The use of Biological Riboswitches and Ribozymes as Scaffolds for Selection and Novel RNA Devices

Ely Blanton Porter

University of Colorado Boulder, elyporter@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/chem_gradetds

Part of the Biochemistry Commons, and the Molecular Genetics Commons

Recommended Citation
https://scholar.colorado.edu/chem_gradetds/143

This Dissertation is brought to you for free and open access by Chemistry & Biochemistry at CU Scholar. It has been accepted for inclusion in Chemistry & Biochemistry Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
The use of Biological Riboswitches and Ribozymes as Scaffolds for Selection and Novel RNA Devices

by

Ely B. Porter

B.S., Biology, The Ohio State University, 2009

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry and Biochemistry

2015
This thesis entitled:
The use of Biological Riboswitches and Ribozymes as Scaffolds for Selection and Novel RNA Devices
written by Ely B. Porter
has been approved for the Department of Chemistry and Biochemistry

_________________________________
Robert T. Batey

_________________________________
James A. Goodrich

Date _________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
The utility of RNA as a sensor for in vivo enzyme evolution, metabolomics monitoring, and bacterial control modules is slowly being realized, but is met with some outstanding challenges. Current work to date on genetically encoded RNA-biosensors has focused on independently developing modular sensors or adaptors. The engineering of adaptor domains has been generally successful in their broad application and modularity, with small molecule fluorescent activators and several strategies for gene regulation now at our disposal. The sensor domains primarily consist of naturally occurring riboswitches and synthetic aptamers generated by in vitro selection. The riboswitch aptamer domains have seen wide application due to their modularity and high affinity, yet their diversity is limited to natural availability. In contrast, hundreds of synthetic aptamers have been created, yet their modularity and application has been met with limited success. Several studies have now shown the faults in assuming that the transition from aptamer to device sensor would be trivial. In this work, a recurrent and model RNA motif is used to introduce robust folding and complexity to a limited library that is selected to bind 5-hydroxytryptophan. The introduced peripheral motif benefits the binding and structure of the RNA, producing an aptamer with higher affinity and specificity than sequences lacking the greater fold. The developed workflow relies on deep sequencing and bioinformatics for the identification of robust elements, reducing the need for tedious validation and allowing the immediate screening of candidate sequences. Additionally, when multiple scaffolds are employed to host limited libraries and parallel
selections are carried out, a suite of high affinity aptamers are obtained that are functionally related, yet distant in sequence space. This allows for an aptameric screening strategy, where a desired adaptor platform may be selected and candidate sensors screened for function, an ability not readily afforded by traditional selection techniques. These strategies ease constraints on and expedite the development of high affinity RNA devices, adding yet another robust component to the synthetic biologist’s toolbox.
Dedication

To Archimedes who welcomed me home with a smile, no matter how good or bad the day,
and to my family for the support you can never ask for.
Acknowledgements

I must begin by thanking Rob Batey for endless advice and guidance while giving me the opportunity to explore the scientific world around me. Debbie Wuttke also deserves a great deal of thanks for her thoughtful and pleasant conversation. Additional thanks to my committee members Roy Parker, Norm Pace, and especially Jim Goodrich for his reading of this thesis. I give credit to the lab members themselves for making this education enjoyable. The daily banter and critical thought of the collective group would make this experience difficult to replicate elsewhere. It was a true pleasure and privilege to mentor the exceptional undergraduate researchers Makenna Morck and Sam Webster. I can only hope I taught them as much as they taught me.

There is a long list of friends to thank. Without them this endeavor would have never been possible. They created a family and a home for me here in Boulder, the type that I would not hesitate to ask anything of them knowing they are always there. The running adventures, mountain escapades, and associated shenanigans are memories I cannot forget. The amount of gratitude my parents and brother deserve cannot be expressed in words. They are always understanding and go beyond any reasonable expectation of support and encouragement.
Contents

Chapter

1 Introduction

1.1 Synthetic biology and the utility of RNA ............................ 1
1.2 Riboswitches are natural parts .......................................... 5
1.3 Plasticity of the aptamer domain ....................................... 10
1.4 Expression platforms as adaptors in synthetic biology .......... 14
1.5 Aptazyme regulatory domains .......................................... 17
1.6 Fluorescent RNA as a coupled adapter and actuator .......... 19
1.7 A case for reprogramming natural riboswitches ................... 22

2 Investigation of the Lysine Aptamer Domain as a Host for Orthogonal Compounds

2.1 The lysine aptamer ...................................................... 26
2.2 Ligand specificity of the lysine riboswitch .......................... 27
2.3 The crystal structure of lysine analogs bound to the lysine riboswitch .... 31
2.4 Reprogramming the lysine aptamer domain ......................... 35
2.5 Optimization of the lysine library sequence and priming regions ... 40
2.6 Discussion .............................................................. 45
2.7 Materials and methods ................................................ 46
## Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Lysine Analog Crystalography Statistics</td>
<td>33</td>
</tr>
<tr>
<td>3.1</td>
<td>xB$_{17}$ Crystalography Statistics</td>
<td>66</td>
</tr>
<tr>
<td>4.1</td>
<td>ITC Table of Affinities</td>
<td>121</td>
</tr>
<tr>
<td>4.2</td>
<td>Broccoli Sensors Show Variable Response Ranges and Affinities</td>
<td>134</td>
</tr>
<tr>
<td>5.1</td>
<td>Proof of Concept $in~vivo$ RNA devices</td>
<td>144</td>
</tr>
</tbody>
</table>
Figures

Figure

1.1 Programming Gene Expression ........................................... 3
1.2 Overview of Selection Method ............................................. 6
1.3 Prevalence of Natural Riboswitches ................................. 8
1.4 Logic of Riboswitch Regulation ......................................... 9
1.5 Ligand Adaptability by Purine Riboswitches ..................... 12
1.6 Biosynthetic Parts and Composability of Riboregulators .......... 15
1.7 A Structural Understanding of a Fluorescent RNA ............... 20

2.1 Lysine Structure and Fluorescent Reporter ....................... 28
2.2 Adaptability of the Lysine Binding Pocket ....................... 30
2.3 Structural Analysis of Lysine Analogs ............................ 34
2.4 Initial Lysine Design and Selection ................................. 36
2.5 SHAPE and ITC Characterization of 5htp and GMP Aptamers .... 38
2.6 Redesign and Expansion of the Lysine Library .................. 41
2.7 Primer Site Optimization and Site Specific Deletions .......... 44

3.1 The Purine Riboswitch Architecture ............................... 53
3.2 Library Sequence, Design, and Selection .......................... 55
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3 Library Sequencing and Cluster Analysis</td>
<td>57</td>
</tr>
<tr>
<td>3.4 Selection Progression and xB Cluster Evolution</td>
<td>59</td>
</tr>
<tr>
<td>3.5 SHAPE Characterization of the $xpt$ selection</td>
<td>61</td>
</tr>
<tr>
<td>3.6 ITC Characterization of the $xpt$ selection</td>
<td>63</td>
</tr>
<tr>
<td>3.7 Density Map of $xB_{17}$</td>
<td>67</td>
</tr>
<tr>
<td>3.8 Overview of the $xB_{17}$ Structure</td>
<td>68</td>
</tr>
<tr>
<td>3.9 Detailed Secondary Structure and Base Pair Interactions</td>
<td>70</td>
</tr>
<tr>
<td>3.10 The New Junction is a Complex Structure</td>
<td>71</td>
</tr>
<tr>
<td>3.11 The $xB_{17}$ Binding Pocket is a T-loop</td>
<td>73</td>
</tr>
<tr>
<td>3.12 SHAPE Signatures are Further Clarified by the Structure</td>
<td>75</td>
</tr>
<tr>
<td>3.13 Strategy for $xB_{17}$ in the $metE$ Platform</td>
<td>79</td>
</tr>
<tr>
<td>3.14 Strategy for $xB_{17}$ Coupled to On Switches</td>
<td>82</td>
</tr>
<tr>
<td>3.15 The Malachite Green Fluorescent RNA Reporter</td>
<td>84</td>
</tr>
<tr>
<td>3.16 The Spinach Fluorescent RNA Sensor</td>
<td>87</td>
</tr>
<tr>
<td>3.17 Revised Minimal Spinach Platform and Optimization</td>
<td>89</td>
</tr>
<tr>
<td>4.1 A new Termostable and Processive RT</td>
<td>107</td>
</tr>
<tr>
<td>4.2 Design of Additional Scaffolds</td>
<td>112</td>
</tr>
<tr>
<td>4.2 Continued</td>
<td>113</td>
</tr>
<tr>
<td>4.3 Thermostable RT and Platform Optimization Reduces Error</td>
<td>114</td>
</tr>
<tr>
<td>4.4 Pairwise Identity Trees of the Three selections</td>
<td>116</td>
</tr>
<tr>
<td>4.5 Sequence Analysis of Candidate Clusters</td>
<td>118</td>
</tr>
<tr>
<td>4.6 Structural Probing and Analysis of $cdG_{17}$</td>
<td>122</td>
</tr>
<tr>
<td>4.7 Structural Probing and Analysis of $Fhh_4$</td>
<td>124</td>
</tr>
<tr>
<td>4.8 Structural Probing and Analysis of Three Purine Scaffolded Aptamers</td>
<td>125</td>
</tr>
</tbody>
</table>
4.9 Aptameric Screen for Transcriptional Attenuators ......................... 127
4.10 P₁ in the context of pbuE is a robust switcht ............................. 129
4.11 Aptameric Screeing Strategy of Broccoli Sensors ......................... 132
4.12 Magnesium Effects on the Broccoli Sensors .............................. 133
A.1 Mapping file example .......................................................... 165
A.2 Overview of sequencing analysis pipeline .................................. 166
A.3 Comparison of the Branched Workflow ................................. 170
Chapter 1

Introduction

1.1 Synthetic biology and the utility of RNA

The past thirty years since the discovery of nucleic acid catalysis has seen an exponential growth in our understanding of RNA as more than its centrality in genetic information transfer [1]. RNA regulatory elements have now been observed in all forms of life exerting their control and function at many levels within the cell. These realizations show that RNA is an ubiquitous and critical part in maintaining cellular homeostasis and responding to numerous environmental stimuli. The functions that RNA possess are vast and diverse, ranging from metabolic housekeeping, to biofilm formation, and antibiotic resistance to name a few [2–5]. The survival of an organism is predicated on its ability to adapt in the face of changing physical and chemical environments with many sophisticated and complex tools available to the cell, including long known and highly capable protein based mechanisms. However, the persistence and diversity of RNA regulatory mechanisms in the face of protein tools thought to be more competent argues that they are a critical component in the life and adaption of an organism [6].

