histogram and density plot is shown. The red line shows the mean value for each individual feature and the blue line the median. a-g: mean greater than the median. h-n: mean less than the median. o: mean and median coincide.

Figure 6 shows the box plot of each feature. A boxplot is a visualization of the five-number summary. The box ranges from the first quartile to the third. Minimum and maximum are at the end of the whiskers. The median is the line in the box. Dots outside the whiskers are potential outliers. Here, the boxplots are used for outlier detection.

Probability features P_sw_acc_RBS_linker, P_sw_acc_RBS_linker_bs_b, P_tr_acc_toe and P_tr_sw_acc_RBS_linker_bs_b have many potential outliers indicated by the dots outside of the whiskers, also defect_avg has many potential outliers. e_RBS_linker, e_tr_sw_bs, e_tr_sw_toe, P_sw_hairpin, defect_sw and defect_tr have potential outliers but much less than the former mentioned features.

The probability features with many potential outliers have upper outliers, P_sw_hairpin and P_tr_acc_toe have much less potential outliers, P_sw_hairpin just one, and those are lower ones.

The energies e_RBS_linker, e_tr_sw_bs and e_tr_sw_toe have only lower potential outliers. The ensemble defect features defect_sw, defect_tr and defect_avg have just upper outliers.

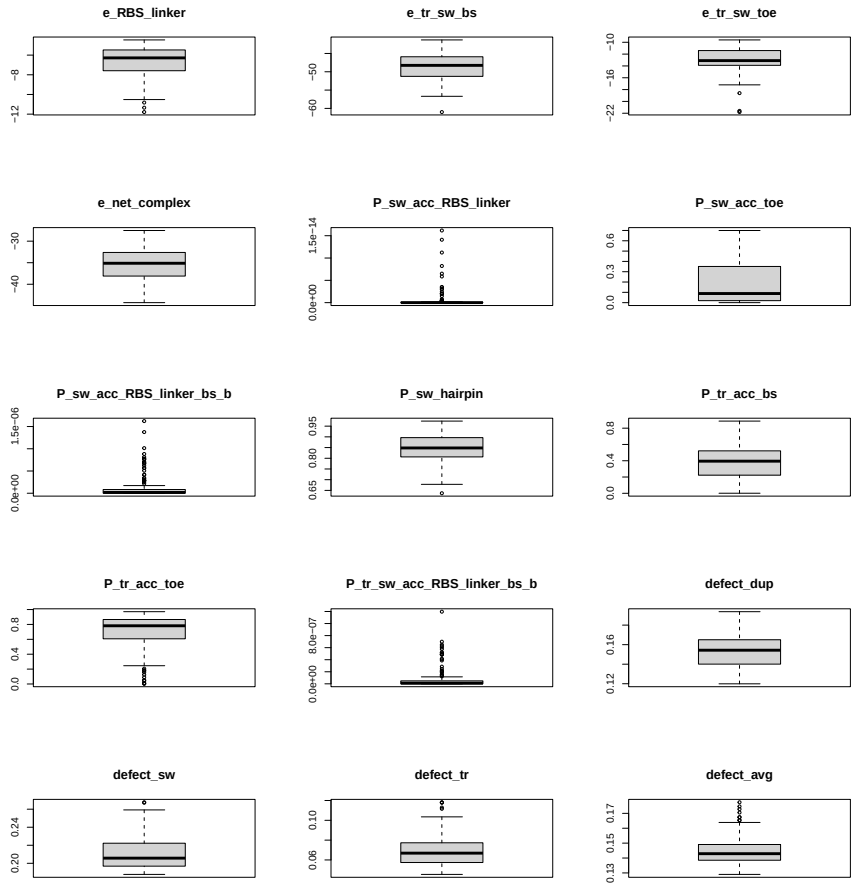


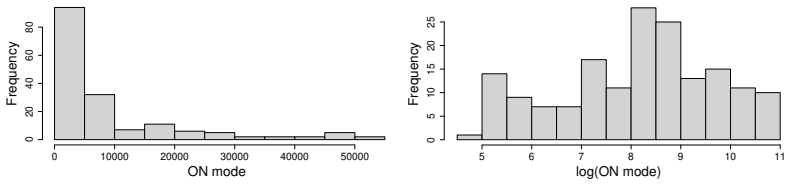
Figure 6: Box plot of each feature: A boxplot is a visualization of the five-number summary. The box ranges from the first quartile to the third. Minimum and maximum are at the end of the whiskers. The median is the line in the box. Dots outside the whiskers are potential outliers.

FEATURE-FEATURE CORRELATIONS Computing feature-feature correlations, the following features had a correlation above 0.75:

P_tr_acc_toe and P_tr_acc_bs show a strong positive correlation with $R = 0.77$. P_sw_acc_RBS_linker_bs_b and P_tr_sw_acc_RBS_linker_bs_b show a strong correlation with $R = 0.96$. Some models, which are later on trained for predictive modeling, can be affected by correlated features. Linear models produce unstable solutions or show poor performance, while other models such as regression trees are unsusceptible to highly correlated features, but multicolinearity can complicate interpretability. Therefore, correlated features with a value greater than 0.75 were removed.

3.2.2 *Dependent Variable ON Mode*

A histogram of the dependent variable ON mode is shown in Figure 7a. ON mode is strong right skewed and all values are positive, therefore ON mode was logarithmically transformed to reduce right skewness and achieve a nearly normal distribution. Figure 7b shows the histogram of the logarithmically transformed ON mode. The predictive models built in (Section 3.5) were trained to predict the log(ON mode), since the correlation analysis (see Section 3.3) showed higher correlations of the features to the log(ON mode) fluorescence than to the other output variables. ON/OFF fluorescence and OFF state fluorescence also show a right skew and therefore are logarithmically transformed for the correlation analysis.



(a) non-transformed ON mode (b) log-transformed ON mode

Figure 7: Histograms of dependent variable (ON mode): (a) non-transformed ON mode (b) log-transformed ON mode

3.2.3 *Conclusion*

Concerning the distributions of the features, all features that compute probabilities were logarithmically transformed (most of them showed a right skew and therefore will also be nearly normally distributed) and a logarithmic transformation will maintain the probabilities positive. The dependent variable ON mode was also logarithmically transformed. Features with a correlation coefficient R greater than 0.75 were removed for the predictive models.

3.3 CORRELATION

To gain further insights into possible, respectively expected relationships between the developed individual features (in Section 3.1) and the output variables (ON mode, OFF mode, ON/OFF mode), all pairwise correlations were calculated for the first generation toehold switches. Due to the right skewed distribution of all output variables they were logarithmically transformed. Additionally, all probability features were logarithmically transformed: P_sw_acc_RBS_linker, P_sw_acc_toe, P_sw_acc_linker_bs_b, P_sw_hairpin, P_tr_acc_bs, P_tr_acc_toe and P_tr_sw_acc_RBS_linker_bs_b. In the further analysis, the logarithmically transformed output variables and the logarithmically transformed probability features are used. Table 11 in the Appendix A shows the numerical values of the correlations of the individual features with the log(dependent variables), sorted by decreasing R of the feature correlations with log(ON mode). In the further analysis mainly the correlations between log(ON mode) and log(OFF mode) fluorescence will be discussed, since the relation between log(ON) and log(OFF mode) and the individual features can be easier understood, than the relation to log(ON/OFF). Therefore, in this section Figure 8 is used to illustrate the correlation coefficient R with log(ON mode) and log(OFF mode). The height of the bars represents R, red bars show the correlation of the log(OFF mode) and blue of log(ON mode) with the individual features. The coefficient of determination R^2 of each feature with the output variables is listed beside the bars. The features are ordered from top to bottom by decreasing absolute value of R of log(ON mode) with each feature.

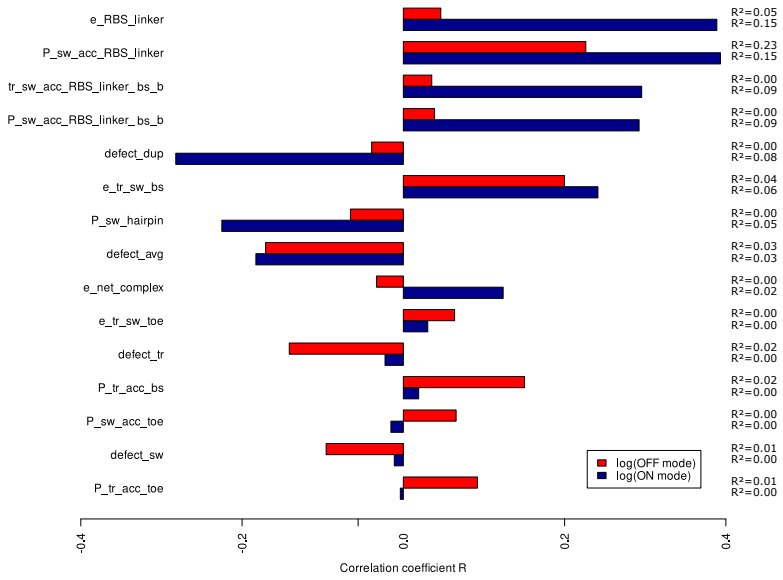


Figure 8: Graphical representation of pairwise correlation coefficients R of log(ON mode) and each individual feature, and log(OFF mode) with each individual feature. Barplot of R of log(OFF mode) with each feature is shown in red, R of log(ON mode) with each feature in blue. Coefficient of determination R^2 between each feature and log(OFF mode), respectively log(ON mode) is listed beside each feature. The features are ordered from top to bottom by decreasing absolute value of R of log(ON mode) with each feature.

Figure 9 shows a scatter plot of the individual features with the log(ON mode) to display the relationships, which will be discussed in the following section. The red line shows the regression line of each feature with log(ON mode).

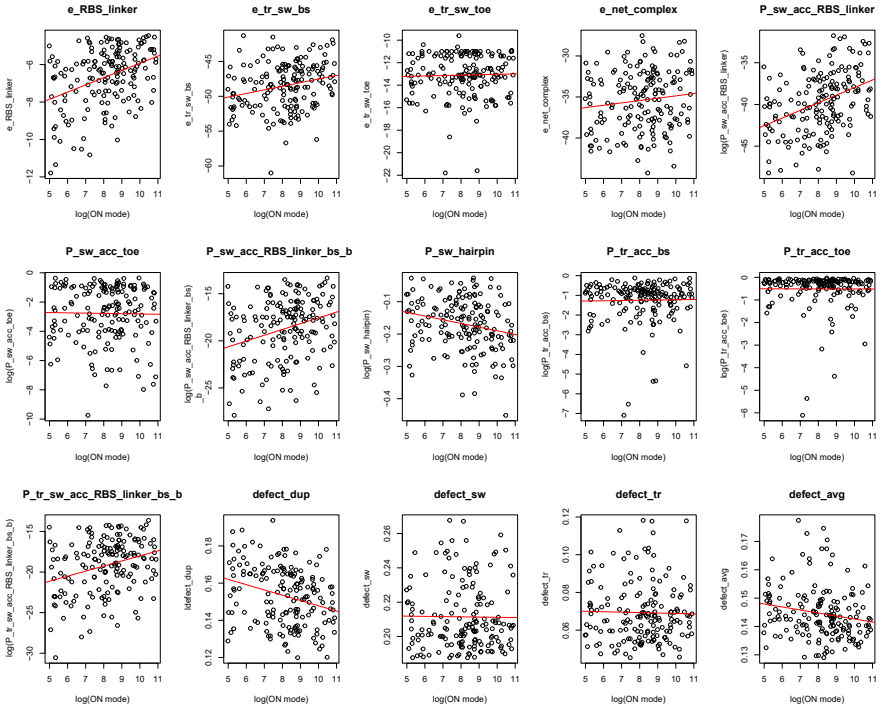


Figure 9: Correlation scatter plot of the individual features with log-transformed dependent variable (ON mode) (and log transformed probability features). The red line shows the linear regression line for each feature pair.

3.3.1 *Correlations between independent variables (features) and dependent variable (output variable)*

Green et al. found the highest correlation between $\Delta G_{RBS-linker}$ and the $\log_{10}$ON/OFF ratio, they report R^2=0.40 [54]. Also in this correlation analysis the log(ON/OFF) state correlates highest with e_RBS_linker. For the log(ON/OFF) ratio R=0.38 and R^2=0.15 (Appendix A, Table 11). For the log(ON mode) the correlation is minimal higher R=0.39, R^2=0.15 and it also correlates highest with the log(ON). For the log(OFF mode) R=0.05 and R^2=0.00 (11th ranked). e_RBS_linker describes the state when the trigger is bound and the RBS is accessible, by only calculating the ensemble free energy of the loop in the hairpin with the RBS up to the linker. When the trigger is bound the remaining sequence (in 3' direction) is unstructured and ΔG gets higher. The more unstructured the RBS-linker region thus the higher the free energy of this region and therefore the higher

the log(ON mode) (see Figure 9), respectively the log(ON/OFF). As expected e_RBS_linker correlates strongly with the log(ON mode), respectively the log(ON/OFF). There is no correlation with the log(OFF mode).

P_sw_acc_RBS_linker correlates at the same level as e_RBS_linker with the log(ON mode), R=0.39 and R^2=0.15. P_sw_acc_RBS_linker correlates with the log(OFF mode) with R=0.23 and R^2=0.05 and has the highest correlation between an individual feature and the log(OFF mode). Since this feature computes the probability that the RBS-linker region stays unpaired in absence of the trigger this feature should stronger correlate with the log(OFF mode) than the log(ON mode). But P_sw_acc_RBS_linker correlates positive with either state and even higher with the log(ON mode) (see Figure 9 for positive log(ON mode) correlation). It could be possible that the designs form such a strong hairpin that there is no base fluorescence in the log(OFF mode). Furthermore the hairpin could be so stable, that it even does not open in the presence of the trigger and there is no switching to the log(ON mode). In order to gain further insights in the relationship between this feature and the log(ON mode) some switch designs were further analysed regarding their secondary structure (see Section 3.4).

P_tr_sw_acc_RBS_linker_bs_b computes the probability that the RBS-linker region stays unpaired in the duplex of switch and trigger RNA when the binding site is paired. Therefore this feature should correlate stronger with the log(ON mode). P_tr_sw_acc_RBS_linker_bs_b correlates with the log(ON mode) with R=0.30 and R^2=0.09 and with the log(OFF mode) with R=0.04 and R^2=0.00. As expected there is a quite strong correlation with the log(ON mode) and no correlation with the log(OFF mode).