In a biological context, RNA is a robust means of effecting numerous cellular activities. One of the first observed examples of RNA mediated control was in the bacterial \textit{trp} operon,
where the TRAP\textsuperscript{1} protein interrogates the intracellular tryptophan pool and conditionally binds the leader sequence of the mRNA [7]. Upon TRAP’s recognition of tryptophan, the protein binds the leader sequence of the operon and induces an RNA mediated termination signal, whereas in its absence the RNA forms an alternative secondary structure and the gene is expressed. This is just but one early example in our understanding of the roles that RNA play in cells, with the list of functions now vast that includes chromatin modification, gene silencing, post-transcriptional regulation and many more.

Along with the realization of RNA as a regulatory element, synthetic biology has concurrently advanced as a discipline. Defined as engineering biology via the synthesis of complex, biologically inspired systems into novel functions that do not naturally occur [10], synthetic biology builds upon the understanding of natural systems and uses the fundamental, simplified parts as tools for the creation of more complex devices and systems. Over time, more parts are discovered and characterized allowing for the creation of novel pathways and functions either \textit{in vitro} or a host organism. These systems can be carried out on a macroscopic scale in a very rational, standardized, and systematic manner once the basic parts are reasonably understood [11,12].

Early synthetic networks were comprised of simple and fundamental control systems from bacteria. They were built from our understanding of promoter intensities [13–15], translation initiation elements [14,16], and previously characterized repressor elements [17]. The culmination of these advances might best be illustrated by the semi-synthetic production of artemisinin, a potent anti-malarial whose production was previously subject to the availability of the plant \textit{Artemisia annua} leading to unstable supply and price fluctuation making the drug unavailable to those most at risk [18]. To remedy this, a synthetic metabolic pathway

\textsuperscript{1} tryptophan RNA-binding attenuation protein
Figure 1.1: Overview and examples of synthetic parts and devices used to control the expression of heterologous enzymes. (a) Three examples are given that illustrate current themes in controlled gene expression. The AraC* system regulating E1 has undergone massive optimization via directed protein evolution in order to minimize cross-talk with the lac repressor [8], a common theme with trans-acting protein modulators. The systems controlling E2 and E3 are a combination of protein and RNA elements working concurrently to modulate enzyme levels in the pathway. E1 and E3 exert their function at the DNA level, while E2 is able to regulate at both the transcriptional (protein) and post-transcriptional (RNA-aptzyme) levels incorporating multiple inputs into a single output while maintaining a compact and modular genetic footprint. (c) These genetic tools modulate protein stoichiometries of the corresponding enzymatic pathway, yet the current availability of these tools remains limited and tedious to create. Figure taken from [9]
was implemented in a host chassis (*S. cerevisiae*) that used many various genes from bacteria, yeast, and plant species to overproduce isoprenoids and ultimately artemisinic acid [19]. While this example illustrates the utility of synthetic biology to ease monetary tensions around a global health problem, there exists additional considerations. One issue is that the synthetic host “bioreactor” relies entirely on pre-existing parts (including promiscuous enzymes and scarce regulatory devices) with intensive screening to optimize the pathway [20]. During construction of the synthetic circuit to produce the drug, some elements required development in *E. coli* [21] before implementation in *S. cerevisiae* in addition to the orthogonal enzyme screening, promoter swapping, and genomic integration of parts that was necessary to produce industrially useful yields [20, 22]. This feat of engineering illustrates the imperfection of current strategies. There is an incomplete understanding of the parts needed to create these pathways and screening is used as a substitute. The outcome of combinatorial screening is still unpredictable and often yields lackluster responses on the order of 2 to 4 fold [23]. Hence there is a critical need for improved methods in enzyme development and an expansive library of modular regulatory devices [9].

A strong case can be made for RNA as a robust and predictable node in a gene regulation engineering platform. RNA control systems provide unique features in contrast to their protein counterparts, namely the ability to exert control in a *cis* regulatory fashion on a single molecule [24]. Since their action is physically constrained to the associated gene partner, RNA devices are more likely to exert independence from each other and other cellular processes, limiting toxicity and unintended effects [25]. RNA is also highly tunable, since it folds into higher order structure predictably according to base-pairing rules [26, 27]. Thus a single parental RNA part can be systematically altered to create a suite of orthogonal riboregulators whose response is variable yet still dependent on the intended
input [28]. RNA also folds into small, compact domains which minimizes the genetic footprint of these elements, aids in their modularity [29], and increases their portability into unrelated organisms [28]. Therefore, the individual parts can be assembled into higher-order devices according to general and programmable rules leading to combinatorial regulation and genetic circuits in a robust and predictable manner, the same for which cannot be said for trans-acting protein counterparts [30].

In addition to RNA meeting all necessary requirements for a generalized and easily engineered gene regulation platform, its greatest strength may yet lie in the power of *in vitro* selection (SELEX)\(^2\) [31–33]. The process of selection is enabled by the ability to synthesize large randomized libraries that can exceed \(>1 \times 10^{15}\) possible solutions. This sampling size allows for the screening of a function and iterative rounds of selection and amplification serve to drive the screened pool to relatively few, but highly competent solutions (Figure 1.2). Selection techniques are able to produce custom aptamers to nearly any imaginable small molecule, infinitely expanding the conceivable stimuli to which RNA can respond

### 1.2 Riboswitches are natural parts

Nature has in fact already evolved and optimized various RNA parts for the control and regulation of genes [29,34,35]. Among these and the most pervasive are small RNA devices termed Riboswitches [36]. These regulatory RNA elements are most often found in the leader sequences of bacterial mRNAs, controlling expression in a cis-fashion. Riboswitches were originally discovered as conserved sequence elements upstream of genes controlling the transport or biosynthesis of systems such as \(S\)-adenosylmethionine (SAM) [37,38], thiamine pyrophosphate (TPP) [39], lysine [40], riboflavin [41], adenosylcobalamin [42] and purines

---

\(^2\) Systematic Evolution of Ligands by Exponential enrichment
Figure 1.2: The in vitro selection of target specific RNA aptamers The starting point for SELEX is a synthetic randomized DNA library, that is then amplified and transcribed into an RNA library. The library is then subjected to repetitive steps of binding, partitioning, competitive elution, amplification, and conditioning. In the first step of the selection cycle, the library and target molecule are allowed to interact and equilibrate. Unbound species are then removed by several washings of the binding complexes, and target-bound oligonucleotides are competitively eluted from the solid phase. The eluted fraction is then enzymatically amplified and conditioned to return to the beginning of the cycle. The result is an enriched pool, that can be further enriched by successive rounds of selection or sequenced for analysis at this point.
The observation of these numerous elements lead to the prescient speculation that the mRNAs directly bind their small molecule effectors [44], which was later shown to be true for all of these systems [36, 38, 45–49]. This conclusively demonstrated that riboswitches are able direct their own expression and respond to physiological stimuli in the absence of additional protein mediators [50].

The list of effector molecules that these RNA recognize and the genes they control is now understood to be extensive (Figure 1.3), including cyclic nucleotide second messengers [5, 52], the levels of amino-acylated tRNA [56, 57], and even small metal anions [58, 59] among many others [60]. Riboswitches are also a significant regulator within the cell controlling greater than 4% of genes in organisms such as B. subtilis under the control of conserved sequence motifs [36]. While riboswitches are most abundant in 5’ leader sequences of bacterial genomes, examples have also been identified in eukaryotic introns and 3’ UTRs [61]. This provides strong evidence that these simple RNA mechanisms are able to function in a variety of contexts regardless of the host organism.

The activity of a riboswitch is imparted through two independent functional domains: a small molecule binding aptamer (or sensor) domain and an expression platform (or adaptor) containing a structural switch that most often acts at the level of transcription or translation [2]. Most riboswitches in bacterial systems that exert their control at the transcriptional level do so via the formation of two alternative stem-loop structures, a rho-independent transcriptional terminator (T) and an anti-terminator (AT) (Figure 1.4). The main communication between these two states is mediated by a switching sequence, or a stretch of nucleotides whose presence is mutually exclusive between the two states and is informed by the presence of the effector molecule. The expression platforms are highly varied and no clear consensus exists when the aptamer domains are aligned [63]. While this makes characterization of the
The term riboswitch was established to define RNAs that control gene expression by binding metabolites without the need for protein factors. More recently, the name has begun to be used for riboswitch-like RNAs that respond to changes in temperature (Johansson 2009; Klinkert and Narberhaus 2009), tRNA binding (Gutiérrez-Preciado et al. 2009), or metal ion binding (Cromie et al. 2006; Dann et al. 2007). Although the functions of the RNAs encompassed by this expanded definition certainly would have been useful in an RNAW orld, the discussion hereafter will be focused on riboswitches that have evolved to respond to small organic compounds.

Riboswitches need to form molecular architectures with sufficient complexity to carry out two main functions: molecular recognition and conformational switching. Simple riboswitches each carry one aptamer that senses a single ligand and one expression platform that usually controls gene expression via a single mechanism. Because only four types of monomers are used by RNA to form selective binding pockets for target metabolites, aptamer sequences and structures tend to be strikingly well conserved over great evolutionary distances (e.g., see Grundy and Henkin 1998; Gelfand et al. 1999; Sudarsan et al. 2003; Nahvi et al. 2004). This sequence and structure conservation serves as the basis for assigning riboswitch representatives to specific classes (Fig. 1).

![Riboswitch class representatives](image_url)

**Figure 1.3: Riboswitches are a prevalent means of genetic regulation in bacteria and are found in all domains of life [51].** The various riboswitches are rank ordered by decreasing genome frequency with filled bars indicating some level of validation for metabolite recognition for which the class is named. Open bars indicate orphan or uncharacterized riboswitches named for commonly associated genes (current as of 2013), though *ydaO* has now been characterized as a c-di-AMP binding riboswitch [52–54] and *pfl* binds ZMP/ZTP while regulating purine biosynthesis [55].
Background and Motivation

Riboswitches are a prevalent means of gene regulation in the bacterial domain generally located in the 5' leader sequence of mRNA. Bacterial riboswitches primarily operate via the binding of a metabolite in the recognition, or aptamer, domain (shown in gray) eliciting a secondary conformational change in a coupled, downstream region of the mRNA known as the regulation domain, or expression platform (cyan). In this example, recognition of the metabolite (orange circle) stabilizes a secondary conformation that promotes polymerase read-through and formation of an anti-terminator (AT). In the absence of ligand, an alternative structure is formed promoting the formation of a Rho-independent terminator stem (T) leading to early termination of the mRNA.

Figure 1.4: Riboswitches display differential structural states based on the presence of an effector molecule. Riboswitches primarily operate via the binding of a metabolite to the aptamer domain (gray) eliciting a secondary conformational change in a coupled, downstream region of the mRNA that exerts the regulator function (known as the expression platform: cyan). In this example, recognition of the metabolite (orange circle) stabilizes a secondary conformation that sequesters the switching sequence (red) and promotes gene expression via an anti-terminator (AT). In the absence of ligand, an alternative structure dominates and inclusion of the switching sequence in a helical element juxtaposed to a polyuridine stretch promotes a rho-independent terminator stem (T) and early termination of the gene transcript. Adapted from [62]
expression platforms difficult, it makes clear that an aptamer domain is not dependent upon the adaptor for its function, suggesting that it is a modular element capable of regulating in a myriad of contexts [34].

1.3 Plasticity of the aptamer domain

Structural and biophysical studies of riboswitches have almost exclusively focused upon the aptamer domain as this domain has both complex tertiary architecture amenable to structure determination and an observable activity in ligand binding. The three-dimensional architectures of almost every major family of riboswitch have been elucidated, with several riboswitch aptamer domains such as SAM-I [64], TPP [65, 66], and preQ1-I [67–69] being extensively characterized by a number of biophysical approaches. With the accumulation of structural information, it has become clear that riboswitches are able to recognize their cognate ligands in a highly specific manner employing many techniques to specifically recognize the ligand, including shape complementarity, electrostatics, and a unique hydrogen bonding pattern [70]. Often, the main determinants of the binding pocket are conferred by a handful of nucleotides, with the rest of the RNA architecture serving to orient these critical nucleotides [71].

While some riboswitches such as glmS are fairly non-specific and capable of responding to multiple ligands in order to more completely survey the cellular milieu [72], other riboswitches must alter their sequence in order to adapt specificity. The purine riboswitch is a prime example of this phenomena. The binding pocket of the purine class consists of three critical bases for recognition organized by a helical junction and tertiary interactions. The hydrogen bonding network established by these three bases (nucleotides 47, 51, and 74 in the B. subtilis xpt-pbuX numbering scheme) forms a base quartet and dictates the specificity
of the RNA [73]. By altering these nucleotides, the specificity within the purine architecture can easily be swapped (Figure 1.5). The xpt guanine-responsive [73] [66] and Vibrio vulnificus add adenine-responsive [66] riboswitches discriminate between their respective ligands based primarily upon the Watson-Crick pairing of position 74 to the ligand in the binding pocket (C for guanine and U for adenine). Recognition of 2′-deoxyguanosine, on the other hand, is achieved primarily through the identity of the nucleotide 51 being cytosine [74] [75].

In the bound state, C51 shifts towards 74 relative to U51 in the guanine-bound aptamer, sterically accommodating the 2′-deoxyribose sugar. Accordingly, nucleotide 47 disengages from its interactions with residue 51 and reorients itself out towards the solvent (Figure 1.5). Due to these structural nuances the Mesoplasma florum 1A riboswitch has a 100-fold preference for 2′-deoxyguanosine over guanosine [76]. Comparison of the wild type aptamer bound to both compounds revealed that the ribose sugar of guanosine adopts the C3′-endo conformation that prevents hydrogen bonding of the 3′-hydroxyl group with C48 (C56 in M. florum numbering) [74]. This causes C48 to be flipped out towards solvent, which was speculated to destabilize the conformation of J2/3, leading to lower affinity for guanosine.

While the extent of 1A’s sequence variation extends beyond the binding pocket, it appears that the identities of just two nucleotides, 51 and 74, govern discrimination between different purine nucleobases and nucleosides [75]. In principle, the C51/U74 combination in a sequence that traditionally recognizes guanine or adenine would enable their recognition of 2′-deoxyadenosine, although a sequence such as this has not yet been observed in biology [76].

While natural riboswitch aptamers perform well within their native regulatory roles, they are generally not useful as a biosynthetic tool as they cannot easily escape their cognate ligand within the cell. Based on the structural information outlined above, the Micklefield group has developed an orthogonal riboswitch capable of responding to a non-natural small
Figure 1.5: Ligand specificity is easily manipulated via the alteration of only a few nucleotides within the core binding pocket. (a) Three dimensional structure of the *M. florum* 2'-deoxyguanosine aptamer domain with critical features annotated. (b) A detailed view of the loop-loop interaction that is highly conserved and critical for ligand binding shows subtle differences between the dG and guanine riboswitch yet maintains overall structure. (c) Analysis of the binding pocket shows nuanced changes between dG and guanine, with only positions 22 and 51 conferring specificity between the two ligands. (d) A comparison of dG and guanosine bound to the dG riboswitch reveals that the ribose sugar of guanosine adopts the C3'-endo conformation, preventing hydrogen bonding with C56 (C50 in *xpt* numbering) allowing the residue to swing out in solution and resulting in a >100 fold specificity between the two compounds. (e) Re-engineering of the adenine *add* riboswitch to specifically recognize ammeline and azacytosine. Conformation and hydrogen bonding networks remain relatively the same with only slight shifts due to nucleotide composition resulting in specificity. Adapted from [70]
molecule using the *add* adenine-responsive riboswitch as a scaffold [77]. Site-directed mutagenesis was used to create 15 variants of the *add* riboswitch that fully randomized positions 47 and 51, two residues within the binding pocket that are critical for ligand recognition. Each variant was then individually cloned upstream of the chloramphenicol acetyl transferase (CAT) resistance gene and screened against a chemical library of 80 heterocycles for an antibiotic resistance dependent upon ligand concentration. One resulting variant from the screen (U47C,U51C: termed M6) was found to specifically bind ammeline (4,6-diamino-2-hydroxy-1,3,5-triazine) and yielded an RNA element capable of regulating protein expression in a dose-dependent manner.

To demonstrate how these orthogonal riboswitches can assist in the regulatory control of multiple genes, the authors then created dual and single promoter operons in which the natural *add* and selected M6 translational “ON” riboswitches regulate the DSRed and eGFP genes, respectively [78]. In both examples, each gene was specifically induced by its cognate effector (2-aminopurine for *add* and ammeline for M6). Despite the riboswitches only differing by two point mutations, they exhibited no cross-talk and exerted independent control of their respective fluorescent proteins as a function of their synthetic small molecule inducer in the dual promoter system. When the *add* riboswitch is placed between the reporter genes as an intercistronic element, it does display a moderate dependence for both effector ligands in order to achieve a maximal response.

This innovative study suggests the reengineering of naturally occurring riboswitch aptamer domains is a viable solution for the creation of novel genetic control elements. However, the approach of screening molecules against an RNA is not a generalizable technique. The strategy is effectively backwards from the desired scheme as there is no control over the effector that is produced. With no control on what the effector will be, it can complicate
downstream applications by possessing unattractive qualities such as toxicity or the inability to enter cells. A more desirable strategy would be to screen and evolve the RNA, which is non-toxic and genetically encodable, towards an effector that has desirable characteristics.

1.4 Expression platforms as adaptors in synthetic biology

While the aptamer domain, defined as the minimal RNA sequence necessary to achieve high affinity binding of the cognate effector ligand, is structurally well characterized, the structural features of the expression platform are largely unexplored. A serious limitation in the analysis of this region is that most alignments of riboswitches, particularly those found in Rfam [80, 81] generally exclude features other than the easily identifiable and alignable aptamer domain. Structural elements in the expression platform, such as terminator/antiterminator or sequester/antisequester hairpins, are proposed based upon computational secondary structural prediction algorithms. While these programs are good at predicting relatively strong elements of secondary structure, such as rho-independent terminators, regulatory elements can vary significantly in their structure impeding identification of consensus elements and conserved motifs [63].

Expression platforms of natural riboswitches can be used to create synthetic ribosensors, and several case examples exist. It is well appreciated that aptamer domains of riboswitches, like their SELEX-derived counterparts, are highly composable such that they can function in variety of contexts including synthetic riboswitches, aptazymes (aptamer inducible ribozymes), and fluorescent biosensors [34]. However, modularity of expression platforms had not been demonstrated until recently [79, 82]. In some riboswitches, the sequence required for high affinity ligand binding to the aptamer domain does not overlap with the alternative secondary structural switch, particularly those in the P1 helix considered to
Figure 1.6: **Synthesis of artificial RNA devices using composable elements.** (a) A riboswitch can be decomposed into highly modular parts. There are three critical parts to an RNA device: the regulator (riboswitch: gray box), the signal (ligand), and an actuator (gene reporter). The riboswitch can be broken down further into a sensor (aptamer) and expression platform (adapter) whose parts may be recombined into new devices with novel function. (b) Allosteric ribozymes (aptazymes) are built from the same parts as riboswitches, with a catalytic RNA domain serving as the modular adapter. (c) The results of the "Mix-and-Match" strategy using existing, natural riboswitch parts and adapters to create novel riboregulators. The T_{50} is largely dictated by the aptamer domain while levels of termination and dynamic range are more a property of the expression platform. Adapted from [24] and [79].
be central to interdomain communication. An example of such an arrangement is the *B. subtilis metE* transcriptional "OFF" riboswitch, in which the two base pairs essential for S-adenosylmethionine (SAM) binding to the aptamer are localized to the 3'-side of the P1 helix while nucleotides that participate in the P1/P-AT switch are localized to the 5'-side [79]. By defining the boundary of the two domains as the region of P1 where these elements meet, the aptamer can be replaced by a number of alternative sensors, both natural and *in vitro* selected with the resultant chimeras functional both *in vitro* and *in vivo*.

From a mechanistic perspective, the adaptability of select expression platforms suggests these RNAs exploit an "encoded co-transcriptional" folding process [83]. In this model of RNA folding, the secondary structure adopted by the RNA is a function of the relative thermodynamic stability of competing helical elements as well as their 5'-to-3' polarity. This principle was first demonstrated in folding of small model RNAs that have isoenergetic, mutually exclusive structures [83]. For these RNAs, the folding landscape is directed by the sequential ordering of the helical elements. In transcriptional "OFF" expression platforms, ordering of the P1/P-AT elements remain fixed, but the relative stabilities of the helical elements are variable [84]. In the absence of ligand, P-AT is the more stable element, enabling its formation to dominate over P1. Conversely, ligand binding to the aptamer at a site juxtaposed to P1 alters the relative stability of the associated helical element and promotes formation of a rho independent terminator. Thus, the expression platform does not necessarily care about the particulars of the ligand-aptamer complex, only that it communicate the binding event to the expression platform via an increase in thermodynamic stability or a kinetic road block [85], a property that many aptamers are able to convey [84].

A more useful modular switch is one that can induce gene expression in response to ligand binding, like the *add* and M6 riboswitches [77, 78]. However, most of the "ON"
switches surveyed for potential modularity have significant sequence overlap between the two domains, preventing their simple decoupling at a defined boundary like the *metE* riboswitch [84]. Instead, the well-characterized adenine-responsive *pbuE* riboswitch from *B. subtilis* was re-engineered to enable its coupling to different aptamers [84]. In the natural *pbuE* riboswitch, the terminator (P-T) invades into L3, such that the alternative secondary structural switch includes sequence elements of the aptamer necessary for both ligand binding and tertiary structure formation. To decouple the two domains, a short sequence was added to the 3’-side of the P1 helix that prevented the invasion of P-T into the aptamer, mimicking the arrangement observed in *metE*. This engineered variant of the *pbuE* expression platform is able to host a variety of aptamers to turn on gene expression. These studies have only explored a small number of expression platforms, and it may be that even more useful regulatory modules can be created from translational attenuators or even eukaryotic riboswitches that regulate splicing or mRNA stability.