P_sw_acc_RBS_linker_bs_b computes the probability that the RBS-linker region and the binding site stays unpaired in the switch RNA when the binding site is constrained to stay unpaired. The trigger is absent. The RBS would then be accessible if the hairpin opens spontaneously, but this is no reaction that is targeted. This feature correlates at similar level as P_tr_sw_acc_RBS_linker_bs_b strong with the log(ON mode) R=0.29 and R^2=0.09 and very low with the log(OFF mode) R=0.04 and R^2=0.00. In absence of the trigger this feature should show a stronger correlation to the log(OFF mode).

P_sw_acc_toe computes the probability that the complete 12 nucleotide long toehold region a stays unpaired. A good toehold supports the switching behaviour in the switch, therefore we would expext a relationship with the log(ON) mode rather than the log(OFF) mode. But in this context the feature is not informative, the correlations are very low. The correlation is very low with the log(ON mode) is $R = -0.02$ and R^2=0.00 and with the log(OFF mode) $R = -0.07$ and R^2=0.00.

P_sw_hairpin computes the probability that the hairpin forms in the switch, therefore the probability should be near 1, if the switch forms as designed. According to the non-logarithmically transformed data (see Table 10), the probabilities range from 0.6368 to 0.9736 with a mean of 0.8474. All designs have a high probability that the hairpin forms. The hairpin module in the switch functions as a translation repressor. When the hairpin unwinds, translation can take place, by making the RBS and start codon accessible. P_sw_hairpin correlates negatively with the log(ON mode), $R = -0.23$ and R^2=0.06 (see Figure 9). The more unlikely the hairpin forms, the higher the log(ON mode). The feature correlates, as expected, rather high negatively with the log(ON mode) and rather low with the log(OFF mode), $R = -0.07$ and R^2=0.00.

The next two features compute accessibilities in the trigger RNA: P_tr_acc_toe computes the probability that region a* is accessible and P_tr_acc_bs that the entire binding site, consisting of a* and b*, is accessible. We would expect that these features correlate with the log(ON mode), since there is no trigger available in the log(OFF mode). Both features show weak correlation with the log(ON mode), while P_tr_acc_bs shows weak correlation with the log(OFF mode), and P_tr_acc_toe a very weak correlation.

The feature defect_dup computes the ensemble defect of the duplex of switch and trigger RNA. The closer the duplex structure to the target the smaller is the ensemble defect. Therefore a small ensemble defect (near 0) should correlate with the log(ON mode). Defect_dup correlates with the log(ON mode) with $R = -0.28$ and R^2=0.08, whereas there is virtually no correlation with the log(OFF mode), $R = -0.04$ and R^2=0.00.

e_tr_sw_bs is the free energy of the 30 nucleotide long binding site in the duplex formed by switch and trigger RNA. It computes the free energy of a designed duplex, which should (almost) perfectly match. Indirectly, this feature describes the GC-content. The positive correlation for the log(ON mode) is R=0.24, R^2=0.06, for the log(OFF mode) R=0.20, R^2=0.04 and for the log(ON/OFF) state R=0.20, R^2=0.04. The higher the free energy the higher all output variables get, whereas correlation is highest for the log(ON mode). This feature correlates rather similar with all states, the higher the free energy the higher each of the output variables. The reasons are unclear.

defect_avg correlates negatively to all output state, at similar level to log(ON) ($R = -0.18$) and log(OFF) ($R = -0.17$), and to log(ON/OFF) with $R = -0.13$.

e_net_complex correlates with R=0.12 and R^2=0.02 with the log(ON mode). e_net_complex is the net free interaction energy between the switch and trigger RNA to build the activated switch complex. The activated complex correlates to the log(ON mode), the higher the energy, the higher the log(ON mode).

3.3.2 *Conclusion*

As expected, e_RBS_linker correlates strongly with the log(ON mode), respectively the log(ON/OFF) ratio. The features P_tr_sw_acc_RBS_-linker_bs_b and P_sw_hairpin correlate as expected with the output variable.
P_sw_acc_RBS_linker, P_sw_acc_RBS_linker_bs_b and e_tr_sw_bs exhibit unexpected correlations to the log(ON mode). P_sw_acc_RBS_linker unexpectedly correlates highest with all output variables and was therefore selected for further analysis. In the next section (Section 3.4) some secondary structures were analysed in an attempt to understand the correlation behaviour of this feature. For other features such as P_sw_acc_toe the correlations are very low and are not informative.

3.4 FEATURE P_SW_ACC_RBS_LINKER - STRUCTURE AND ACCESSIBILITY

The feature P_sw_acc_RBS_linker was chosen for further analysis due to the strong correlation to all output variables (ON mode R=0.39 and log(OFF) mode R=0.23, see Section 3.3.1). Therefore, the secondary structures of some toehold switches were predicted in order to examine if there is a structure-dependent explanation for the correlation to all output variables. As explained P_sw_acc_RBS_linker computes the probability that the RBS-linker region stays unpaired in the switch RNA in absence of the trigger, therefore the switch does not get activated and there should be no log(ON mode) fluorescence and thus no correlation to the log(ON mode). But, the feature should correlate with the log(OFF) mode fluorescence as it does. But the correlation is higher to the log(ON mode) fluorescence. The translation repressing hairpin could be so strong that there is no base fluorescence in the log(OFF mode) and no switching behaviour in the switch in presence of the trigger resulting in the unusual correlation.

At first secondary structures of switch RNAs alone were analysed, then some switch RNAs were analysed with their cognate trigger, since there were no obvious structures from the switch RNAs alone that explain the correlations to all output variables. Figure 10 shows the scatter plot of log(P_sw_acc_RBS_linker) and log(ON mode). The further analysed switches are colored and labeled with their number (according to their ON/OFF performance in decreasing order). Switches with low accessibility and low log(ON mode) are red colored (switches 161, 141, 140 and 164). Switch 17 exhibits the minimum accessibility (but high log(ON mode)) dark red colored. Switches with best/worst log(ON/OFF) are blue colored (switches 1, 2 and 168). Switches with lowest and highest log(ON mode) are orange colored (switch 165 and 5). Switches with low accessibility and moderate

log(ON mode) are green colored (switches 66 and 84). The secondary structures of the following switches with their cognate triggers were also computed: 1, 17, 141, 164, 165 und 168.

In the following, these switches are discussed. For more detailed structure analysis see Appendix A (Section A.4 and Section A.5).

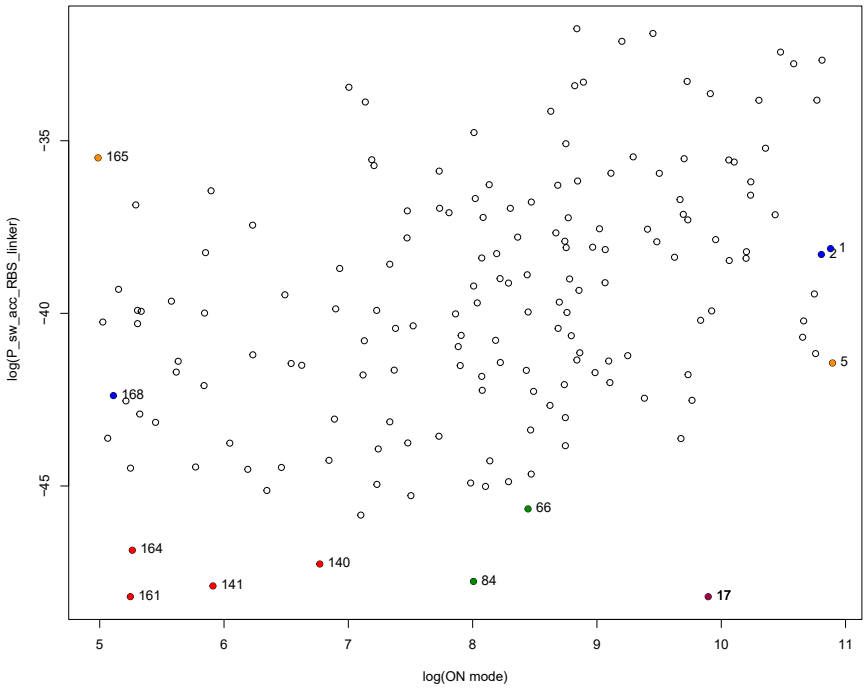


Figure 10: Scatter plot of log(P_sw_acc_RBS_linker) and log(ON mode), switches with low accessibility and low log(ON mode) (red circle), minimum accessibility (but high log(ON mode)) (dark red circle), best/worst log(ON/OFF) (blue circle), lowest and highest log(ON mode) (orange circle), low accessibility and moderate log(ON mode) (green circle)

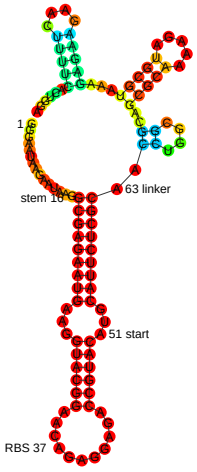
Table 2 shows the summary of the secondary structure prediction. Important positions that were analysed are the region a (toehold), which ranges from nucleotide 4 - 15, region b, which is at the bottom of the hairpin stem and ranges from nucleotide 16 - 33. The RBS starts at position 37 and ranges up to 44. The startcodon is at position 51. The linker region starts at 63.

Switch	Monomer			Duplex		Comment
	a	b	RBS unpaired	AUG unpaired	alteration	
161	last nt pairs with linker	1nt elongated		NC		low access.
141	x	x	-	-	x	low access.
140	x	x		NC		low access.
164	last nt pairs with linker	1nt elongated	+	-	x	low access.
17	pairs with linker region	x	-	x	x	minimal access.
1	folds back onto linker, truncated	elongated	+	x	x	best ON/OFF
2	x	x		NC		second ON/OFF
168	x	x	-	-	-	worst ON/OFF
165	folds back onto linker	x	x	x	x	lowest ON mode, high access.
5	truncated	elongated		NC		highest ON mode, moderate access.
66	truncated	elongated		NC		low access., moderate ON
84	x	x		NC		low access., moderate ON

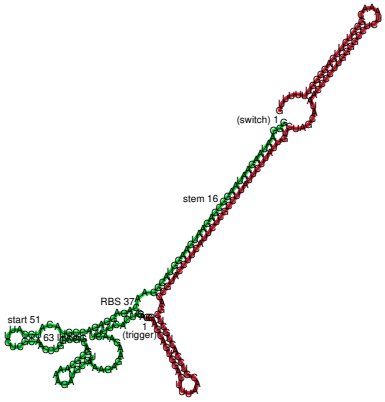
Table 2: Summary of secondary structure analysis. For the switch monomer there are the categories 'a' and 'b'. Since RBS and start codon were always at designed positions, they are not listed. For the duplex of switch and trigger RNA the regions 'RBS' and 'start codon' are listed, and the column 'alteration' marks if there are additional structure modifications. The last column 'comment' explains what is distinctive for the chosen switches. NC marks that the secondary structure was not calculated. A 'x' marks that the designed target structure is formed, a '-' that there are alterations or there is a further explanation. NC (not computed) marks that the secondary structure was not calculated.

3.4.1 *Low accessibility of RBS-linker region and low log(ON mode)*

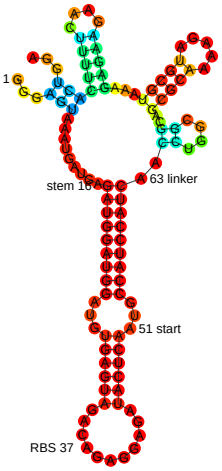
First some switches with low accessibility and low log(ON mode) were chosen for single stranded RNA secondary structure prediction: the secondary structures of the switches 161, 141, 140, 164 were investigated. The switches 141 (Figure 11a) and 140 (Figure 11c) form the target structure, whereas switches 164 (Figure 12a) and 161 (Figure 12c) do not. In the dimer secondary structure of switch 141 (Figure 11b) neither the RBS nor the AUG are unpaired, whereas in switch 164 (Figure 12b) at least the RBS is unpaired, but not the start codon.



(a) sw141

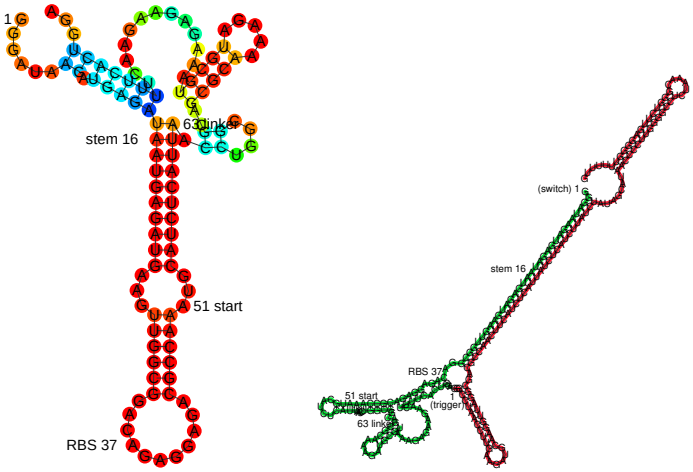


(b) sw+tr141

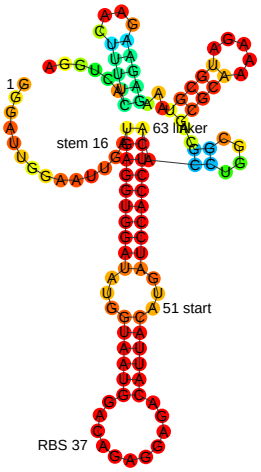


(c) sw140

Figure 11: Centroid structures of switches 141 and 140: Both switches have a low RBS-linker region accessibility and low log(ON mode). (a) Secondary structure of switch monomer 141 (b) Secondary structure of switch-trigger duplex 141 (c) Secondary structure of switch monomer 140
Some positions are labelled for orientation: '1' ist the first nt. The other numbers label the intended positions of certain domains, which can differ from the actual secondary structure. 'stem 16' the stem of the hairpin starts at nt 16. 'RBS 37' the RBS starts at nt 37. '51 start' the start codon is at nt 51. '63 linker' the linker region starts at nt 63.



(a) sw164 (b) sw+tr164

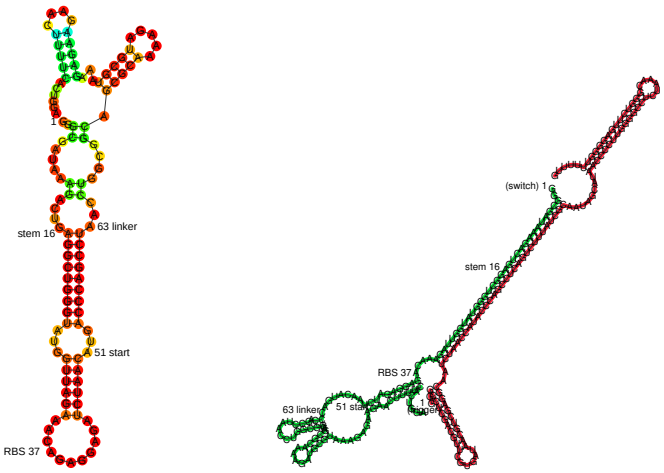


(c) sw161

Figure 12: Centroid structures of switches 164 and 161: Both switches have a low accessibility and low log(ON mode). (a) Secondary structure of switch monomer 164 (b) Secondary structure of switch-trigger duplex 164 (c) Secondary structure of switch monomer 161
Some positions are labelled for orientation: '1' ist the first nt. The other numbers label the intended positions of certain domains, which can differ from the actual secondary structure. 'stem 16' the stem of the hairpin starts at nt 16. 'RBS 37' the RBS starts at nt 37. '51 start' the start codon is at nt 51. '63 linker' the linker region starts at nt 63.

3.4.2 *Minimal accessibility of RBS-linker region and high log(ON mode)*

Switch 17 (Figure 13a) was also analysed, since this switch has the lowest accessibility overall, although it has a high log(ON mode). The secondary structure of switch 17 in absence of the trigger does not form the target structure. According to the secondary structure of switch 17 the toehold region forms unintended basepairs, which might result in a bad accessibility for the trigger, whereas the RBS and the startcodon are accessible as intended. And, the overall functionality of the switch is good. In the predicted secondary structure of switch 17 with its trigger (Figure 13b) the RBS is not unpaired (7 of 8 nucleotides are paired), but the start codon is unpaired.

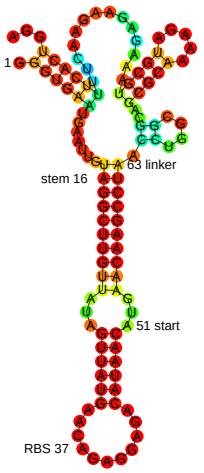


(a) sw17 (b) sw+tr17

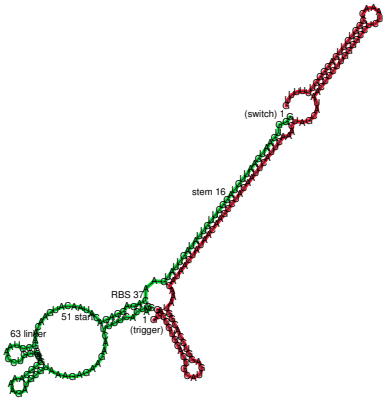
Figure 13: Centroid structures of switch 17: This switch has minimal accessibility but high log(ON mode). (a) Secondary structure of switch monomer 17 (b) Secondary structure of switch-trigger duplex 17 Some positions are labelled for orientation: '1' ist the first nt. The other numbers label the intended positions of certain domains, which can differ from the actual secondary structure. 'stem 16' the stem of the hairpin starts at nt 16. 'RBS 37' the RBS starts at nt 37. '51 start' the start codon is at nt 51. '63 linker' the linker region starts at nt 63.

3.4.3 *Lowest and highest log(ON/OFF mode)*

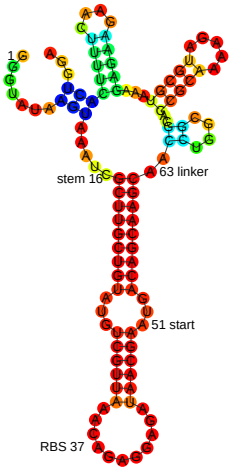
Additionally, some switches with distinguishing charcteristics were analysed. Switch 1 and switch 2 were chosen for their overall best log(ON/OFF) and switch 168 for the worst log(ON/OFF). According to the structure prediction of switch 1 without trigger it does not form the target structure (Figure 14a). But in the dimer structure the RBS and the start codon are unpaired (Figure 14b). Switch 2 forms the target structure when only the switch RNAs secondary structure is predicted (Figure 14c). The single stranded secondary structure of the switch 168 forms the target structure, but in the duplex neither the RBS nor the start codon are unpaired (Figure 15a and Figure 15b).



(a) sw1

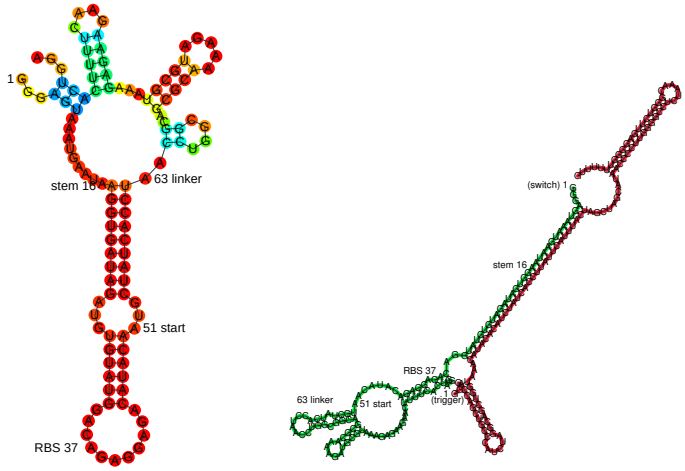


(b) sw+tr1



(c) sw2

Figure 14: Centroid structures of switches 1 and 2: Switch 1 has the best (highest) log(ON/OFF) and switch 2 the second best log(ON/OFF) (a) Secondary structure of switch monomer 1 (b) Secondary structure of switch-trigger duplex 1 (c) Secondary structure of switch monomer 2
Some positions are labelled for orientation: '1' ist the first nt. The other numbers label the intended positions of certain domains, which can differ from the actual secondary structure. 'stem 16' the stem of the hairpin starts at nt 16. 'RBS 37' the RBS starts at nt 37. '51 start' the start codon is at nt 51. '63 linker' the linker region starts at nt 63.

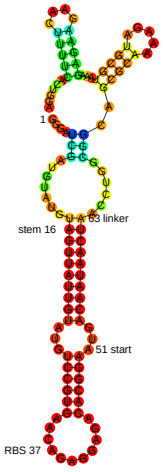


(a) sw168 (b) sw+tr168

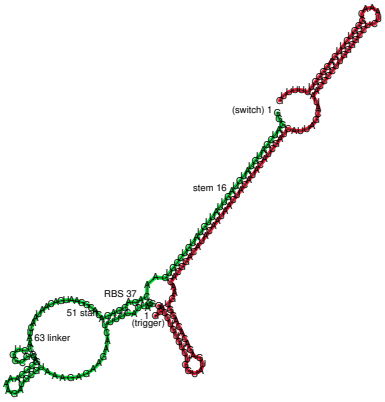
Figure 15: Centroid structures of switch 168: This switch has the worst (lowest) log(ON/OFF). (a) Secondary structure of switch monomer 168 (b) Secondary structure of switch-trigger duplex 168
Some positions are labelled for orientation: '1' ist the first nt. The other numbers label the intended positions of certain domains, which can differ from the actual secondary structure. 'stem 16' the stem of the hairpin starts at nt 16. 'RBS 37' the RBS starts at nt 37. '51 start' the start codon is at nt 51. '63 linker' the linker region starts at nt 63.

3.4.4 *Lowest and highest log(ON mode)*

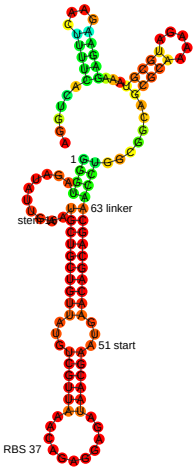
Switch 165 has the lowest log(ON mode) (but high accessibility) and switch 5 has the highest log(ON mode) and a mediocre accessibility. Switch 165 alone does not form the target structure (Figure 16a), but in the dimer (Figure 16b) the RBS and the start codon are unpaired. Switch 5 without trigger does not form the target structure (Figure 16c).



(a) sw165



(b) sw+tr165

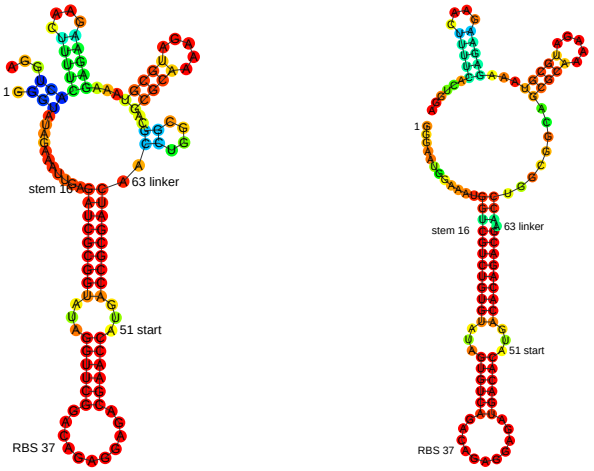


(c) sw5

Figure 16: Centroid structures of switches 165 and 5: Switch 165 has the lowest log(ON mode) and high accessibility. Switch 5 has the highest log(ON mode) (and moderate accessibility). (a) Secondary structure of switch monomer 165 (b) Secondary structure of switch-trigger duplex 165 (c) Secondary structure of switch monomer 5 Some positions are labelled for orientation: '1' ist the first nt. The other numbers label the intended positions of certain domains, which can differ from the actual secondary structure. 'stem 16' the stem of the hairpin starts at nt 16. 'RBS 37' the RBS starts at nt 37. '51 start' the start codon is at nt 51. '63 linker' the linker region starts at nt 63.

3.4.5 *Low accessibility and mediocre log(ON mode)*

Switches 66 and 84 were also analysed due to their low accessibility and mediocre log(ON mode) fluorescence. Switch 66 (Figure 17b) without trigger does not form the target structure, wheres switch 84 does (Figure 17a).



(a) sw84 (b) sw66

Figure 17: Centroid structures switches 84 und 66: Both switches have a low accessibility and mediocre log(ON mode). (a) Secondary structure of switch monomer 84 (b) Secondary structure of switch monomer 66
Some positions are labelled for orientation: '1' ist the first nt. The other numbers label the intended positions of certain domains, which can differ from the actual secondary structure. 'stem 16' the stem of the hairpin starts at nt 16. 'RBS 37' the RBS starts at nt 37. '51 start' the start codon is at nt 51. '63 linker' the linker region starts at nt 63.

3.4.6 *Conclusion*

Due to the examined switch RNA secondary structures it is not obvious why the feature P_sw_acc_RBS_linker shows a strong correlation with the log(ON mode). There are switches with low accessibility as switch 141 and 164, but according to their secondary structure prediction in absence of their triggers, 141 forms the target structure, while switch 164 does not, the region a forms a basepair with the linker. Switch 1 has the best log(ON/OFF) fluorescene and does not form the target structure in absence of the trigger. But in the dimer RBS and start codon are accessible. Switch 165 has one of the low-

est log(ON/OFF) fluorescences but shows same secondary structure properties. Switch-trigger 17 which shows a good fluorescence has an almost completly paired RBS in the duplex, but has a high fluorescence level, and the switch alone does not form the target structure. According to the secondary structures it is not obvious why some of the duplexes show a high fluorescence level while the others do not.

3.5 MACHINE LEARNING - BUILDING PREDICTIVE MODELS

Based on the the created thermodynamic features (see Table 1) different algorithms were trained to build models that predict the fluorescence level of an RNA sequence design. Data pre-processing before model building can be required depending on the type of model used for prediction, or improve predictions. Some models as linear regression are sensitive to characteristics of the predictor data, while tree-based methods are not [92]. Due to the gained insights from exploration and visualization of the data (see Section 3.2) pre-processing of the data was performed: based on the plot (Figure 5) the features that compute probabilities were logarithmically transformed (as for the correlation analysis Section 3.3). Further, correlated features were removed, a cutoff of 0.75 was used and linear combinations were removed. The data was also centered and scaled to standardize the features. RMSE was used for optimization in the model training process. RMSE has the advantage that it has the same unit as the dependent variable. Therefore a baseline RMSE was computed, the mean value of the output variables of log(ON mode) of each training set was calculated and assumed as predicted value. Then, the RMSE between this predicted (mean) values and the true values of the current test set were calculated; this procedure is performed k times. These values were used as baseline to evaluate the performance of the tested algorithms. A good model has to outperform the baseline. In order to determine whether the differences between two models are statistically significant a paired t-test was performed. The t-test was computed for the pairs of the k folds between the baseline and the different algorithms (significance level $\alpha = 0.05$).

PREDICTING LOG(ON MODE) Models were trained with different feature sets to predict the log(ON mode). log(ON mode) fluorescence was used since the correlations between the single features and the log(ON mode) were higher than those for the log(ON/OFF) fluorescence. Furthermore it is easier to interprete the influence of features on the log(ON mode) rather than a ratio.