1.5 **Aptazyme regulatory domains**

Since the current iteration of riboswitch platforms have only begun to be explored and still rely on host machinery limiting their portability, intense development has been underway in an alternative platform, the hammerhead ribozyme [86]. Preceding the discovery of a natural allosteric ribozyme [87], Tang, J. and Breaker, R.R. rationalized that since natural catalytic RNAs can greatly affect the expression levels of coupled proteins it might be possible to modulate this activity by appending an RNA that recognizes a small molecule [88]. To accomplish this, the bimolecular hammerhead ribozyme was engineered for reduced kinetics and to host the ATP aptamer [89] within stem II, which had previously been identified as a tunable region for catalytic activity [90,91]. The resulting aptazyme was
capable of repressing substrate strand cleavage in the presence of ATP at a concentration of ligand that approximated the affinity of the appended aptamer. Unfortunately, the minimal hammerhead variant used in this study is difficult to apply in vivo due to the platform’s dependence on high magnesium concentrations, above what is present in a cellular context.

Building upon this, the theophylline aptamer [92] was appended to the hammerhead ribozyme via a simplified screening strategy. The resulting aptazyme showed a 40-fold rate enhancement in the presence of ligand yet the maximal rate was many orders of magnitude slower than the native ribozyme [93]. The Smolke group was then able to show that insertion of a theophylline aptazyme into the 3’ UTR of target mRNA in yeast was able to modulate gene expression in a conditionally dependent manner [94]. As an additional proof of concept, the theophylline aptazyme was used to evolve a novel caffeine demethylase enzyme in yeast. The aptazyme was successful at correlating intracellular theophylline concentrations with GFP expression allowing for high-throughput screening of enzyme libraries via FACS [95]. After several iterations, the improved caffeine demethylase exhibited a 33 fold rate enhancement and 22 fold increase in product selectivity in vivo. Aptazyme development seems to have culminated in the ability to regulate mammalian gene expression, where the addition of theophylline in a drug like manner is able to induce human and mouse T cell proliferation in an in vivo murine model [96].

It’s interesting to note that all of these aptazyme experiments are conducted with a highly compromised, minimal hammerhead ribozyme that lacks sequence elements outside of the catalytic core which confer higher rates of activity and a lowered magnesium dependence [97]. Introduction of these peripheral elements inhibits the design of aptazymes as they remove the conformational instability necessary to couple ligand sensing [90] [91] [97]. Aptazymes are also limited by their dynamic range [96,98,99], making their efficacy in vivo
challenging since there is a trade off between low background and high signal to noise. Therefore, the existence of a truly modular platform capable of moderating gene expression in a broad spectrum of hosts is sorely lacking.

1.6 Fluorescent RNA as a coupled adapter and actuator

RNA has now been ascribed the function of receptor and adaptor in the absence of any protein counterparts, and recent reports have shown the ability of RNA to act as its own actuator. In many of the studies described above, the actuator is a fluorescent protein, a tool used pervasively throughout biology [100]. Fluorescent RNA began development in the late 1990’s with the development of aptamers against sulforhodamine [101] and tetramethyl-rosamine [102]. While the fluorophores are capable of retaining fluorescence when bound to RNA, this method is not entirely useful in biological or engineering practices due to high background fluorescence of free dye and the reversible nature of binding leading to rapid degradation of signal once unbound dye is removed. Research focus then turned towards an RNA element that could induce the fluorescence of a small molecule, such that the unbound compound exhibited low low background.

Studies of self-actuating RNAs began with the selection of a malachite green aptamer. The malachite green aptamer was originally developed as a tool to localize the effects of chromophore-assisted laser inactivation (CALI) in cells [109]. It was shown that by tagging an mRNA in the 5' UTR with the aptamer and then irradiating with high intensity light, free radicals could be generated that cleaved the mRNA. Eventually the aptamer was shown to regulate gene expression in yeast by inhibition of ribosomal scanning [110], yet the desired CALI mechanism was not viable in cells since hydroxy-radical generation by malachite green was not localized to the mRNA in vivo, leading to strong bactericidal effects [111].
Figure 1.7: A structural understanding of an RNA mimic of GFP. The initial structure prediction and scheme for linking aptamers (cyan) to the fluorescent adapter spinach (black) (a). The structure of this RNA was taken for granted and persisted for years as a current review [103] and spinach variant [104] still cite this as a correct structure despite lacking any experimental confirmation. Due to this persistence, studies to increase the thermal stability were largely unsuccessful as regions believed to be involved in base pairing (blue boxes) were actually critical for ligand recognition (b) [105]. This was made evident by the recent crystal structure elucidated the correct secondary structure (c) where bases involved in the formation of a G-quartet (red) were a predicted part of stems 2 and 4 in the original paper [106]. The crystal structure revealed a simple, compact, and common motif [107] (d) that forms a planar platform where the ligand (DFHBI) is housed (e) and allows for the construction of an optimal, truncated version of the aptamer [108]. The structure also revealed a previously unknown dependence on potassium as the formation of the G-quartet prefers this metal ion for its formation (f). Figure adapted from references.
Other groups soon realized that while triphenylmethane dyes, such as malachite green, have extremely low quantum yields, their fluorescence could be rescued by viscous or cold environments that restricted de-excitation vibrations and that the malachite green aptamer could produce the same effect upon binding [112]. The quantum yield of free malachite green is roughly $8 \times 10^{-5}$, yet when bound by RNA it leaps to 0.187, a $>2300$-fold improvement over free fluorophore [112]. Additionally, in the vein of the pioneering aptazyme studies [88] [93], the ATP, FMN, and theophylline aptamers were appended next to the malachite green binding site and shown to regulate fluorescence in a ligand dose dependent manner [113]. The best of these sensors was able to increase fluorescence more than 30 fold over the ligand free state. While these experiments set the stage for RNA to switch on fluorescence, the application of the malachite green aptamer is still limited by the generation of hydroxy-radicals at wavelengths necessary to excite and image the fluorophore [109], likely affecting cellular state.

Many RNA aptamers have since been developed with the ability to switch on the fluorescence of an otherwise dark molecule [114] [115]. The most relevant of these is an RNA mimic of GFP termed “Spinach” [106] (Figure 1.7). A series of GFP chromophore analogs were synthesized and selected against, resulting in aptamers capable of switching on fluorescence upon binding across a wide spectrum of colors [106]. One of the chromophores, 3,5- difluoro-4-hydroxybenzylidene imidazolinone (DFHBI), is cell permeable, non-toxic, has a quantum yield of 0.72, possesses spectral qualities that facilitate the use of standard GFP filters, and exhibited the remarkable ability to resist photobleaching. Unfortunately, the molar brightness was much lower than that for GFP variants and exhibited fast fluorescence intensity decay [116]. Despite the drawbacks, the same authors were able to show that natural and synthetic aptamers were able to switch on the fluorescence of DFHBI [117]. The
natural SAM riboswitch aptamer was appended in a similar manner to previous studies and produced \( \sim 20 \)-fold increase in fluorescence upon addition of ligand. The SAM sensor was also able to monitor the flux of the metabolite in \textit{E. coli} over the period of several hours.

Despite the initial intrigue with Spinach, few applicable sensors have been developed \cite{118, 119}. The Jaffrey lab put forth several efforts to improve the effectiveness of Spinach, including a comprehensive mutational analysis (Spinach2 \cite{105}) and reselecting \textit{in vivo} using a FACS based method (Broccoli \cite{104}), yet applications have remained scarce. This is likely due to the spinach aptamer being structurally misunderstood for years, leading to the misinformed design and frustration with fluorescent sensors \cite{107}. It wasn’t until more structural information was available that the full binding characteristics and minimal binding motif were elucidated \cite{108} (Figure 1.7). This illustrates a critical hinderance to the development of RNA biosensors where a greater understanding of their molecular function is needed to create robust RNA elements, a step that requires considerable effort, time, and specialties that can’t often be provided by a single group.

1.7 A case for reprogramming natural riboswitches

While many examples of adaptor RNAs capable of relating the state of a receptor RNA to an actuator in a dose dependent manner exist, the metabolite recognition in these studies is almost exclusively performed by a small number of RNA aptamers \cite{120}. With \textit{in vitro} selection techniques promising the facile production of sensor domains \cite{31, 32}, it is an intriguing point that so very few have seen broad application with many failed attempts \cite{113}, especially \textit{in vivo} \cite{121}. The culprit may be a function of the selection itself, where preformed and highly structured states are produced in the aptamer, and this property differs vastly from natural riboswitch aptamers \cite{122–124}.
The need for a destabilized basal unbound state has best been explored in the context of the neomycin aptamer [125]. The aptamer was engineered into the 5’ UTR of yeast and shown to regulate the expression of a fluorescent protein, yet the vast majority of aptamers and constructs screened were completely ineffective and the strategy to attain a single synthetic riboswitch was laborious [126]. Despite the high sequence conservation between the aptamers used in this study, only two produced ligand dependent response. However, the authors quickly realized that they had a highly related population that contained active and inactive variants, which could allow for the dissection of small nuances that have a critical effect on the regulatory ability of RNA elements. NMR and UV-melting analysis was used to show that large conformational differences between the ground and ligand bound state along with a dramatic thermal stabilization accompanying ligand binding were responsible for dose dependent regulation [127]. Interestingly, the aptamers possessing the highest affinity and greatest abundance in the original selection were not able to regulate [126] and found to be highly pre-formed in the unbound state with non-canonical base pairs and a structured loop contributing to organization. [127]. A flexible unbound state is also found in riboswitches, which ensure a global organization of RNA structure by recognizing all functional groups provided by the target [128,129]. In addition to the disordered apo state, neomycin aptamers capable of regulation mimic other riboswitch trends such as conformational capture [71,130] and distal sequence elements not involved with direct ligand interaction proving critical for efficiency [131,132].

A detailed knowledge of molecular determinants responsible for riboswitch activity is invaluable and the availability of structural information for aptamers such as theophylline, tetracycline, and neomycin is likely responsible for their success in synthetic biology. However, acquiring this information is laborious and time consuming often requiring years to
decades before elements are fully understood and implementable [106, 107, 125] [131] with the lack of this information often leading to the hasty implementation of RNA elements [117] and irreproducible results [133]. While RNA synthetic biology is still in the phase of engineering basic parts [134], and with \textit{in vitro} selection proving by no means to be simple [135], there is still room for improvement upon the selection process itself. In the following chapters, I will describe my work using existing riboswitch scaffolds to incorporate robust sequence elements into selection that enables the facile characterization of resulting aptamers, aiding in the design of novel RNA parts for synthetic biology.