3.5.1 *Training set error*

In order to find out if there are potential problems with overfitting the training set error was calculated. Therefore all data was used to train models, and predictions were made on the same data set. As metrics RMSE and R^2 were calculated (listed in Table 3). The RMSE is used to compare the models. Random forest and bagged tree have the lowest RMSEs, when the training set error is calculated. The RMSE for random forest is 0.63 (R^2 0.93), for bagged tree RMSE is 1.00 (R^2 0.68). The metrics for the linear regression model are also reported

here, since it showed one of the best performances of all tried algorithms in all settings. RMSE of the linear model is 1.30 and R^2 0.31. If we now consider the R^2, which is a measure for how much of the variance is explained, random forest can explain 93% of the variation, bagged tree 66% and the linear model can only explain 31% of the variation. But compared to the results, when k-fold cross-validation without further tuning is applied (see Table 4), both random forest and bagged tree algorithms have the highest RMSE with 1.4666 and 1.4683, respectively, and the lowest R^2 (0,1704 and 0.1807). The linear regression model has an RMSE of 1.3802 and R^2 of 0.2653, the training set error and the test set error are much closer than those for random forest and bagged tree. Random forest and bagged tree perform well when predictions are made on the trained data again, but do not predict well on new data, whereas the linear model has similar estimates for the training set error and the test set error computed by using k-fold cross-validation. In general, more complex models such as random forest and bagged tree can have high variance, which causes over-fitting. In contrast simpler models as linear regression models can have high bias, they have a tendency to under-fit due to unflexibility. The model can not capture the true relationship [92]. But, when comparing the results using k-fold crossvalidation, the linear model seems to predict quite well.

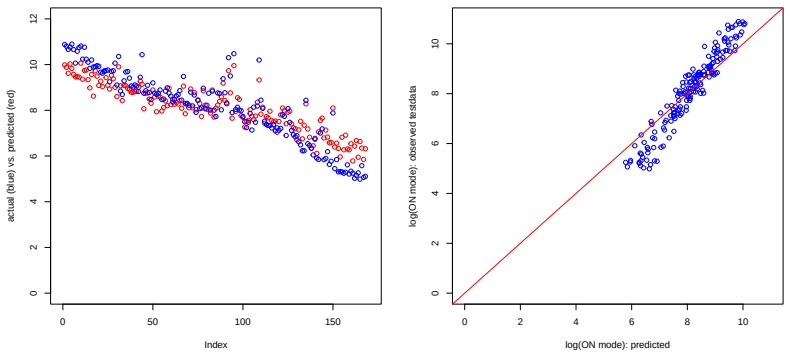
	RMSE	R^2
lm	**1.3014**	**0.3144**
pls	1.4119	0.1930
enet	1.4986	0.2098
svmRadial	1.3087	0.3860
treebag	1.0124	0.6619
rf	**0.6286**	**0.9298**
gbm	1.2711	0.3731

Table 3: Evaluation metrics of training set: RMSE and R^2 of different models predicting log(ON mode) using all data for training (removed correlated features and linear combinations, centered and scaled)

The following plots (Figure 18) were made to visualize and evaluate the predictions on the training set: A plot showing the actual and predicted log(ON mode) fluorescence sorted by index, a plot showing the log(observed) and log(predicted) values and a plot of the residuals for the log(predicted) values.

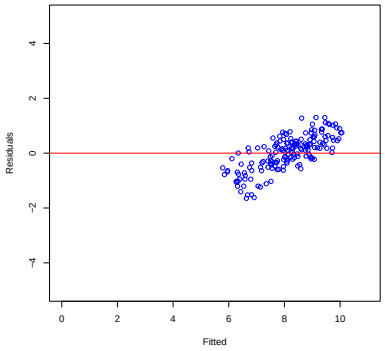
RANDOM FOREST The plots are shown for the random forest model, which had the lowest RMSE (0.63) when predictions are made on the training set. Compared to the RMSE from the k-fold cross-validation (RMSE 1.4521) it is obviously low. On the basis of predictions using

resampling, the train error for random forest is optimistic. The predicted versus actual values sorted by the index (Figure 18a), and the actual versus predicted values (Figure 18b) show that the predicted values and the actual values are not uniformly accurate. The model tends to overpredict low values, while it underpredicts high values. The residuals for the predicted values (see Figure 18c) do not show random scatter around 0, but this can be due to discretization and random noise.



(a) actual vs. predicted log(ON) sorted by index

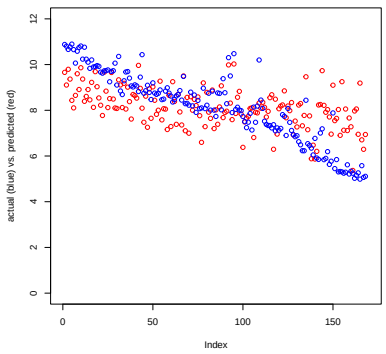
(b) actual vs. predicted log(ON)



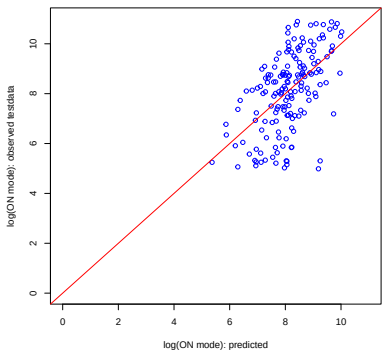
(c) residuals vs. fitted log(ON)

Figure 18: Diagnostic plots of training set error for random forest to evaluate and visualize the performance of the model: (a) actual vs. predicted log(ON) sorted by index (b) actual vs. predicted log(ON) (c) residuals vs. fitted log(ON)

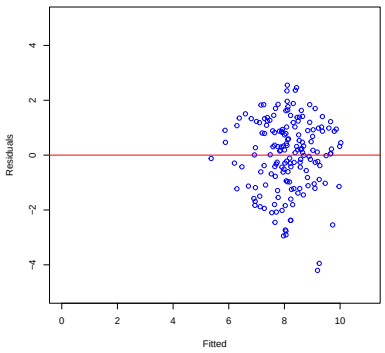
LINEAR MODEL The plots are also shown for the linear model which had a good performance, according to the RMSE, with all features, best subset and single feature, and the RMSE in-sample is also lower as for the cross-validation. Comparing the actual and predicted log(ON) sorted by index from random forest (Figure 18a) and linear model (Figure 19a), it is noticeable that the predictions from random forest do not scatter that far away from the actual values as those of the linear model. This is also apparent in the actual versus predicted plots. As well, the linear model has a tendency to overpredict low values and underpredict high values, but the predicted values spread in an almost horizontal band around the observed values. Figure 18b shows the plot for random forest and Figure 19b for the linear model. The fitted versus residuals plots for the linear model (Figure 19c) appears more randomly scattered around 0 than the plot for random forest (Figure 18c). Random forest shows constant variance, but not randomly scattered around 0, it is shifted around 0. For the linear model, the variance increases with fitted values to a certain point and then decreases again. Again, this pattern can be an artefact from discretization and random noise of the experimental data.



(a) actual vs. predicted log(ON) sorted by index



(b) actual vs. predicted log(ON)



(c) residuals vs. fitted log(ON)

Figure 19: Diagnostic plots of training set error for linear model to evaluate and visualize the performance of the model: (a) actual vs. predicted log(ON) sorted by index (b) actual vs. predicted log(ON) (c) residuals vs. fitted log(ON)

3.5.2 *All Features - k-fold cross-validation*

In order to access the performance of the different machine learning algorithms on new data repeated cross-validation was performed. Repeated cross-validation was used as resampling with 5 repeats and 10 folds. Using a numeric output variable in caret [85], data is split based on percentiles into groups sections. Using k-fold cross-validation it ensures that the test sets are independent since each sample is included only once in a fold. To compare the performance of the different algorithms RMSE was computed for all algorithms. Due to the use of identically resampled data sets, a paired t-test was used to evaluate if the differences between the tested models are statistically significant [92]. The RMSE for predicting the log(ON mode) fluorescence of the toehold switch designs, when correlated features are removed (there were no linear combinations), are shown in Table 4. The features P_sw_acc_RBS_linker_bs_b, P_tr_sw_acc_RBS_linker_bs_b, defect_sw and P_tr_acc_toe were removed to avoid highly correlated features (cutoff 0.75).

	RMSE	R^2
lm	**1.3802**	**0.2653**
pls	1.4012	0.2445
enet	**1.3772**	**0.2635**
svmRadial	1.4211	0.2334
treebag	1.4666	0.1704
rf	1.4474	0.1883
gbm	1.4683	0.1807

Table 4: k-fold cross-validation: RMSE and R^2 of different models predicting log(ON mode) using basic feature set (removed correlated and linear combinations, centered and scaled)

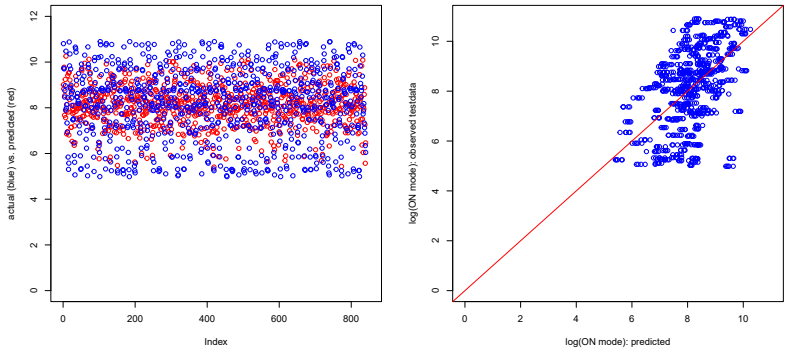
The baseline for log(ON mode) was 1.5579 (as reported in Table 5). The mean RMSE is smaller than the baseline for all tested algorithms. An asterisk marks that the RMSEs of an algorithm are significantly better than the baseline RMSEs. Using the mean RMSE for evaluation, all algorithms perform better than the baseline algorithm. Elasticnet has the lowest RMSE with 1.3772 and R^2 is 0.2635 the second best algorithm is a linear model with 1.3802 and R^2 is 0.2653. All other allgorithms have RMSE higher than 1.4. Elasticnet and the linear model explain 26-27% of the variation in the predicted value, and their RMSEs are at similar level. Further, the diagnostic plots resemble each other for all algorithms. Thus, for further analysis the linear model is shown, since linear models are good to understand and interprete. One drawback of the interpretability of linear regression models is, that the interpretation/increase of one feature forces all

other features to stay the same. It may be unrealistic that the increase of one feature is isolated from the increase of another feature.

	baseline RMSE	RMSE	t-value	p-value	significance
lm	**1.5579**	**1.3802**	**4.8678**	**0.0000**	*
pls	1.5579	1.4012	4.0827	0.0001	*
enet	1.5579	1.3772	4.8229	0.0000	*
svmRadial	1.5579	1.4211	3.7746	0.0002	*
treebag	1.5579	1.4666	2.7634	0.0040	*
rf	1.5579	1.4474	3.2735	0.0010	*
gbm	1.5579	1.4683	2.4504	0.0089	*

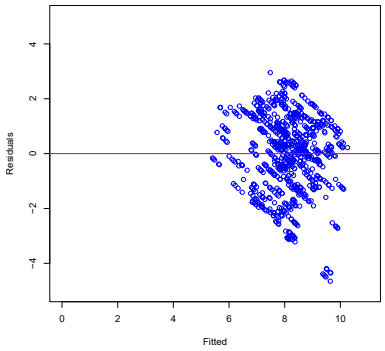
Table 5: Statistical significance test of predictions with k-fold cross-validation with basic feature set: Paired t-test. An asterisk marks that a model is statistically significant better (significance level $\alpha = 0.05$) than the baseline algorithm

Evaluating the performance of the models based on the RMSE, as mentioned before, all models outperform the baseline. To further evaluate and visualize diagnostic plots are made for the linear model. Figure 20a shows the actual and predicted log(ON) sorted by index. Most of the predicted values concentrate between 7 and 9, there are values predicted between 6 and 10, though the actual values are between 5 and 11. The model does underpredict high values and overpredict low values. This is also apparent from Figure 20b. The residuals versus fit plot (Figure 20c) does not show random scatter around 0, the variance increases with fitted values to a certain point and then decreases again.



(a) actual vs. predicted log(ON) sorted by index

(b) actual vs. predicted log(ON)



(c) residuals vs. fitted log(ON)

Figure 20: Diagnostic plots of linear model k-fold cross-validation with basic feature set to evaluate and visualize the performance of the model: (a) actual vs. predicted log(ON) sorted by index (b) actual vs. predicted log(ON) (c) residuals vs. fitted log(ON)

The linear regression model predicts the output variable log(ON mode) as a weighted sum of the input features. Figure 21 shows the variable importance for the linear model and Listing 3 the coefficients of the linear model (this are the same results as for the model of the

training set prediction). The variable importance for the linear models is the absoulte value of the t-statistics for each feature. The features are ordered by decreasing importance: the most significant features are e_net_complex (***) and e_RBS_linker (***), P_sw_acc_toe (*) and e_tr_sw_bs (*) and P_sw_hairpin (*). The other features have no significance code, because $p \geqslant 0.05$ for the other features. The high feature importance of e_RBS_linker is in accordance with the former results of the correlation analysis. e_net_complex has the highest feature importance, since this feature is the net free interaction energy of the activated complex it is reasonable that this feature is important to predict log(ON mode). Since the toehold region should facilitate the binding of the trigger, the probability that the toehold is accessible (P_sw_acc_toe) it is plausible that this is an important feature. As for the correlation analysis the underlying mechanismus why e_tr_sw_bs is a relevant feature is unclear. P_sw_hairpin has a high importance for the mode, according to the purpose of the hairpin, it is an translation repressor in absence of the trigger, which is an essential part in the regulation of the toehold switches. The interpretation of a feature is, e. g., if the e_RBS_linker increases by 1 kcal/mol the predicted fluorescence increases by 0.5883, while all other features stay constant. Whereas, when the feature e_net_complex increases, the fluorescence decreases by 0.9350, while all other features stay the same.

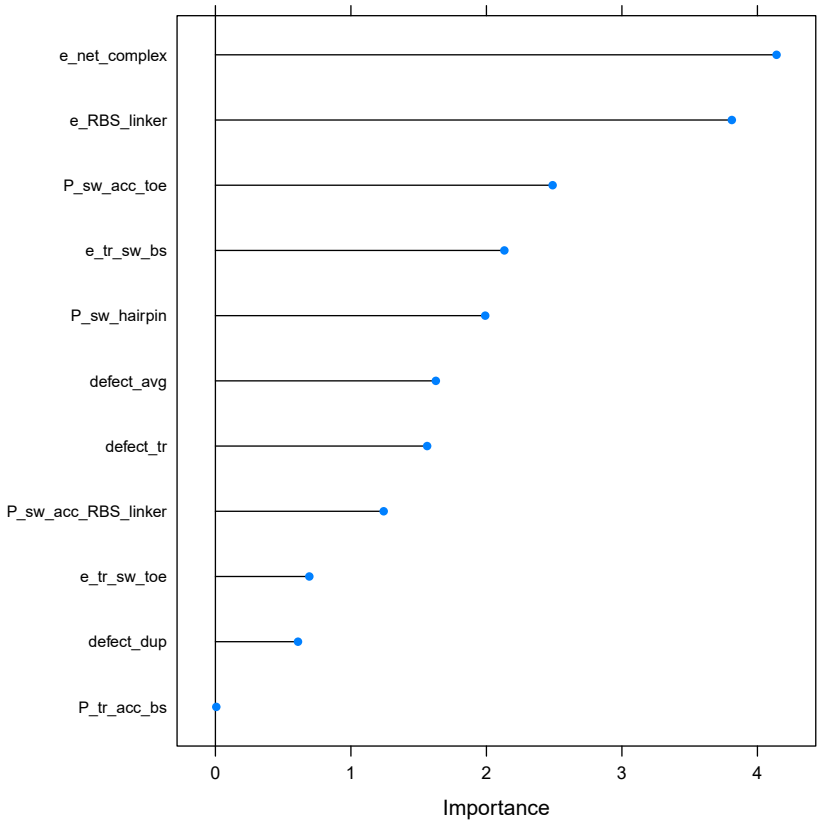


Figure 21: Linear model: Variable importance (k-fold cross-validation with basic feature set)

Listing 3: Coefficients of linear model using basic feature set

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2025 -0.9007  0.1757  0.9568  2.5504

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.138154   0.104195  78.105  < 2e-16 ***
e_RBS_linker         0.588304   0.154366   3.811 0.000199 ***
e_tr_sw_bs           0.504093   0.236394   2.132 0.034537 *
e_tr_sw_toe          0.122770   0.177134   0.693 0.489283
e_net_complex       -0.935092   0.225804  -4.141 5.64e-05 ***
P_sw_acc_RBS_linker  0.202215   0.162842   1.242 0.216179
P_sw_acc_toe        -0.556977   0.223818  -2.489 0.013878 *
P_sw_hairpin        -0.225574   0.113289  -1.991 0.048211 *
P_tr_acc_bs         -0.001109   0.170012  -0.007 0.994803
defect_dup          -0.141894   0.232827  -0.609 0.543120
defect_tr            0.417355   0.267097   1.563 0.120183
defect_avg          -0.480098   0.295127  -1.627 0.105809
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.351 on 156 degrees of freedom
Multiple R-squared:  0.3144,    Adjusted R-squared:  0.266
F-statistic: 6.503 on 11 and 156 DF,  p-value: 7.45e-09
```

3.5.3 *Best subset - k-fold cross-validation*

Using `regsubsets` from the `leaps` package [88] the best subset by sequential replacement for a regression model was computed (when correlated and linear combinations are removed). Using the CP-criterion a model with 8 features is the best subset: e_RBS_linker, e_tr_sw_bs, e_net_complex, P_sw_acc_RBS_linker, P_sw_acc_toe, P_sw_hairpin, defect_tr, defect_avg. Compared to the single feature correlation with the log(ON mode) (see Table 11) 5 of the features with the highest correlation to the log(ON mode) are within the set of 8 features. e_RBS_linker is the first feature included in the best subset, e_RBS_linker also has the highest correlation to the output variable.

Table 6 shows the RMSE and R^2 of the different trained algorithms with the best subset features. Elasticnet and the linear model perform best according to the RMSE (1.3543, R^2 is 0.2890), the second best is a partial least squares model with RMSE 1.3817 and R^2 is 0.2625. This three algorithms improved their performance according to their RMSEs and R^2. Furthermore, support vector machine with radial basis function kernel improved, all other algorithm got worse.

	RMSE	R^2
lm	**1.3543**	**0.2890**
pls	1.3817	0.2625
enet	**1.3543**	**0.2890**
svmRadial	1.4412	0.2043
treebag	1.4796	0.1663
rf	1.4621	0.1752
gbm	1.4768	0.1769

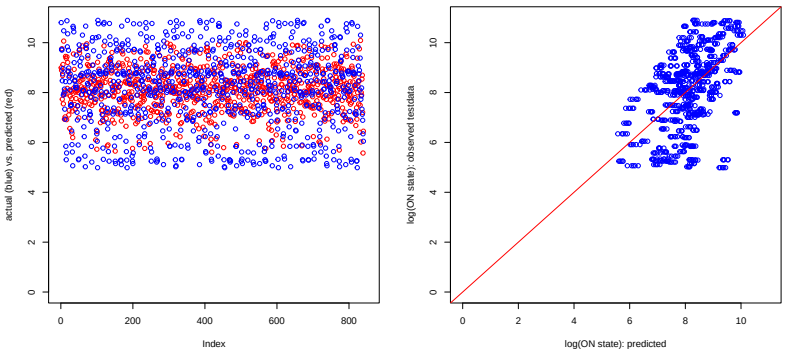
Table 6: k-fold cross-validation: RMSE and R^2 of different models predicting log(ON mode) using best subset feature set (removed correlated and linear combinations, centered and scaled)

Table 7 shows the results of the t-test for the algorithms trained with the 8 features from the best subset. Again, all algorithms, though some got worse, outperform the baseline.

	baseline RMSE	RMSE	t-value	p-value	significance
lm	**1.5579**	**1.3543**	**5.5408**	**0.0000**	*
pls	1.5579	1.3817	4.5927	0.0000	*
enet	1.5579	1.3543	5.5413	0.0000	*
svmRadial	1.5579	1.4412	3.3219	0.0008	*
treebag	1.5579	1.4796	2.2404	0.0148	*
rf	1.5579	1.4621	2.7662	0.0040	*
gbm	1.5579	1.4768	2.2673	0.0139	*

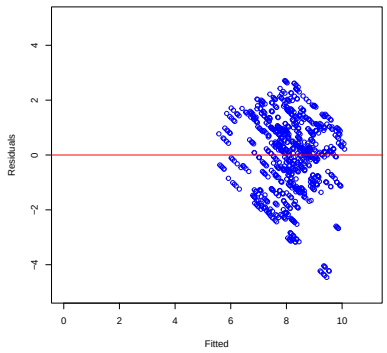
Table 7: Statistical significance test of predictions with k-fold cross-validation with best subset feature set: Paired t-test, an asterisk marks that a model is statistically significant better (significance level $\alpha = 0.05$) than the baseline algorihm

The diagnostic plots for the linear model are evaluated. Although the RMSE showed an (minor) improvement from 1.3802 to 1.3543 and R^2 0.2927 to 0.2890 the plots from the linear model with the best subset features resemble those of the full feature set. Figure 22a shows the actual and predicted log(ON mode) sorted by index. As for the full feature set (Figure 20a), there are predicted values between 6 and 10, but most of the predicted values concentrate between 7 and 9. The actual values are between 5 and 11. The model does underpredict high values and overpredict low values, which can also be seen in Figure 22b. Figure 22c (the residual versus fit plot) does not show random scatter around 0.



(a) actual vs. predicted log(ON) sorted by index

(b) actual vs. predicted log(ON)



(c) residuals vs. fitted the log(ON)

Figure 22: Diagnostic plots of linear model k-fold cross-validation with best subset feature set to evaluate and visualize the performance of the model: (a) actual vs. predicted log(ON) sorted by index (b) actual vs. predicted log(ON) (c) residuals vs. fitted log(ON)

The variable importance for the best subset features using a linear model are shown in Figure 23. Listing 4 shows the summary of the linear model. The features ordered by decreasing importance: e_RBS_linker (***) and e_net_complex (***) are the most significant

features again, but in this setting e_RBS_linker has a minimal higher significance. All features except P_sw_acc_RBS_linker are statistically significant. Former in this thesis, this feature was analysed in more detail, due to its unexpected correlation high correlation to the log(ON mode) and the log(OFF mode), respectively.

Multiple R^2 for the whole feature set ($R^2 = 0.3144$) and the best subset feature set ($R^2 = 0.31$) are almost the same. But the adjusted R^2 differ more, adjusted R-squared for best subset is 0.2753 and 0.266 for the full feature set. Therefore, the bestsubset linear model is minimal better. The adjusted R^2 takes into account the number of added variables, while R^2 increases as more variables are added. In contrast to the feature importance with all features, here also defect_tr and defect_avg are statistically significant for the model.

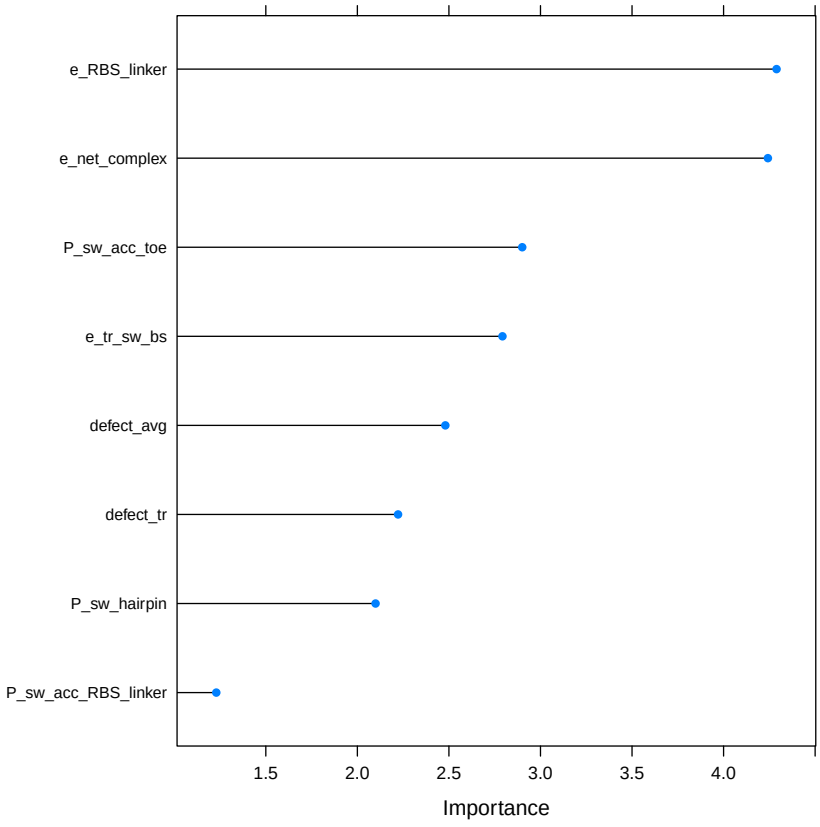


Figure 23: Linear model: Variable importance (k-fold cross-validation with best subset feature set)

Listing 4: Coefficients of linear model using best subset feature set

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1394 -0.9545  0.1541  0.9457  2.6368

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.1382     0.1035  78.602  < 2e-16 ***
e_RBS_linker          0.6138     0.1431   4.289 3.10e-05 ***
e_tr_sw_bs            0.5380     0.1927   2.792  0.00587 **
e_net_complex        -0.8380     0.1976  -4.242 3.75e-05 ***
P_sw_acc_RBS_linker   0.1845     0.1500   1.230  0.22070
P_sw_acc_toe         -0.5469     0.1886  -2.900  0.00426 **
P_sw_hairpin         -0.2271     0.1082  -2.099  0.03736 *
defect_tr             0.4764     0.2144   2.222  0.02768 *
defect_avg           -0.5613     0.2263  -2.480  0.01416 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.342 on 159 degrees of freedom
Multiple R-squared:   0.31,     Adjusted R-squared:  0.2753
F-statistic: 8.929 on 8 and 159 DF,  p-value: 4.516e-10
```

3.5.4 *Single feature: e_RBS_linker*

Since the feature e_RBS_linker seems to have a high impact on the functionality of the toehold switch designs (as already Green et al. stated and also this work confirms), and it is one of the features with the highest correlation to the log(ON mode), models were built with only this single feature. The metrics for the build models are shown in Table 8. The linear model has the lowest RMSE with 1.4454 and R^2 with 0.1921. These values are the worst results for any tried linear model. Using the full feature set (after pre-processing) results in a RMSE of 1.3802 and R^2 of 0.2653, with the best subset feature set the RMSE is 1.3543 and R^2 0.2890 and in-sample the RMSE is 1.30 and R^2 0.31.

But as the results of the paired t-test in Table 9 show the linear model and boosted tree still outperform the baseline, but support vector machine with radial basis function kernel and bagged tree do not anymore.

Figure 24a shows the actual and predicted log(ON mode) of the linear model sorted by index of the models build with the single feature e_RBS_linker. The linear model only predicts values between 6 and 9, whereas most of the values concentrate around 7 to 9. The actual values vary between 5 and 11. In contrast to the former models

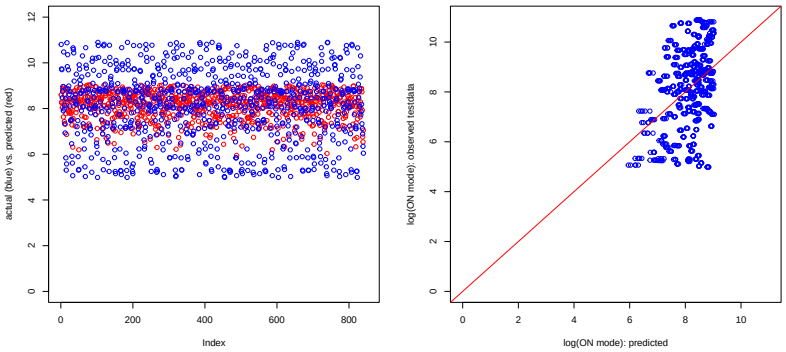
	RMSE	R^2
lm	**1.4454**	**0.1921**
svmRadial	1.5349	0.1062
treebag	1.5573	0.1032
gbm	1.4877	0.1536

Table 8: k-fold cross-validation: RMSE and R^2 of different models predicting log(ON mode) using single feature e_RBS_linker (centered and scaled)

	baseline RMSE	RMSE	t-value	p-value	significance
lm	**1.5579**	**1.4454**	**3.3341**	**0.0008**	*
svmRadial	1.5579	1.5349	0.7471	0.2293	-
treebag	1.5579	1.5573	0.0155	0.4938	-
gbm	1.5579	1.4877	2.1726	0.0173	*

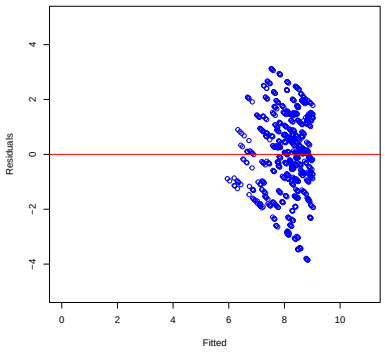
Table 9: Statistical significance test of predictions with k-fold cross-validation with single feature e_RBS_linker: Paired t-test, an asterisk marks that a model is statistically significant better (significance level $\alpha = 0.05$) than the baseline algorihm

with more features, this model predicts values in a narrower band. The other linear models with all features or best subset predict values between 6 and 10. From Figure 24b, actual versus predicted log(ON mode) plot, it can be seen that the model underpredicts high values. The residuals versus fit plot (Figure 24c) shows non-constant scatter around 0. The variance increases with fitted values to a certain point and then decreases.