The work begins with our investigations into the lysine riboswitch, which is a relatively large RNA that recognizes its ligand within a five way junction supported by complex tertiary element. The RNA was found to bind potential ligands in a very rigid fashion, usually requiring the ligand to adapt to the RNAs preformed and rigid hydrogen bonding network. Therefore, a limited library was introduced with the intent of repurposing the RNA to recognize non-cognate ligands. While the strategy produces aptamers, the scaffold would often rearrange or employ sequence elements outside of the library to aid in ligand recognition. While these traits of the scaffold inhibit further applications, several iterations of selection aided in development of the technique and informed the later successful selections.

With the lysine scaffold proving to complex and unique of a structure, the focus was turned to a more pervasive and model architecture, the purine scaffold. The limited libraries in this context produce a robust aptamer, that lends itself readily to high levels of characterization. The new aptamer retains many properties of the parental aptamer, including global dynamics and the 3D architecture. Despite the resemblance to natural riboswitches and its high affinity, the aptamer resisted engineering into modular devices. While the reason is unclear, it lead to the realization that producing a single robust aptamer was not sufficient
for the production of biosensor.

The scaffolded selection had proven to provide the context for a well behaved aptamer, so additional scaffolds were introduced to produce multiple unique solutions. The scaffolds produced distinct solutions that were related in their function and recognition of the ligand. With multiple solutions obtained from parallel selections it introduces the ability for aptameric screening. This eases the constraint placed on previous device development which was limited to invoking a response out of a single solution, which is tedious if at all possible. The suite of aptamers produces transcriptional attenuators and fluorogenic sensors that function at or even beyond the previous limit of the adaptor.

These results lay the groundwork for the production of additional RNA devices for use in RNA synthetic biology. It had previously been reasoned that traditional selections would meet the demands of sensor production, but the structurally simple solutions it produces are not sufficient to transition into devices. This is evidenced by so few aptameric devices present in the literature, with \textit{in vivo} sensors still relegated to the proof of concept stage. It is hoped that the experiments and studies described herein will advance the field and aid in the future development of robust sensory devices.
Chapter 2

Investigation of the Lysine Aptamer Domain as a Host for Orthogonal Compounds

2.1 The lysine aptamer

The lysine riboswitch is a widely distributed class of riboswitch [2](Figure 1.3) and is found primarily in Firmicutes and Fusobacteria upstream of genes involved in the regulation and biosynthesis of this essential amino acid [136]. Most of these switches are involved in regulation of the diaminopimelate (DAP) pathway, responsible for cell wall biosynthesis and the production of other amino acids. The importance of the DAP pathway in cell survival and the absence of a homologous pathway in mammals has lead to its targeting in the search for novel antimicrobial agents [137–139]. One result of these efforts was the compound S-2-aminoethylcysteine (AEC), and studies to elucidate the mechanism of the antimicrobial resistance pointed to the operon of the lysC gene [140, 141]. This gene, an aspartokinase, is strongly down regulated in the presence of excess lysine [142] and AEC resistant microbes localized mutations to the untranslated region upstream of the gene that resulted in its constitutive expression. In the presence of these mutations, lysine was no longer able to attenuate the expression of this gene and it was later shown that these mutations in the operon are residues critical for the recognition of ligand by the lysine riboswitch [143, 144].

With these realizations, the investigation of binding principles for the lysine riboswitch
had already begun, but had not yet lead to a clear picture of how this RNA interacts with its cognate ligand [143, 145, 146]. Often, binding studies were done with an eye towards screening for antibiotic compounds, leaving an incomplete understanding of how the specific functional handles of the lysine metabolite are recognized or how bulky moieties at these positions are accepted by the RNA [51]. To address the discrepancies of specificity, evolve our understanding of molecular recognition, and investigate the ability of the RNA aptamer to adapt to non-cognate ligands, crystallography and fluorescent binding studies were employed.

2.2 Ligand specificity of the lysine riboswitch

The apparent equilibrium dissociation constant ($K_{D, \text{app}}$) for the compounds investigated in this study were established using a two-piece fluorescent reporter of the *B. subtilis lysC* riboswitch [147]. We constructed a synthetic RNA containing a site specific incorporation of 2-aminopurine (2AP) at the non-conserved G193 position within the five-way junction (Figure 2.1c). Solution studies had previously indicated that the chemical environment of position 193 undergoes a significant change upon ligand binding [144, 145] and it is solvent exposed in the bound state [143, 144]. This likely shows the residue transitioning from an inter-base stacked position in the ligand unbound state to a solvent exposed/unassociated state in the ligand bound state. Thus, this position is an ideal candidate for a reporter that is highly dependent upon the local chemical environment for its signal, such as 2AP whose fluorescence is high in the absence of external interaction yet can be quenched by any base stacking interaction via collisional processes [148–150].

The two-piece reporter RNA exhibits an observable increase in fluorescence upon the addition of lysine (Figure 2.1d) consistent with the nucleotide flipping into solution as established previously in the investigation of the Mg$^{2+}$ dependence of the aptamer [147]. A titra-
Figure 2.1: **Overview of the lysine riboswitch 3 dimensional structure and site specific 2AP labeling.** (a) Crystal structure of the lysine riboswitch upstream of the *T. maritime asd* gene (PDB ID: 3D0U). The three coaxial stacks are shown in different colors and consists of five paired regions supported by distal tertiary interactions. Lysine is shown in purple with the switching sequence of the 3’ side of P1 highlighted in red. (b) Overlay of the five way junction binding pocket of two independently solved crystal structures (3D0U, colored; 3DIL, gray) showing the specific hydrogen bonding pattern responsible for the recognition of lysine which includes a critical ordered potassium ion. (c) Cartoon of the constructed, two-piece reporter construct used for the binding studies described here in where G193 is replaced with the fluorescent analog 2AP. Colored nucleotides correspond to the lysine binding pocket from (b). (d) Representative titration of the fluorescent reporter illustrating the decrease in affinity and signal maxima when potassium is omitted from solution.
tion of lysine against a fixed concentration of the reporter RNA at 37 °C revealed a $K_{D,app}$ for lysine of $53 \pm 5 \mu M$, a value significantly weaker than previously reported [145, 151]. Our number is valid, however, as the previous studies were conducted under artificially high Mg$^{2+}$ concentrations, and ITC reveals that the binding constant is within $\sim$3-fold of the wild type switch under our experimental conditions. As further validation of the reporter we tested the construct in the absence of potassium, as the crystal structure revealed the RNAs dependence on this ion for recognition of the $\alpha$-carboxyl group of lysine [143]. Forcing the substitution of potassium for sodium resulted in a $>10$-fold decrease in the affinity of the reporter and failure to reach levels of fluorescence equivalent to that in the presence of potassium (Figure 2.1d). Thus potassium does not simply aid in binding of the ligand, but stabilizes a conformational shift in the RNA, an observation consistent with previous studies finding a potassium-dependent stop in primer extension assays [143]. These experiments suggest that potassium is recruited to the ligand-RNA complex and is not responsible for preformed structure, thus a non-cognate ligand must be able to recapitulate this interaction for productive binding.

We then used the validated 2AP reporter to screen key aspects of the lysine ligand to investigate the energetic contribution of each functional handle (Figure 2.2). The initial crystallographic efforts suggested two essential interactions between the $\varepsilon$-amino group and the RNA. The $\varepsilon$-amino is recognized by an electrostatic interaction with a non-bridging phosphate of G110 and a water mediated interaction with the 2'-hydroxyl of G143 (W1, Figure 2.1b). The importance of this water was tested with the compound $N$-methyl lysine (NML) which maintains the charge necessary for the G110 interaction but sterically displaces W1. The reporter system reveals an $\sim$34-fold loss in affinity, confirming the importance of the water mediated hydrogen bonding network. The electrostatic interaction is also non-
Figure 2.2: Functional handles of lysine ascribe a different function for the RNA. (a) A schematic of the various compounds tested in these studies in order to investigate the role each functional handle has on the RNA’s affinity and regulation. (b) Table of equilibrium dissociation constants as observed by the 2AP fluorescent construct. Modification to the main chain carboxylate has a modest effect on binding, leading to the idea that this region of the ligand can be built out and still maintain affinity and regulation. (c) Comparison of ligand $K_D$ and $T_{50}$ at various nucleotide concentrations. Modifications to the $\alpha$-carboxyl group retain regulatory ability, while perturbation of the $\alpha$- or $\varepsilon$-amino group drastically reduces or abolishes regulatory function. The measurement of $K_D$ via fluorescence spectroscopy was my contribution to this work.
essential for binding as tested by the compounds \(N\-\varepsilon\)-formyl lysine (FL) and \(N\-\varepsilon\)-acetyl lysine (ACL) which add bulk to the \(\varepsilon\)-amino group and abolish the formal charge yet are still able to bind with observable affinity (Figure 2.2b). The majority of the energetic loss of these compounds is likely due to their additional steric bulk, as \(N\-\varepsilon\)-iminoethyl lysine adds bulk but maintains the localized charge and binds with similar affinity. Similarly, the addition of very large moieties completely abolishes binding, indicating that the binding pocket around the \(\varepsilon\)-amino group is rigid and unaccommodating. Further modification to the ligand localized to the main-chain atoms reveal the the RNA is more plastic in this region. Removal of the \(\alpha\)-amino group all together (6-aminocaproic acid; 6ACA) is still able to bind the RNA at a reduced affinity, and modifications to the \(\alpha\)-carboxyl group (Figure 2.2 red) are well tolerated with only very moderate losses in affinity relative to lysine. These bulky \(\alpha\)-carboxyl groups likely displace the potassium ion otherwise necessary for the productive binding of lysine, though they are not capable of completely recapitulating the energetic loss. While the RNA is more accepting of changes to the \(\alpha\)-carboxyl position, there is a limit as a peptide chain greater than one additional glycine is not permitted, as in the case of Lys-Gly-Gly (KG2). These data taken together with additional structural probing reveals that the binding site is rigid and adopts a similar, if not identical, conformational state regardless of the ligand.

2.3 The crystal structure of lysine analogs bound to the lysine riboswitch

With previous lysine analog crystallographic efforts focusing on the \(\varepsilon\)-amino group and side chain derivatives and our data suggesting that main-chain alterations are fairly well tolerated within the lysine riboswitch binding pocket despite its overall rigidity, further crystallographic studies were needed to reveal how this RNA is able to accommodate chemical diversity at these positions. To do so, we solved the crystal structure of the *T. maritima*
lysC lysine riboswitch bound to the Lys-Gly dipeptide (KG) and 6-aminocaproic acid. These two analogs interrogate either the $\alpha$-carboxyl or $\alpha$-amino groups independently allowing us to observe how they orient themselves to best fit the fixed hydrogen bonding network. The crystals diffracted to $\sim$3.0 Å resolution and the models were generated via molecular replacement with the ligand free structure (PDB ID: 3D0X) having residues responsible for the hydrogen bonding network removed. Superimposition of the refined models revealed that the RNA adopts an identical fold regardless of ligand with the average deviance of RNA atoms between lysine bound and analog bound structures being less than the coordinate error of 0.38 Å (Figure 2.3).

The KG-RNA complex shows clear ligand density in the $F_O$-$F_C$ difference map, with additional density extending 3.6 Å away from the central cavity when compared to the lysine structure allowing for the placement of the glycine functional group. The terminal carboxyl group of the dipeptide now forms an additional hydrogen bond with the 2'-hydroxyl of G143 at the base of P4. This new interaction replaces an ordered solvent molecule involved in the coordination of the K$^+$ ion in the lysine bound structure. This, along with the water critical in the recognition of the $\varepsilon$-amino group, suggests that water-mediated contacts are important in multiple facets of the binding interface.

The absence of the hydrogen bonding network between the $\alpha$-amino group and G143 presented by the 6ACA ligand appears to be compensated for by water and a shift of the carboxyl group in the structure. Two distinct peaks were present in the $F_O$-$F_C$ difference map near the absent $\alpha$-amino group of lysine which were modeled as ordered solvent. These water molecules form a new hydrogen bonding network along with the shifted carboxyl group of 6ACA that recognizes the sugar edge of G143. These structures show no evidence for the adaptation of the RNA binding pocket, thus it is the responsibility of the ligand to form
Table 2.1:

<table>
<thead>
<tr>
<th>Ligand in complex</th>
<th>Lysine-glycine (KG)</th>
<th>6-aminocaproic acid (6ACA)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Collection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>$P_{32}$</td>
<td>$P_{32}$</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a, b, c$ (Å)</td>
<td>119.9,119.9,58.7</td>
<td>119.9,119.9,58.7</td>
</tr>
<tr>
<td>$a, \beta, \gamma$ (°)</td>
<td>90,90,120</td>
<td>90,90,120</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>1.5418</td>
<td>1.5418</td>
</tr>
<tr>
<td>Resolution (Å)*</td>
<td>19.6-3.0 (3.1-3.0)</td>
<td>59.8-3.0 (3.1-3.0)</td>
</tr>
<tr>
<td>Rsym $^\circ$</td>
<td>0.113 (0.321)</td>
<td>0.098 (0.338)</td>
</tr>
<tr>
<td>$I/\sigma I$</td>
<td>5.1 (1.9)</td>
<td>5.4 (1.7)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>97.8 (96.5)</td>
<td>97.7 (99.9)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>1.92 (1.93)</td>
<td>1.59 (1.58)</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)*</td>
<td>19.6-3.0 (3.2-3.0)</td>
<td>26.7-3.0 (3.1-3.0)</td>
</tr>
<tr>
<td>No. Reflections</td>
<td>18687</td>
<td>17611</td>
</tr>
<tr>
<td>Rwork / Rfree $^\circ$</td>
<td>19.5/23.4</td>
<td>20.7/22.3</td>
</tr>
<tr>
<td>No. Atoms</td>
<td>3503</td>
<td>3490</td>
</tr>
<tr>
<td>RNA</td>
<td>3479</td>
<td>3479</td>
</tr>
<tr>
<td>Ligand</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>Water</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td><strong>Atomic Displacement Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>48.1</td>
<td>55.5</td>
</tr>
<tr>
<td>RNA</td>
<td>48.1</td>
<td>46.7</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>53.8</td>
<td>36.7</td>
</tr>
<tr>
<td>Water</td>
<td>31.8</td>
<td>50</td>
</tr>
<tr>
<td>r.m.s deviations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.68</td>
<td>0.914</td>
</tr>
<tr>
<td>Maximum likelihood coordinate error (Å)</td>
<td>0.38</td>
<td>0.34</td>
</tr>
</tbody>
</table>

a Values for the highest resolution shell are shown in parentheses.
b Rmerge = $\sum |I_i - \langle I \rangle | / \sum |I_i|$.
c Refinement was against all data within the stated resolution range, with a random 10% omitted for use in calculation of Rfree.
d Rwork = $\sum |F_o| - |F_c|/\sum |F_o|$ for the working set. Rfree is the same for the test set.
Supplementary Figure 5. Simulated annealing omit maps (green) contoured at the 3.0 level; only the density surrounding the ligand is shown. The maps were calculated for (a) KG and (b) 6ACA compounds using an unliganded model based on the free state of the structure (PDB 3D0U). A distinct peak was observed in the 6ACA map suggested that a water replaces the missing main chain amine created a new set of hydrogen-bonding interactions between G114 and the main chain moieties. Density for the W1 water was also present beyond the side chain indicating its presence in these complexes. The figure shows the refined structural model of each analog overlaid with the omit map for clarity.

Figure 2.3: Lysine riboswitch scaffold reinforces a strict network of hydrogen bonding interactions. Structure of (a) KG (PDB ID: 4ERL) and (b) 6ACA (PDB ID: 4ERJ) molecules bound to the *T. maritima* lysine riboswitch overlaid with the bound lysine structure (light gray) from [143]. The RNA adopts an identical conformation independent of the bound molecule and supports the chemical probing studies. Simulated annealing omit maps (c) and (d) are shown for each compound.
novel hydrogen bonding interactions, recruit solvent, and avoid steric clashes to adapt to the rigid RNA that presents an established network of functional groups for recognition.

2.4 Reprogramming the lysine aptamer domain

The idea of screening for novel effector molecules against the lysine scaffold seems difficult, with the results of the above study providing no clear path for the implementation of any novel effector molecules. Given this, the RNA is now well understood in the community with many orthogonal studies into its recognition, structure, and dynamics down to single molecule resolution [152]. If the lysine riboswitch could be evolved to recognize a novel effector, this wealth of information and level of understanding should aid in the timely characterization of the RNA. Therefore, to begin the investigation into novel RNA control elements, I sought to reprogram the *B. subtilis* *lysC* lysine riboswitch aptamer to specifically recognize non-cognate ligands with a limited library.

The crystal structures were used as a basis for the selection of residues that would be randomized to create the library. While the structure is based on *T. maritima*, we chose the mesophilic *B. subtilis* sequence for the basis of our study as it likely possess greater conformational dynamics than the thermophilic sequence used for crystallography. The use of a highly thermostable RNA such as that from a thermophile might preclude downstream applications, as it has been shown that highly preformed states in the absence of ligand do not translate well into sensory RNA devices [153]. An ~8 Å shell of nucleotides that surrounds the cognate ligand where selected on the basis that this level of randomization grossly preserves secondary and tertiary structure, yet provides enough diversity to allow the inflexible RNA to host a novel compound. The resulting library consisted of twenty-four variable positions (Figure 2.4) corresponding to $2.8 \times 10^{14}$ possible nucleotide combinations.
Figure 2.4: **Evolving the lysine riboswitch** (a) Structure of the lysine riboswitch (PDB ID: 3D0U) with an 8 Å shell of residues highlighted (cyan) surrounding the ligand (orange) resulting in twenty-four varied positions. (b) Ten rounds of selection were carried out according to the materials and methods (Section 2.7) with increased stringencies noted (arrows) and resulting in nearly 40% competitive elution for the 5htp selection. (c) Projection of conservation onto the structure from (a) showing possible covariation at the base of P3 and P4 (orange) and highly conserved nucleotides around the proposed binding pocket (green). (d) A **WebLogo** illustrating the sequence conservation and orientation of the varied regions. Possible covariation is shown by the orange bars.
and roughly 14% total sequence randomization.

The initial targets were guanosine monophosphate (GMP), 4-amino-L-phenylalanine (4AF), and 5-hydroxy-L-tryptophan (5htp). These targets were chosen due to a combination of their ability to bind RNA in previous studies [154, 155], their proposed in vivo functionality [156], and industrial relevance [157]. The ligands were coupled to a solid sugar support and ten rounds of selection were carried out (Figure 2.4b) with increasing wash stringencies and counter-selection against adenosine monophosphate (AMP), L-phenylalanine, or L-tryptophan, respectively. The selection for 4AF failed and will be addressed later, the GMP and 5htp selections were successful in producing RNA capable of binding to the column and competitively eluting. During the selection, the counter selections resulted in a marked decrease in the amount of RNA retained and competitively eluted from the column, however the retention rebounded after a single round and showed even higher levels of retention and elution under the same conditions. Thus suggesting that the pool was now more selective for the respective ligand.

Individual variants were cloned and sequenced after ten rounds of selection revealing a single class with a high level of sequence conservation for the 5htp variants and essentially a single sequence for GMP (Section 2.7). In addition to the essentially fixed nucleotides for the 5htp selection, projection of sequence identities onto the crystal structure and comparison of the nucleotide degeneracy reveals two areas of possible covariation, one each at the base of P3 and P4 (Figure 2.4c orange nucleotides). This, along with maintainence of the sequence elements necessary for peripheral structure, would suggest the architecture was preserved and that the helices were rebuilt to retain the orientation necessary for tertiary contacts.

A single, abundant sequence that matched the consensus profile for each selection was then chosen for validation and characterization. Selective 2'-hydroxyl acylation analyzed by
Figure 2.5: SHAPE and ITC analysis reveals specificity, affinity, and a necessary primer (a) SHAPE probing reveals ligand dependent changes in J2/3 and J3/4. Sequence 12 of the 5htp selection would appear to have a slight preference for 5htp (5) over L-tryptophan (W) and serotonin (S) while showing no affinity for lysine (K). (b) Sequence 21 of the GMP selection again shows modification upon addition of ligand in the J2/3 and J3/4 region of the RNA. This sequence shows no specificity for the phosphorylation levels at the 5’ position, yet would appear to be highly dependent upon the Watson-Crick face for productive binding based on its indifference for AMP and inosine monophosphate (IMP). (c) ITC of sequence 5htp-12 shows only modest affinity for the ligand and removal of the 3’ primer (a region critical for use in biosensors) used for RT-PCR shows no affinity for 5htp (d). (e) The GMP-21 aptamer was able to successfully bind GMP in the absence of primer sequences, though the affinity decreases by roughly five fold.
primer extension (SHAPE [158]) was used to investigate the overall structure of the two sequences and screen for selectivity (Figure 2.5a). The technique relies on the modification of the ribose hydroxyl, the reactivity of which is highly dependent upon local nucleotide flexibility [159]. If the 3'-phosphodiester is conformationally constrained, e.g. by base pairing interaction, the formation of the deprotonated, nucleophilic oxyanion form of the 2'-hydroxyl group is disfavored, leading to a low degree of modification by the N-methylisatoic anhydride electrophile. If, however, the nucleotide is conformationally dynamic, e.g. an unpaired region, it is capable of accessing a more reactive state, leading to greater formation of the 2'-O-adduct. This 2'-O modification is obstructive to a reverse transcriptase and can thus be used to monitor a nucleotides reactivity in a primer extension assay at single nucleotide resolution.