(a) actual vs. predicted log(ON) sorted by index

(b) actual vs. predicted log(ON)



(c) residuals vs. fitted log(ON)

Figure 24: Diagnostic plots of linear model k-fold cross-validation with single feature e_RBS_linker set to evaluate and visualize the performance of the model: (a) actual vs. predicted log(ON) sorted by index (b) actual vs. predicted log(ON) (c) residuals vs. fitted log(ON)

3.5.5 *Conclusion*

Just evaluating the performance of the models based on the RMSE all models, except support vector machine with radial basis function kernel and bagged tree with the single feature e_RBS_linker (Table 9), perform significantly better than the baseline algorithm. But these improvements in the prediction evaluated on the RMSE are minimal. In general, experimental raw data can have random noise, which masks patterns in the data. Further, the discretization of raw data can hide important information. Furthermore, it is possible that the used features do not capture enough important information to make good predictions. As the predictions with the single feature e_RBS_linker only predict values in a smaller range than the other feature sets, it is obvious that there should be a focus on good feature selection. It is possible that another feature selection already would improve the predictions. Above all, since all toehold switch designs base on one optimized design and have certain fixed sequences, all designs can be actually good, but the features can not distinguish between these, resulting in minimal predictive power of the models.

APPENDIX

A.1 FEATURE SELECTION AND COMPUTATION

A.1.1 *Prior features by Green et al.*

C. Definitions of thermodynamic parameters

Class	Name of thermodynamic term	Description
Free energy of individual RNAs	ΔG switch	Free energy of the longer switch RNA including the GGG leader sequence and the first 29-nts of GFPmut3b following the linker
	ΔG trigger	Free energy of the entire trigger RNA including the GGG leader, 5′ hairpin, and transcriptional terminator
	ΔG complex	Free energy of the bimolecular complex formed by the switch and trigger RNAs used for ΔG switch and ΔG trigger
	ΔG min. switch	Free energy of the minimal switch RNA element from the GGG leader to the last base of the stem (immediately before linker sequence)
	ΔG min. trigger	Free energy of the minimal 30-nt trigger sequence
	ΔG min. complex	Free energy of the bimolecular complex formed by the switch and trigger RNAs used for ΔG min. switch and ΔG min. trigger
Deviation from ideal secondary structures	Dev. ΔG switch	The difference in free energies between the ideal secondary structure and the actual secondary structure for the *longer* switch RNA sequence
	Dev. ΔG trigger	The difference in free energies between the ideal secondary structure and the actual secondary structure for the *complete* trigger RNA sequence
	Dev. ΔG complex	The difference in free energies between the ideal secondary structure and the actual secondary structure for the bimolecular complex formed between the longer switch RNA and the complete trigger RNA sequence
	Dev. ΔG min. switch	The difference in free energies between the ideal secondary structure and the actual secondary structure for the *minimal* switch RNA sequence
	Dev. ΔG min. trigger	The difference in free energies between the ideal secondary structure and the actual secondary structure for the *minimal* trigger RNA sequence
	Dev. ΔG min. complex	The difference in free energies between the ideal secondary structure and the actual secondary structure for the bimolecular complex formed between the minimal switch RNA and trigger RNA sequence
Net reaction free energies	ΔG toehold binding	The free energy of the *n*-base pair duplex formed by the binding of the *n*-nt toehold to its reverse complement
	ΔG trigger binding	The free energy of the 30-base pair duplex formed by the binding of the 30-nt minimal trigger RNA to its reverse complement
	Net ΔG complex	The net free energy of the interaction between the longer switch RNA and complete trigger RNA to form the activated switch complex (i.e. Net ΔG complex = ΔG complex - ΔG switch - ΔG trigger)
	Net ΔG min. complex	The net free energy of the interaction between the minimal switch RNA and trigger RNA to form the minimal activated switch complex
RBS/mRNA secondary structure	ΔG(RBS to GFP) or $\Delta G_{RBS\text{-}GFP}$	The free energy of the RNA sequence starting from the first base of the loop of the switch RNA to the 29th base of the GFPmut3b sequence
	ΔG(RBS to linker) or $\Delta G_{RBS\text{-}linker}$	The free energy of the RNA sequence starting from the first base of the loop of the switch RNA to the last base of the 21-nt linker sequence
	ΔG(AUG to GFP) or $\Delta G_{AUG\text{-}GFP}$	The free energy of the RNA sequence starting from the first base of the AUG start codon in the switch stem to the 29th base of the GFPmut3b sequence
	ΔG(RBS to AUG) or $\Delta G_{RBS\text{-}AUG}$	The free energy of the RNA sequence starting from the first base of the loop of the switch RNA to the last base of the AUG start codon
Stability of top of toehold switch stem region	ΔG stem top ***n*** *(with $1 \leq n \leq 16$)*	The free energy calculated for the sequence running from the ***n***-th base from the bottom of the switch RNA stem, through the loop, to its counterpart (also ***n***-bases from the stem bottom) on the other side of the stem
Stability of bottom of toehold switch stem region	ΔG stem bot. ***n*** (with $1 \leq \boldsymbol{n} \leq 16$)	The free energy of the sequence consisting of the bottom ***n***-bases on each side of the switch stem, joined using a loop with the sequence $(A)_{10}$

Figure 25: Thermodynamic features from Green et al. [54]

A.1.2 *Structure constraints used for secondary structure probability features*

$$P_{specific\_structure} = \frac{Z_{X_{specific\_structure}}}{Z_x} = e^{\frac{G_{X_{specific\_structure}} - G_X}{RT}}$$

$X_{specific\_structure}$ refers to the 1. constraint
x refers to 2. constraint here.

A.1.2.1 *P_sw_acc_RBS_linker*

```
1. constraint:
.....................................xxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxx.............................
0                  20                  40                  60
    x: RBS-linker region constrained to stay unpaired

2.constraint: unconstrained
```

A.1.2.2 *P_sw_acc_toe*

```
1. constraint:
...xxxxxxxxxxxx.....................................
................................................
0                  20                  40                  60
    x: region a (toehold) constrained

2.constraint: unconstrained
```

A.1.2.3 *P_sw_acc_RBS_linker_bs_b*

```
1. constraint:
...xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx...xxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxx.............................
0                  20                  40                  60
    ...xxx...xxxxxxx...: binding site and RBS-linker region
    constrained to stay unpaired

2. constraint
...xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.........................
................................................
0                  20                  40                  60
    x: binding site constrained to stay unpaired
```

A.1.2.4 *P_sw_hairpin*

```
1. constraint:
..............(((((((((...((((((...........))))))...))))))
)))........................................
```

```
0                 20                  40                  60
    (: hairpin constrained to basepair

2. constraint: unconstrained
```

A.1.2.5 *P_tr_acc_bs*

```
1. constraint:
.............................xxxxxxxxxxxxxxxxxxxxxxxxxxx
xxx.............................................
0                 20                  40                  60
    x: binding site constrained to stay unpaired

2. constraint: unconstrained
```

A.1.2.6 *P_tr_acc_toe*

```
1. constraint:
.............................................xxxxxxxxx
xxx.............................................
0                 20                  40                  60
    x: region a (toehold) constrained to stay unpaired

2. constraint: unconstrained
```

A.1.2.7 *P_tr_sw_acc_RBS_linker_bs_b*

```
1. constraint:
.............................(((((((((((((((((((((((((
(((.................................................&...))
)))))))))))))))))))))))))xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx..........................
0                 20                  40                  60
    &:  removed for computation but for clarity to distinguish
        trigger (1. sequence in constraint) and switch
        (2. sequence in constraint)
    (): constraint for basepairing between switch and trigger
        binding site
    x:  RBS-linker region to stay unpaired

2. constraint: unconstrained
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
e_RBS_linker	-1.18E+01	-7.57E+00	-6.28E+00	-6.64E+00	-5.48E+00	-4.45E+00
e_tr_sw_bs	-6.10E+01	-5.12E+01	-4.82E+01	-4.85E+01	-4.59E+01	-4.13E+01
e_tr_sw_toe	-2.18E+01	-1.39E+01	-1.31E+01	-1.31E+01	-1.14E+01	-9.60E+00
e_net_complex	-4.43E+01	-3.81E+01	-3.51E+01	-3.54E+01	-3.26E+01	-2.75E+01
P_sw_acc_RBS_linker	1.16E-21	5.35E-19	4.73E-18	5.55E-16	7.56E-17	1.62E-14
P_sw_acc_toe	5.88E-05	1.93E-02	8.85E-02	1.81E-01	3.50E-01	7.00E-01
P_sw_acc_RBS_linker_bs_b	7.60E-13	1.05E-09	1.78E-08	1.19E-07	8.21E-08	1.62E-06
P_sw_hairpin	6.37E-01	8.07E-01	8.48E-01	8.47E-01	8.96E-01	9.74E-01
P_tr_acc_bs	8.32E-04	2.24E-01	3.95E-01	3.83E-01	5.21E-01	8.87E-01
P_tr_acc_toe	2.22E-03	6.10E-01	7.82E-01	6.99E-01	8.66E-01	9.71E-01
P_tr_sw_acc_RBS_linker_bs_b	5.42E-14	6.11E-10	1.03E-08	7.78E-08	4.91E-08	1.20E-06
defect_dup	1.20E-01	1.40E-01	1.54E-01	1.53E-01	1.65E-01	1.94E-01
defect_sw	1.88E-01	1.97E-01	2.06E-01	2.11E-01	2.22E-01	2.68E-01
defect_tr	4.54E-02	5.75E-02	6.69E-02	6.93E-02	7.72E-02	1.18E-01
defect_avg	1.29E-01	1.39E-01	1.43E-01	1.45E-01	1.49E-01	1.77E-01
on_mode	1.46E+02	1.25E+03	4.01E+03	9.01E+03	1.00E+04	5.39E+04
off_mode	1.01E+02	1.62E+02	1.85E+02	2.74E+02	2.45E+02	2.74E+03
fold_change_mode	5.98E-01	5.88E+00	1.84E+01	4.11E+01	4.30E+01	2.92E+02

Table 10: Summary statistics for each feature: minimum value, value of 1st quartile, median value, value of 3rd quartile and maximum value

A.2 DATA ANALYSIS - DESCRIPTIVE STATISTICS AND VISUALIZATION

A.2.1 *Independent Variables (Features)*

Table 10 shows the numeric values of the summary statistics of each feature, as calculated in Section 3.2.

	ON mode		OFF mode		ON/OFF	
	R	R^2	R	R^2	R	R^2
e_RBS_linker	0.39	0.15	0.05	0.00	0.38	0.15
P_sw_acc_RBS_linker	0.39	0.15	0.23	0.05	0.33	0.11
P_tr_sw_acc_RBS_linker_bs_b	0.30	0.09	0.04	0.00	0.29	0.09
P_sw_acc_RBS_linker_bs_b	0.29	0.09	0.04	0.00	0.29	0.08
defect_dup	-0.28	0.08	-0.04	0.00	-0.28	0.08
e_tr_sw_bs	0.24	0.06	0.20	0.04	0.18	0.03
P_sw_hairpin	-0.23	0.05	-0.07	0.00	-0.21	0.04
defect_avg	-0.18	0.03	-0.17	0.03	-0.13	0.02
e_net_complex	0.12	0.02	-0.03	0.00	0.14	0.02
e_tr_sw_toe	0.03	0.00	0.06	0.00	0.01	0.00
defect_tr	-0.02	0.00	-0.14	0.02	0.03	0.00
P_sw_acc_toe	-0.02	0.00	0.07	0.00	-0.04	0.00
P_tr_acc_bs	0.02	0.00	0.15	0.02	-0.03	0.00
defect_sw	-0.01	0.00	-0.10	0.01	0.02	0.00
P_tr_acc_toe	0.00	0.00	0.09	0.01	-0.04	0.00

Table 11: R and R^2 of each feature with log(ON mode), ordered by decreasing absolute value of R of log(ON mode)

A.3 CORRELATION

Table 11 is a compact summary of R and R^2 numeric values of the features, as calculated in Section 3.3. Figure 8 is a graphical representation of this data.

A.4 SINGLE STRAND ANALYSIS - RNAFOLD

A.4.1 *Switches with low accessibility of RBS-linker region and low ON mode*

Both switch 161 and switch 141 have a low probability for the accessibility of the RBS-linker region. The structures of switch 140 (Figure 11c) and switch 141 (Figure 11a) form the target structure, whereas switches 161 (Figure 12c) and 164 (Figure 12a) have a shortened toehold region a (in the centroid structure). The region a is one nucleotide shortened, therefore the bottom of the stem of the hairpin is elongated by one nucleotide. This one nucleotide basepairs with the first nucleotide of the linker. There are minimal changes in the design, which affect the toehold and the stem, but the RBS and the start codon are not affeced thereby and are formed as designed. The MFEs for those four switches are also similar, for switch 140 -28.60 kcal/mol, switch 141 29.40 kcal/mol, switch 161 -29.40 kcal/mol and switch 164 -28.40 kcal/mol.

A.4.2 *Minimal accessibility of RBS-linker region and high ON mode*

Examining the structure of switch 17 (see Figure 13a), it stands out that there are single basepairs upstream to the hairpin. The 3rd nucleotide of the leader sequence and toehold region a basepair with the linker region at 3 positions, but the hairpin forms as designed. Therefore the RBS and the startcodon also form the target structure. The MFE of switch 17 is -31.40 kcal/mol.

A.4.3 *Lowest and highest ON/OFF mode*

Switch 1 does not form the target structure (Figure 14a), while the second best switch 2 (Figure 14c) and the worst switch 168 (Figure 15a) do.

A.4.4 *Lowest and highest ON mode*

The secondary structure of the switches with the highest and the lowest on mode were analyzed. Switch 5 has the highest ON mode and a moderate accessibility. The structure is shown in Figure 16c. MFE is -28.90 kcal/mol. The GGG leader sequence plus the first nucleotide of the toehold fold back onto the linker region. This results in a truncated toehold, it is 10 instead of 12 nucleotides long. The stem of the hairpin starts one nucleotide more upstream. Also part b of the binding site is truncated, it is 9 nucleotides instead of 18 nucleotides long. The bulge, RBS and startcodon at the designed positions.

Switch 165 has the lowest on mode and a high accessibility. The structure is shown in Figure 16a. MFE is -26.20 kcal/mol. Also the toehold region of switch 165 folds back onto the linker region, nucleotides 5 to 7 basepair with the linker region. Therefore the toehold is truncated, it is 7 nucelotides long not 12. The stem starts one nucleotide earlier. The bulge, RBS and the startcodon are again at the designed positions.

A.4.5 *Low accessibility and mediocre ON mode*

The secondary structures of the switches 84 and 66 were also analyzed, since these switches also have a low accessibility. But their on mode is moderate. Switch 84 forms the designed secondary structure (Figure 17a), but switch 66 has some alterations (Figure 17b). In switch 66 the toehold is only 9 nucleotides long, not 12. Therefore the stem starts at nucleotide 13 not 16, and there is an unpaired nucleotide in the stem. Therefore region b of the binding site is 20 nucleotides long, not 18. The RBS and the startcodon are at the target positions, but the unpaired linker region starts 4 nucleotides more downstream. The MFE of switch 66 is -32.20 kcal/mol and of switch 84 -32.00 kcal/mol.

A.5 RNACOFOLD - DIMER STRUCTURE PREDICTION, RNA-RNA INTERACTION

The following dimer interactions were computed: Switch 1 and trigger 1 which have the highest ON/OFF fluorescence are shown in Figure 14b. Switch 168 and trigger 168 with the lowest on/off fluorescence are shown in Figure 15b.

Switch 1 and trigger 1 show the highest on/off fluorescence. Switch and trigger form the designed 30 nucleotide long binding site, but the leader sequence basepairs with the terminator of the switch. The RBS and the startcodon stay unpaired.

Switch 168 and trigger 168 show the lowest on/off fluorescence. In contrast to switch-trigger1 the leader sequence does not fold back onto the terminator of the trigger. The 30 nucleotide long binding site forms as designed. The RBS is not unpaired it basepairs with the linker region. Additionally, the UG of the startcodon also form basepairs with nucleotides of the linker region.

Switch 17 which has the overall lowest accessibility, but a good performance regarding the ON mode and the ON/OFF. The last G of the leader sequence basepairs with the trigger (but this time not with the terminator of the trigger). The binding site is one nucleotide elongated in the switch-trigger duplex. 7 of the 8 nucleotides from the RBS form basepairs with parts of the linker region, but the startcodon is unpaired.

The dimer structure of switch 164 and trigger 164 shows that only the first G of the leader sequence stays unpaired, the other to basepair with the trigger. The binding site with the trigger is longer. The RBS is unpaired, but the G of the startcodon forms a basepair.

The dimer of switch 141 and trigger 141 forms basepairs already at the 2 first Gs of the leader sequence. But the binding site starts as designed at nucleotide 4 of the switch and is 30 nucleotides long. The RBS is not free, it forms baepiar with the linker region. Only the A of the startcodon is free UG form basepairs with other nucleotides.

Switch 165, which has the lowest on mode, but a high accessibility. The dimer structure of switch 165 and trigger 165 shows that the third G of the leader basepairs with the trigger, therefore the binding site is 31 nucleotides long. The RBS is unpaired and the starcodon also.

To sum up, there are again duplexes of switch and trigger as switch-trigger 165 which have an accessible RBS and starcodon, but show a low fluorescence, while switch and trigger 1 also have an accessible RBS and starcodon and show a good fluorescence. Switch 17 which shows a good fluorescence has an almost complety paired RBS in the duplex but has a high fluorescence level. According to the secondary structures it is not obvious why some of the duplexes show a high fluorescence level while the others do not.

ABSTRACT

The aim of this thesis is to define numerical features that describe individual RNA sequence designs in numerical manner. This should enable to predict the functionality of such sequence designs computationally in advance, before analysis in the laboratory. The features were determined by a knowledge-based approach and calculated using several RNA related tools. To evaluate the predictive performance of the defined features machine learning algorithms were applied. The underlying data are de novo designed *toehold switch* RNA sequences with corresponding fluorescence measurements from verification experiments from Green et al. [54]. The main part of this thesis covers the computation and description of the features, and a correlation analysis of the defined features with the ON state fluorescence of the RNA sequence designs. Due to some unexpected correlations one feature was chosen for further analysis. The feature P_sw_acc_RBS_linker was chosen, in order to examine if there is a structure-dependent explanation for the correlation of this feature with the output variable. Based on some specified characteristics some RNA switches were chosen and further analysed. Therefore, secondary structures of individual single stranded switches were computed, and secondary structures were predicted with their cognate trigger RNA. Finally, the defined features were used to build predictive models that enable to compute the functionality of the RNA switches. Most of the models outperformed the baseline model (significance level $\alpha = 0.05$).

ZUSAMMENFASSUNG

Das Ziel dieser Arbeit ist es, numerische Features zu definieren, die individuelle RNA-Sequenzdesigns auf numerische Art beschreiben. Dadurch soll es ermöglicht werden, die Funktionalität solcher Sequenzdesigns im Vorhinein zu berechnen, bevor sie im Labor analysiert werden. Die Features wurden anhand eines wissensbasierten Ansatzes bestimmt und mit verschiedenen RNA-bezogenen Tools berechnet. Um die Vorhersagekraft der definierten Features zu beurteilen, wurden verschiedene Algorithmen des maschinellen Lernens eingesetzt. Die zugrundeliegenden Daten sind de novo designte *toehold switch* RNA-Sequenzen mit den zugehörigen Fluoreszenzmessungen aus Verifikationsexperimenten von Green et al. [54]. Der Hauptteil dieser Arbeit umfasst die Berechnung und Beschreibung der Features, sowie eine Korrelationsanalyse der definierten Features mit der ON-State-Fluoreszenz der RNA-Sequenzdesigns. Aufgrund mehrerer unerwarteter Korrelationen wurde ein Feature ausgewählt, das näher untersucht wurde. Das Feature P_sw_acc_RBS_linker wurde ausgewählt, um eine mögliche strukturbedingte Erklärung für die Korrelation des Features mit der Outputvariable zu untersuchen. Einige RNA-Switch-Designs wurden aufgrund bestimmter Charakteristika ausgewählt und weiter analysiert. Dafür wurden die Sekundärstrukturen für individuelle einzelsträngige Switches berechnet und für einige wurden die Sekundärstrukturen mit der zugehörigen Trigger-RNA berechnet. Abschließend wurden die definierten Features verwendet, um prädikative Modelle zu bauen, die es ermöglichen die Funktionalität der RNA-Switches zu berechnen. Die meisten Modelle waren statistisch signifikant (Signifikanzniveau $\alpha = 0.05$) besser als das Baselinemodell.

BIBLIOGRAPHY

[1] F. H. Crick. "On protein synthesis." eng. In: *Symposia of the Society for Experimental Biology* 12 (1958), pp. 138–163. ISSN: 0081-1386.

[2] P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. "The Structural Basis of Ribosome Activity in Peptide Bond Synthesis." en. In: *Science* 289.5481 (Aug. 2000), pp. 920–930. ISSN: 0036-8075, 1095-9203.

[3] F. H. Crick. "The origin of the genetic code." eng. In: *Journal of Molecular Biology* 38.3 (Dec. 1968), pp. 367–379. ISSN: 0022-2836.

[4] L. E. Orgel. "Evolution of the genetic apparatus." eng. In: *Journal of Molecular Biology* 38.3 (Dec. 1968), pp. 381–393. ISSN: 0022-2836.

[5] C. R. Woese. *The genetic code: the molecular basis for genetic expression*. Harper & Row, 1967. 216 pp.

[6] W. Gilbert. "Origin of life: The RNA world." en. In: *Nature* 319.6055 (Feb. 1986), pp. 618–618. ISSN: 0028-0836.

[7] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. "Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena." English. In: *Cell* 31.1 (Nov. 1982), pp. 147–157. ISSN: 0092-8674, 1097-4172.

[8] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. "The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme." English. In: *Cell* 35.3 (Dec. 1983), pp. 849–857. ISSN: 0092-8674, 1097-4172.

[9] G. Storz. "An Expanding Universe of Noncoding RNAs." en. In: *Science* 296.5571 (May 2002), pp. 1260–1263. ISSN: 0036-8075, 1095-9203.

[10] K. Morris and J. Mattick. "The rise of regulatory RNA." In: *Nat Rev Genet* 15.6 (June 2014), pp. 423–437. ISSN: 1471-0056.

[11] S. M. Fica, N. Tuttle, T. Novak, N.-S. Li, J. Lu, P. Koodathingal, Q. Dai, J. P. Staley, and J. A. Piccirilli. "RNA catalyses nuclear pre-mRNA splicing." en. In: *Nature* 503.7475 (Nov. 2013), pp. 229–234. ISSN: 0028-0836.

[12] S. R. Eddy. "Non–coding RNA genes and the modern RNA world." en. In: *Nature Reviews Genetics* 2.12 (Dec. 2001), pp. 919–929. ISSN: 1471-0056, 1471-0064.

[13] L. He and G. J. Hannon. "MicroRNAs: small RNAs with a big role in gene regulation." In: *Nature Reviews Genetics* 5 (July 1, 2004), p. 522.

[14] L. S. Waters and G. Storz. "Regulatory RNAs in Bacteria." In: *Cell* 136.4 (Feb. 2009), pp. 615–628. ISSN: 0092-8674.

[15] E. G. H. Wagner, S. Altuvia, and P. Romby. "12 - Antisense RNAs in Bacteria and Their Genetic Elements." In: *Homology Effects*. Ed. by J. C. Dunlap and C. ting Wu. Vol. 46. Advances in Genetics. Academic Press, 2002, pp. 361–398.