The SHAPE technique revealed ligand dependent modifications in the varied, joining regions of the RNA suggesting that junction was responsible for housing the ligand. The 5htp sequence showed a moderate preference for 5htp over tryptophan and serotonin as observed by the weaker modifications for the two orthogonal ligands. The GMP sequence, however, showed no preference for the phosphorylation state of the nucleotide, but strong discrimination along the Watson-Crick face as it does not recognize AMP or inosine monophosphate (IMP). This is consistent with previous selections against GMP and GTP where the poly-anionic RNA has difficulty specifically recognizing the phosphates [154, 155]. As expected, neither sequence retains the architecture’s cognate binding affinity for lysine. Comparison of the global reactivity pattern between these RNA and the previously published lysC aptamer [143, 144, 160] show some similarities, particularly in the region preceding J2/3, yet there are a number of differences between the RNAs, leading to the uncertainty that the overall scaffold was maintained.
An isothermal titration calorimeter (ITC) was used to establish the binding affinities of the selected RNA. Briefly, ITC is a thermodynamic technique that quantitatively monitors the enthalpy upon the mixing of two solutions. In these studies, one solution contains the RNA and the other the small metabolite. The technique is able to inform upon the binding affinity, change in enthalpy, and stoichiometry under equilibrium conditions in solution and free of labels. The selected RNAs are able to bind to their respective target ligands as established by ITC, albeit with a rather modest affinities. The 5htp aptamer that contains the primer regions as present during the selection shows an affinity of $130 \pm 30 \mu\text{M}$. Unfortunately, when these primer regions are removed and replaced with a single, elongated P1, the aptamer loses all affinity for the metabolite (Figure 2.5c,d). Various other candidate sequences from the selection were investigated for binding, even an additional class revealed from round 6 sequencing, but none were able to bind in the absence of the priming region. In order for these RNA parts to be applicable the P1 stem must be adapted into a communication module, which the 5htp sequence appears unable to do. The GMP sequence, however, is able to bind in the absence of its priming region (Figure 2.5e) with a $K_D$ of $60 \pm 10 \mu\text{M}$. This value is slightly weaker than the sequence that possesses the primer regions ($K_D = \sim15 \mu\text{M}$), which is >500 fold higher than aptamers derived from a near random pool [154].

2.5 Optimization of the lysine library sequence and priming regions

The selection described above was able to produce competent binding sequences, but not in a manner suitable for downstream applications. The culprit for the 5htp selection would appear to be an artifact of the primer regions interacting with the RNA in a way that promotes efficient binding to the ligand. This is, in fact, a common phenomena within the selection field, occurring in a surprising number of selections and various ideas to solve
Figure 2.6: Expansion and redesign of the lysine architecture does not lead to a successful selection (a) Projection of the expanded selection onto the crystal structure as in Figure 2.4a. (b) Detailed view of sequences randomized for the *B. Subtilis* secondary structure. Cyan residues show those randomized for the initial selection and those in orange for this iteration. The gray box shows the appended P1 stem designed to limit its interaction with the rest of the sequence and serves as the priming sites for amplification. (c) Selection against all three compounds progressed as expected for seven rounds. However, successive rounds failed to competitively elute RNA from the respective columns. (d) Sample sequencing analysis reveals no convergence for the 5htp selection making interpretation of the sequences for further characterization impossible.
the problem, which usually involves hiding the primer regions in base pairs distal from the library [161, 162]. Therefore, I employed a combination stem-primer region adapted from Illangasekare et al. (2010) [163] (Figure 2.6). Since the priming regions are complimentary and extend away from the constant region and library, this design allows for the productive amplification of the pool while lowering the risk that the fixed region will interact with selected RNA sequence in a manner that is necessary for ligand binding.

An additional problem lies in the failure of the 4af selection, which could not be enriched at any stage of the experiment. The 4af aptamer has proven difficult for RNA to recognize before, requiring a large amount of structure and complexity [156]. Given that higher affinity aptamers are larger, possess more intricate structural motifs, and are more informationally complex [155], we had hoped that selections within a predefined scaffold would decrease the amount of information content needed to elicit productive binding. Since this appears not to be the case, I sought to expand the library from twenty-four nucleotides to thirty-eight, bringing the total sequence variation to 22% and a total sequence diversity to $7.6 \times 10^{22}$ possibilities. While this level of randomization moves beyond what can be sampled by current synthesis techniques, it more closely resembles the percentage previous studies used that were successful in altering the binding specificity of RNA [164, 165].

The selection yielded observable competitive elution of RNA at early rounds for all three selections (Figure 2.6c) and showed a complete resistance to the challenge of counter-selection at rounds 6 and 7. The pools were cloned and sequenced after round 7 to reveal no clear consensus and no redundant sequences (Figure 2.6d). The individual sequences are so diverse, in fact, that they cannot be clustered by UCLUST [166] at 90% identity, which is unusual given that the sequences are by design 78% identical in the initial pool. This suggests that the pool had not settled on a clear solution, and with no clear starting point
for characterization, I decided to carry out additional rounds of selection. However, despite multiple attempts that included starting at even earlier rounds, the levels of RNA competitively eluted could never be recovered (Figure 2.6c). The reason for the selections failing is not clear, however it is likely that artifact sequences incapable of binding to the column, yet highly proficient at amplifying relative to binding competent sequences dominated the pool.

While the priming regions may not have been the issue with the previous selection as there was nothing to implicate them in binding, they did not produce a robust solution as before and were not a variable we were willing to experiment with further. Therefore I designed a hybrid stem-primer region that incorporated aspects form both strategies: known robust primer regions from iteration one and a long, destabilized stem from the second to create a platform in which the constant priming regions are kept distant from the selected RNA and amplifies efficiently (Figure 2.7a).

The *B. subtilis* lysine riboswitch was selected against 5htp in both the twenty-four and thirty-eight nucleotide library iterations containing the new amplification platform. Additionally, a control selection consisting of twenty-four nucleotide library in the newly developed platform was conducted in parallel. The control sequence included a small, structured stem loop to provide some level of architecture, but at a level much less than provided by the peripheral sequence of the lysine architecture. Interestingly, the selections progressed more rapidly than before, having observable elution by round 2 for both the lysine libraries, while the control sequence was not able to elute observable RNA before round 4 (Figure 2.7b). Counterselection against L-tryptophan was performed at rounds 5 and 6 and resulted in only a modest loss in the levels of RNA eluted by 5htp. The pools were sequenced at round 7 to reveal yet another artifact that arose during the selection, large swaths of sequence deletion at very specific junctions that kill the architecture of the RNA. Two islands of sequence on
Figure 2.7: **Primer site optimization reveals a substantial problem with preservation of the architecture.** (a) Elongation of the P1 stem optimized for length and low thermal stability. Robust and previously established primer sites were used in this iteration, with the site of annealing for the reverse transcription primer outlined in blue. (b) Elution profile of the two lysine selections with a non-scaffolded control selection performed in parallel. Counter selections were performed at rounds 5 and 6 as indicated and described in the methods section 2.7. (c) Sequencing analysis of the lysine scaffolds. Both selections regardless of library size were subject to large regions of deletion that occurs with remarkable precision. (d) This large scale deletion occurred various times independently as suggested by the diversity that is still present in variable regions of the scaffold, yet convergence in J4/5 and other regions suggests that this pool has decided on a solution and is likely functional.
the 3' side of P2 were deleted (Figure 2.7c), making formation of the L2-L3 and P2-L4 interactions highly unlikely. The specificity and accuracy with which this happens is remarkable, with the deletion repeatedly occurring immediately after a stretch of adenosine residues, and preserving a small island of sequence in P2 before deleting an additional 7 residues. These sequences are likely functional, as sequence convergence is seen in region J4/5 and J3/4 is invariant, yet this truncation must have occurred on multiple, separate occasions as other regions within the junction show high sequence heterogeneity (Figure 2.7d). Given the destruction of the intended architecture, and no clear sequence conservation in the control selection, the interpretation of this experiment is again complicated.

2.6 Discussion

Reprogramming the lysine scaffold proved to be difficult, likely due the the size of the RNA, 176 nucleotides in these experiments without priming regions, and complexity of the scaffold, which contains multiple distal tertiary interactions and long base-paired regions. These two factors influence the ability of the RNA to replicate, with the main pressure on the reverse transcription process that likely has difficulty moving through such highly structured and thermo-stable regions of RNA [167]. Therefore, the external pressure on the selection is immense and sequences capable of more efficient replication, such as those that contain mutation inducing global misfolding or reorganization, dominate the final pools. Given the caveats presented by the lysine architecture and it proving to be a fragile scaffold in selection, it was abandoned. That is not to say the the selective process in and of itself was not successful in its own way, the result was simply pleiotropic and what we obtained was, in fact, what we selected for: RNA capable of surviving affinity chromatography and amplifying robustly, with no pressure on scaffold maintenance. Absent of the likely issue of
the reverse transcription, I have successfully developed a platform and procedure capable of producing robust RNA parts as I will describe in subsequent chapters of this thesis.

2.7 Materials and methods

Preparation of RNA for fluorescent studies
A 2AP-modified oligonucleotide with the sequence GGAGUCUUUCUUGGAG-2AP-GCUAU CUCUCC was chemically synthesized by Dharmaco, Inc. All other RNAs in this study were synthesized using T7 RNAP and purified according to previously established protocols. Purified RNAs were stored in 5 mM Tris-HCl (pH 8.0) and 0.5 mM EDTA buffer and stored at -20 °C prior to use. To assure proper annealing of the two-piece RNA construct, we engineered the sequence derived from the B. subtilis lysC aptamer using the following 5' and 3' primers to PCR amplify from a cloned plasmid containing the entire promoter and 5' leader segments of this gene (5'-TAATACGACTACTATAGGCGTGAGCAGACTCTTTTTTTGGA GAGATAGAGGTGCGAAC and 3'-TAAACGACCCGCCACGTAACCTTATTTACATTCCG ACAGTTCTTTTCTGAGG) to produce an unlabeled RNA with the sequence GGAGAGAGAGGUGCGAACUUCAAGAGUAUGCCUUUGGAGAAAGGAAUGGAUUCUGUGAAAA AGGCCGAAAGGGGAGCGGUCGCGGCAAGCAAAAAACCCCAUCCGGUAAUUAUUGC UGGCCGUGCAUUGAAUAAUGUAGGCGUGUCAAGAAAGACUCC. The two-piece RNA used in the titration experiments was constructed by annealing in a buffer containing 50 mM Na-Hepes (pH 8.0) and 100 mM NaCl. The unlabeled RNA was added to a final concentration of 3 μM to a 1 μM stock of 2AP-labeled RNA and heated in a PCR block to 85 °C for 2 min then cooling at a rate of 0.1 °C/s with a 2-min hold at 10° intervals down to 4 °C. Annealed products were analyzed by native polyacrylamide gel electrophoresis to ensure that the quality of the annealing reaction was optimal. Under these conditions, the labeled oligonucleotide was verified to incorporate with high efficiency into a single species corresponding to the appropriate size product.