[16] S. Brantl. "Regulatory mechanisms employed by cis-encoded antisense RNAs." eng. In: *Current Opinion in Microbiology* 10.2 (Apr. 2007), pp. 102–109. ISSN: 1369-5274.

[17] J. Chappell, K. E. Watters, M. K. Takahashi, and J. B. Lucks. "A renaissance in RNA synthetic biology: new mechanisms, applications and tools for the future." In: *Current Opinion in Chemical Biology*. Synthetic biology • Synthetic biomolecules 28 (2015), pp. 47–56. ISSN: 1367-5931.

[18] J. A. J. Arpino, E. J. Hancock, J. Anderson, M. Barahona, G.-B. V. Stan, A. Papachristodoulou, and K. Polizzi. "Tuning the dials of Synthetic Biology." In: *Microbiology* 159.Pt 7 (July 2013), pp. 1236–1253. ISSN: 1350-0872.

[19] P. E. M. Purnick and R. Weiss. "The second wave of synthetic biology: from modules to systems." eng. In: *Nature Reviews. Molecular Cell Biology* 10.6 (June 2009), pp. 410–422. ISSN: 1471-0080.

[20] S. Findeiß, M. Wachsmuth, M. Mörl, and P. F. Stadler. "Chapter One - Design of Transcription Regulating Riboswitches." In: *Methods in Enzymology*. Ed. by D. H. Burke-Aguero. Vol. 550. Riboswitches as Targets and Tools. Academic Press, 2015, pp. 1–22.

[21] E. Andrianantoandro, S. Basu, D. K. Karig, and R. Weiss. "Synthetic biology: new engineering rules for an emerging discipline." en. In: *Molecular Systems Biology* 2.1 (Jan. 2006), n/a–n/a. ISSN: 1744-4292.

[22] A. S. Khalil and J. J. Collins. "Synthetic Biology: Applications Come of Age." In: *Nature reviews. Genetics* 11.5 (May 2010), pp. 367–379. ISSN: 1471-0056.

[23] J. Liang, R. Bloom, and C. Smolke. "Engineering Biological Systems with Synthetic RNA Molecules." In: *Molecular Cell* 43.6 (Sept. 2011), pp. 915–926. ISSN: 1097-2765.

[24] J. Kim et al. "De novo-designed translation-repressing riboregulators for multi-input cellular logic." In: *Nature Chemical Biology* 15.12 (Dec. 2019), pp. 1173–1182. ISSN: 1552-4469.

[25] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. "ViennaRNA Package 2.0." In: *Algorithms for Molecular Biology* 6 (2011), p. 26. ISSN: 1748-7188.

[26] M. H. Bailor, X. Sun, and H. M. Al-Hashimi. "Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation." en. In: *Science* 327.5962 (Jan. 2010), pp. 202–206. ISSN: 0036-8075, 1095-9203.

[27] A. R. Banerjee, J. A. Jaeger, and D. H. Turner. "Thermal unfolding of a group I ribozyme: the low-temperature transition is primarily disruption of tertiary structure." In: *Biochemistry* 32.1 (Jan. 1993), pp. 153–163. ISSN: 0006-2960.

[28] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. "NUPACK: Analysis and design of nucleic acid systems." eng. In: *Journal of Computational Chemistry* 32.1 (Jan. 2011), pp. 170–173. ISSN: 1096-987X.

[29] J. S. Reuter and D. H. Mathews. "RNAstructure: Software for RNA Secondary Structure Prediction and Analysis." In: *BMC bioinformatics* 11 (Mar. 2010), p. 129. ISSN: 1471-2105.

[30] Y. Ding and C. E. Lawrence. "A Statistical Sampling Algorithm for RNA Secondary Structure Prediction." In: *Nucleic Acids Research* 31.24 (Dec. 2003), pp. 7280–7301. ISSN: 1362-4962.

[31] J. H. Havgaard, R. B. Lyngsø, and J. Gorodkin. "The FOLDALIGN Web Server for Pairwise Structural RNA Alignment and Mutual Motif Search." In: *Nucleic Acids Research* 33.Web Server issue (July 2005), W650–653. ISSN: 1362-4962.

[32] K. Sato, M. Hamada, K. Asai, and T. Mituyama. "CENTROIDFOLD: A Web Server for RNA Secondary Structure Prediction." In: *Nucleic Acids Research* 37.Web Server issue (July 2009), W277–280. ISSN: 1362-4962.

[33] C. B. Do, C.-S. Foo, and S. Batzoglou. "A Max-Margin Model for Efficient Simultaneous Alignment and Folding of RNA Sequences." In: *Bioinformatics (Oxford, England)* 24.13 (July 2008), pp. i68–76. ISSN: 1367-4811.

[34] S. Janssen and R. Giegerich. "The RNA Shapes Studio." In: *Bioinformatics (Oxford, England)* 31.3 (Feb. 2015), pp. 423–425. ISSN: 1367-4811.

[35] M. Parisien and F. Major. "The MC-Fold and MC-Sym Pipeline Infers RNA Structure from Sequence Data." In: *Nature* 452.7183 (Mar. 2008), pp. 51–55. ISSN: 1476-4687.

[36] M. Seetin and D. Mathews. "RNA Structure Prediction: An Overview of Methods." In: *Methods in molecular biology (Clifton, N.J.)* 905 (June 2012), pp. 99–122.

[37] A. Churkin, M. D. Retwitzer, V. Reinharz, Y. Ponty, J. Waldispühl, and D. Barash. "Design of RNAs: comparing programs for inverse RNA folding." In: *Briefings in Bioinformatics* 19.2 (Mar. 1, 2018), pp. 350–358. ISSN: 1477-4054.

[38] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. "Design of Multistable RNA Molecules." In: *RNA (New York, N.Y.)* 7.2 (Feb. 2001), pp. 254–265. ISSN: 1355-8382.

[39] S. Badelt. "Control of RNA Function by Conformational Design." In: (2016).

[40] S. Hammer, B. Tschiatschek, C. Flamm, I. L. Hofacker, and S. Findeiß. "RNAblueprint: Flexible Multiple Target Nucleic Acid Sequence Design." In: *Bioinformatics* 33.18 (Sept. 2017), pp. 2850–2858. ISSN: 1367-4803.

[41] C. Berens and B. Suess. "Riboswitch engineering — making the all-important second and third steps." In: *Current Opinion in Biotechnology*. Analytical Biotechnology 31 (Feb. 2015), pp. 10–15. ISSN: 0958-1669.

[42] A. A. Nielsen, T. H. Segall-Shapiro, and C. A. Voigt. "Advances in genetic circuit design: novel biochemistries, deep part mining, and precision gene expression." In: *Current Opinion in Chemical Biology*. Synthetic biology • Synthetic biomolecules 17.6 (Dec. 2013), pp. 878–892. ISSN: 1367-5931.

[43] S. Brantl and E. G. H. Wagner. "Antisense RNA-mediated transcriptional attenuation: an in vitro study of plasmid pT181." en. In: *Molecular Microbiology* 35.6 (2000), pp. 1469–1482. ISSN: 1365-2958.

[44] W. Winkler, A. Nahvi, and R. R. Breaker. "Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression." en. In: *Nature* 419.6910 (2002), pp. 952–956. ISSN: 0028-0836.

[45] F. J. Isaacs, D. J. Dwyer, C. Ding, D. D. Pervouchine, C. R. Cantor, and J. J. Collins. "Engineered riboregulators enable post-transcriptional control of gene expression." eng. In: *Nature Biotechnology* 22.7 (July 2004), pp. 841–847. ISSN: 1087-0156.

[46] J. M. Callura, C. R. Cantor, and J. J. Collins. "Genetic switchboard for synthetic biology applications." en. In: *Proceedings of the National Academy of Sciences* 109.15 (Apr. 2012), pp. 5850–5855. ISSN: 0027-8424, 1091-6490.

[47] J. B. Lucks, L. Qi, V. K. Mutalik, D. Wang, and A. P. Arkin. "Versatile RNA-sensing transcriptional regulators for engineering genetic networks." en. In: *Proceedings of the National Academy of Sciences* 108.21 (May 2011), pp. 8617–8622. ISSN: 0027-8424, 1091-6490.

[48] V. K. Mutalik, L. Qi, J. C. Guimaraes, J. B. Lucks, and A. P. Arkin. "Rationally designed families of orthogonal RNA regulators of translation." en. In: *Nature Chemical Biology* 8.5 (2012), pp. 447–454. ISSN: 1552-4450.

[49] G. Rodrigo, T. E. Landrain, and A. Jaramillo. "De novo automated design of small RNA circuits for engineering synthetic riboregulation in living cells." en. In: *Proceedings of the National Academy of Sciences* 109.38 (Sept. 2012), pp. 15271–15276. ISSN: 0027-8424, 1091-6490.

[50] J. Chappell, M. K. Takahashi, and J. B. Lucks. "Creating small transcription activating RNAs." en. In: *Nature Chemical Biology* 11.3 (Mar. 2015), pp. 214–220. ISSN: 1552-4469.

[51] L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, and W. A. Lim. "Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression." In: *Cell* 152.5 (Feb. 28, 2013), pp. 1173–1183. ISSN: 0092-8674.

[52] R. Lutz and H. Bujard. "Independent and Tight Regulation of Transcriptional Units in Escherichia Coli Via the LacR/O, the TetR/O and AraC/I1-I2 Regulatory Elements." In: *Nucleic Acids Research* 25.6 (Mar. 1, 1997), pp. 1203–1210. ISSN: 0305-1048.

[53] V. A. Rhodius et al. "Design of Orthogonal Genetic Switches Based on a Crosstalk Map of σs, Anti-σs, and Promoters." In: *Molecular Systems Biology* 9.1 (Jan. 2013), p. 702. ISSN: 1744-4292.

[54] A. Green, P. Silver, J. Collins, and P. Yin. "Toehold Switches: De-Novo-Designed Regulators of Gene Expression." In: *Cell* 159.4 (Nov. 6, 2014), pp. 925–939. ISSN: 0092-8674.

[55] J. Chappell, M. K. Takahashi, S. Meyer, D. Loughrey, K. E. Watters, and J. Lucks. "The centrality of RNA for engineering gene expression." en. In: *Biotechnology Journal* 8.12 (2013), pp. 1379–1395. ISSN: 1860-7314.

[56] I. Dotu, J. A. Garcia-Martin, B. L. Slinger, V. Mechery, M. M. Meyer, and P. Clote. "Complete RNA inverse folding: computational design of functional hammerhead ribozymes." en. In: *Nucleic Acids Research* 42.18 (Oct. 2014), pp. 11752–11762. ISSN: 0305-1048, 1362-4962.

[57] J. N. Zadeh, B. R. Wolfe, and N. A. Pierce. "Nucleic acid sequence design via efficient ensemble defect optimization." en. In: *Journal of Computational Chemistry* 32.3 (Feb. 2011), pp. 439–452. ISSN: 1096-987X.

[58] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. "Fast folding and comparison of RNA secondary structures." In: *Monatshefte für Chemie / Chemical Monthly* 125.2 (Feb. 1, 1994), pp. 167–188. ISSN: 1434-4475.

[59] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. "From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures." In: *Proceedings. Biological Sciences* 255.1344 (Mar. 1994), pp. 279–284. ISSN: 0962-8452.

[60] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. "Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration I. Neutral Networks." In: *Monatshefte für Chemie / Chemical Monthly* 127.4 (Apr. 1996), pp. 355–374. ISSN: 1434-4475.

[61] C. Reidys, P. F. Stadler, and P. Schuster. "Generic Properties of Combinatory Maps: Neutral Networks of RNA Secondary Structures." In: *Bulletin of Mathematical Biology* 59.2 (Mar. 1997), pp. 339–397. ISSN: 0092-8240.

[62] A. Esmaili-Taheri, M. Ganjtabesh, and M. Mohammad-Noori. "Evolutionary Solution for the RNA Design Problem." In: *Bioinformatics (Oxford, England)* 30.9 (May 2014), pp. 1250–1258. ISSN: 1367-4811.

[63] A. Saragliadis, S. S. Krajewski, C. Rehm, F. Narberhaus, and J. S. Hartig. "Thermozymes: Synthetic RNA Thermometers Based on Ribozyme Activity." In: *RNA biology* 10.6 (June 2013), pp. 1010–1016. ISSN: 1555-8584.

[64] G. Nechooshtan, M. Elgrably-Weiss, A. Sheaffer, E. Westhof, and S. Altuvia. "A pH-responsive Riboregulator." In: *Genes & Development* 23.22 (Nov. 2009), pp. 2650–2662. ISSN: 1549-5477.

[65] S. H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. F. Stadler, and I. L. Hofacker. "Partition Function and Base Pairing Probabilities of RNA Heterodimers." In: *Algorithms for molecular biology: AMB* 1.1 (Mar. 2006), p. 3. ISSN: 1748-7188.

[66] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. "Thermodynamics of RNA-RNA Binding." In: *Bioinformatics (Oxford, England)* 22.10 (May 2006), pp. 1177–1182. ISSN: 1367-4803.

[67] J. A. Garcia-Martin, P. Clote, and I. Dotu. "RNAiFOLD: A Constraint Programming Algorithm for RNA Inverse Folding and Molecular Design." In: *Journal of Bioinformatics and Computational Biology* 11.2 (Apr. 2013), p. 1350001. ISSN: 1757-6334.

[68] E. I. Ramlan and K.-P. Zauner. "Design of Interacting Multi-Stable Nucleic Acids for Molecular Information Processing." In: *Bio Systems* 105.1 (July 2011), pp. 14–24. ISSN: 1872-8324.

[69] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin. "Coding-sequence determinants of gene expression in Escherichia coli." In: *Science (New York, N.Y.)* 324.5924 (Apr. 2009), pp. 255–258. ISSN: 0036-8075.

[70] K. Bentele, P. Saffert, R. Rauscher, Z. Ignatova, and N. Blüthgen. "Efficient translation initiation dictates codon usage at gene start." In: *Molecular Systems Biology* 9 (June 2013), p. 675. ISSN: 1744-4292.

[71] B. Yurke, A. J. Turberfield, A. P. Mills, F. C. Simmel, and J. L. Neumann. "A DNA-fuelled molecular machine made of DNA." In: *Nature* 406.6796 (Aug. 2000), pp. 605–608. ISSN: 0028-0836, 1476-4687.

[72] K. Pardee, A. Green, T. Ferrante, D. Cameron, A. DaleyKeyser, P. Yin, and J. Collins. "Paper-Based Synthetic Gene Networks." In: *Cell* 159.4 (Nov. 2014), pp. 940–954. ISSN: 00928674.

[73] J. Kim, P. Yin, and A. A. Green. "Ribocomputing: Cellular Logic Computation Using RNA Devices." In: *Biochemistry* 57.6 (Feb. 13, 2018), pp. 883–885. ISSN: 0006-2960, 1520-4995.

[74] A. A. Green, J. Kim, D. Ma, P. A. Silver, J. J. Collins, and P. Yin. "Complex cellular logic computation using ribocomputing devices." In: *Nature* 548.7665 (Aug. 2017), pp. 117–121. ISSN: 0028-0836, 1476-4687.

[75] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer-Verlag, 2006. ISBN: 978-0-387-31073-2.

[76] Z. Ghahramani. "Probabilistic machine learning and artificial intelligence." eng. In: *Nature* 521.7553 (May 2015), pp. 452–459. ISSN: 1476-4687.

[77] X. Zhang, M. L. Acencio, and N. Lemke. "Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review." In: *Frontiers in Physiology* 7 (Mar. 2016). ISSN: 1664-042X.

[78] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Nov. 9, 2012. 679 pp. ISBN: 978-1-118-58600-6.

[79] I. Guyon and C. Aliferis. "Causal Feature Selection." In: (Oct. 2007).

[80] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection." In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[81] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. "Fast folding and comparison of RNA secondary structures." In: *Monatshefte für Chemie / Chemical Monthly* 125".2 (1994), pp. 167–188. ISSN: 1434-4475.

[82] J. M. Schaeffer, C. Thachuk, and E. Winfree. "Stochastic Simulation of the Kinetics of Multiple Interacting Nucleic Acid Strands." In: *DNA Computing and Molecular Programming - 21st International Conference, DNA 21, Boston and Cambridge, MA, USA, August 17-21, 2015. Proceedings*. 2015, pp. 194–211.

[83] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021.

[84] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker. "The Vienna RNA Websuite." In: *Nucleic Acids Research* 36 (suppl_2 July 1, 2008), W70–W74. ISSN: 0305-1048.

[85] M. Kuhn. "Building Predictive Models in R Using the caret Package." In: *Journal of Statistical Software, Articles* 28.5 (2008), pp. 1–26. ISSN: 1548-7660.

[86] M. Kuhn. *caret: Classification and Regression Training*. R package version 6.0-90. 2021.

[87] T. M. Mitchell. *Machine Learning*. 1st ed. USA: McGraw-Hill, Inc., 1997. 432 pp. ISBN: 978-0-07-042807-2.

[88] T. L. based on Fortran code by Alan Miller. *leaps: Regression Subset Selection*. R package version 3.1. 2020.

[89] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure1." In: *Journal of Molecular Biology* 288.5 (May 1999), pp. 911–940. ISSN: 0022-2836.

[90] M. J. Serra and D. H. Turner. "Predicting thermodynamic properties of RNA." In: *Methods in Enzymology*. Energetics of Biological Macromolecules 259 (Jan. 1995), pp. 242–261. ISSN: 0076-6879.

[91] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. "Incorporating Chemical Modification Constraints into a Dynamic Programming Algorithm for Prediction of RNA Secondary Structure." In: *Proceedings of the National Academy of Sciences* 101.19 (May 2004), pp. 7287–7292.

[92] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer Science & Business Media, May 17, 2013. 595 pp. ISBN: 978-1-4614-6849-3.