Fluorescence spectroscopy
All ligand titrations were carried out in 50 mM Tris-HCl (pH 8.0) and 100 mM NaCl and 5 mM MgCl2 and 10 mM KCl with a 100-nM RNA unless indicated otherwise. Changes in the fluorescence of the 2AP reporter as a function of ligand concentration were measured using a Tecan 96-well plate reader for all room temperature experiments and in a Chirascan fluorescence spectrometer for 37 °C experiments. Data were fit to the equation

\[ \Delta F = F_{\text{min}} + (F_{\text{max}} - F_{\text{min}}) \times \left( \frac{[L]^n}{[L]^n + K_D} \right) \]  

(2.1)

where \( \Delta F \) is the measured change in fluorescence, \([L]\) is the ligand concentration, \(F_{\text{min}}\) is the initial fluorescence, and \(F_{\text{max}} - F_{\text{min}}\) represents the total magnitude of the change in measured fluorescence change for the titration.

Crystallographic data collection and processing
Crystallization of the 6ACA and KG ligands was performed with a variant of the T. maritima asd lysine riboswitch that was previously described (PDB ID: 3D0U) [144]. The RNA
was transcribed, purified, and refolded using previously described protocols [168] and was concentrated to 400 μM in 10 mM Na-Hepes (pH 7.0) and 5 mM MgCl$_2$. RNA and ligand were then mixed at a 10:1 ratio to give a final ligand concentration of 5 mM and allowed to incubate for 30 min prior to setting up crystallization drops. Mother liquor [10 mM Na-Hepes (pH 7.0), 2 M Li$_2$SO$_4$, and 5 mM MgCl$_2$] was mixed with the RNA-ligand solution in a 1:1 volumetric ratio, and crystallization was performed by hanging-drop vapor diffusion. Diffraction data were collected on a copper rotating anode source (Rigaku RU200 and Rigaku RU2HR) with a Rigaku MSC IV++ area detector at 100K. Data were scaled and averaged using D*TREK [169] as part of the CrystalClear software package (Rigaku MSC). Structures of the 6ACA- and KG-bound complexes were solved by molecular replacement with the unliganded lysine aptamer (PDB ID: 3D0X) using the PHENIX software suite [170]. Model building was performed with Coot [171] and additional refinement was performed with the routines implemented in PHENIX. All figures of structural models were prepared using PyMOL [172].

**Lysine library constructs**

**First iteration of selection**

**T7 promoter to append:** GCGCGCAATTCTAATACGACTCAGTATAGGAGGACTTCCGGTCAAAAGATAG

**RT primer:** GAACCGGACCGAAGCCCG

**Forward Library template**

| 5htp Sequence | GGAGGACTTCGGGTCAAAAGATAGGTCTGTACTGACAGTTCCAGGTTGAGAAAGATGAGTCTGGAAAAGGGCAAGGCTCGGCGTCCGTGGCAATAAAGGCTCATTATCGGGTTGCTTTGGCCCTTTGGGCCCTTTTCTTCGGTCCGGTTC |
| Reverse Library template | GAACCGGACCGAAGCCCGATTTGACAACGAGATTAGNNNNNAAGAAAATGATTTCATTGNNNCCCTACATTATTCAATGCACGGNNNGCAATAATACGGATCGGGTTTTATT |

**Second Iteration Library with 38 nucleotide library**
Column Synthesis
As described in Section 3.6

Selection Protocol
The selection protocol for the third iteration is as follows.

R1:
350 μL of acetylated sepharose was equilibrated in selection buffer (10 mM HEPES pH 7.0; 250 mM NaCl; 50 mM KCl; 10 mM MgCl₂; 0.1 mg/mL tRNA) and 1 nmol of library RNA in 350 μL of selection buffer was incubated at RT for 30 min. The column was then drained and washed once with 350 μL of selection buffer. The flow through (750 μL total) was added to pre-equilibrated 5htp derivatized sepharose 4B column and incubated for 45 minutes. The column was then drained and washed three times with selection buffer before elution with 10mM 5htp in selection buffer (two 1 hour incubations in 350 μL; total 700 μL eluted volume). The eluted fractions were then concentrated to 50 μL in a 0.5 mL Ultracel 10kMWCO filter (Millipore) and ethanol precipitated in 0.3 M sodium acetate (pH 5.0), 5 μg glycogen, and brought to a final concentration of 75% ethanol before storage at -70 °C for 30 minutes. The solution was then pelleted at 13000 x g at 4 °C, decanted, and dried under vacuum. The dried pellet was then reconstituted with 0.7 mM each dNTP, 7 μM RT-PCR primer, and to a total volume of 14 μL before heating to 65 °C for 5 minutes and incubation on ice for
10 minutes. The solution was then brought up to 1x SuperScript III FS buffer conditions with 5 mM DTT and 200 units SuperScript III (Life Technologies) in a total volume of 20 μL before a 15 minute extension at 54 °C. The entire 20 μL reverse transcription solution was then PCR amplified in a total volume of 500 μL using Taq polymerase conditions. The amplified pool was then transcribed by adding 100 μL of the PCR reaction to a 1 mL transcription reaction containing 40 mM Tris-HCl, pH 8.0, 25 mM DTT, 2 mM spermidine, 0.01% Triton X-100, 4 mM each rNTP pH 8.0, 0.08 units inorganic phosphatase (Sigma-Aldrich, lyophilized powder), and 0.25 mg/mL T7 RNA polymerase and incubated at 37 °C for 2 hours. A 100 μL transcription reaction for 32P labeled RNA was performed under similar condition with the exception that the rNTPs were lowered to 2 mM for UTP, CTP, and GTP while ATP was reduced to 200 μM and 100 μCi of ATP-α-32P was supplemented. Transcription samples were gel purified as described in the library construction section with the gel loading conditions scaled accordingly. Subsequent washing procedure for all selections (I, II, III)

<table>
<thead>
<tr>
<th>Round</th>
<th>[RNA] pmol</th>
<th>Washes</th>
<th>Counter Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>2</td>
<td>400</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>400</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

**Isothermal Titration Calorimetry**

All RNAs tested were exchanged into the selection buffer (10 mM HEPES pH 7.0; 250 mM NaCl; 50 mM KCl; 10 mM MgCl₂) and washed three times in 10k MWCO filter (EMD Millipore). The ligand was brought up from a dry solid directly into the binding buffer and concentration established on a NanoDrop 2000 (Thermo Scientific; ε²₇₅ = 8000 mol⁻¹ cm⁻¹ for 5-hydroxy indol moiety). The RNA was diluted to between 50-100 μM and the ligand was diluted to 10 fold that of the RNA. Titrations were performed at 25 °C using an iTC-200 micro calorimeter (Microcal, Inc.; GE). Data was analyzed and fitting was performed with the Origin 5.0 software suite (Origin Laboratories).

**Chemical Probing**

RNA was prepared as described previously [168]. Structure cassettes flanking the 5’ and 3’ ends of the RNA were added to facilitate reverse transcription and NMIA modification was performed using the established protocols [159] at 25 °C. RNA was probed at 100 nM in 100 mM Na-HEPES [pH 8.0], 100 mM NaCl, and 6 mM MgCl₂. Ligand concentration was 500 μM where indicated. Gel images were analyzed by SAFA [173] and ImageJ (NIH).
The selections involving the lysine riboswitch architecture as previously described in this body of work did not meet our expectations likely due to the size, complexity, and structural constraints placed on the ligand binding site by peripheral structure of the RNA. The lysine scaffold seems to be a unique, rigid solution that was settled upon solely for the purpose of binding lysine. This idea is supported by its scarce representation in genomes and low success rate in analog binding studies, implying its limitations for repurposing. In contrast, there are two folds that are highly pervasive throughout biology: the three-way junction supported by a distal tertiary interaction and the H-type pseudoknot. These folds play host to a wide diversity of functions, much like the various folds found in the protein world [174]. However, unlike protein classification, the prediction of an RNA fold is often not implicit in a function, making the use of an ubiquitous RNA fold as a host for alternative functions attractive.

Several topologies of the RNA pseudoknot exist, with the best understood being the H-type pseudoknot. The fold orchestrates many diverse functions that include the formation of catalytic cores in self-splicing RNA [175] and telomerase [176], small metabolite binding [177], and alteration of gene expression via frame-shifts [178, 179]. In the simplest of cases, the
structure consists of a stem loop, where the unpaired residues of the loop form intramolecular pairs with bases present outside of the stem region forming an intertwined second stem loop. These two stems may then coaxially stack to form a single, stable, and elongated structure with one continuous strand of the helix and one that has been bifurcated. A consequence of this structure is that the terminal ends of the motif are no longer located in proximal space, often protruding from the fold at opposite ends. This makes interfacing with conventional adaptor and actuator RNA elements difficult as they are often spliced into a single stem of the RNA juxtaposed to the binding pocket. Further complicating the issue is that the interaction of the loop with residues outside of the stem can host the function of the RNA, making the engineering of H-type pseudoknots less attractive.

On the other hand, the three-way junction often requires the 3' and 5' ends of the RNA to be proximal to one another in a stem and has proven to be a modular, engineerable domain. This scaffold is noted for its ability to host a variety of activities, including small-molecule binding and catalysis, and is embedded as an essential character in ribonucleoprotein complexes such as the ribosome, RNase P, and the signal recognition particle. Structurally, it is described as a parallel Y shape, where the base (P1) forms a coaxial stack on one of the arms with the second arm establishing tertiary contacts distal from the junction that serve to orient it in parallel with the stacked helixes [180]. This distal tertiary interaction often dictates the final shape of the RNA, acting as a hinge for the formation of the junction region. The hinge mechanism allows for the state of the junction to effectively communicate with distal elements and vice versa, a feature relevant to the biological function of many of these RNAs [181–184].

A chief representative and model system for the three-way junction is the purine riboswitch [185]. Global organization of the purine riboswitch family is governed by interac-
tions between L2 and L3 serving to pack P2 against P3 (Figure 3.1). Under near physiological magnesium ion concentrations (0.5 - 1 mM), this interaction provides \( \sim 4.1 \text{ kcal/mol} \) to the ligand binding energy in the \( xpt/pbuX \) guanine riboswitch [183] and at least 2.9 kcal/mol in the \( pbuE \) adenine riboswitch [186]. At the core of this interaction within all members of the purine clan is the formation of two invariant G-C Watson-Crick pairs between L2 and L3 (“pseudoknot”, Figure 3.1a) and formation of this tertiary element is critical for high affinity recognition of ligands [187]. Since many riboswitches act co-transcriptionally, they contain robust folding elements such that they are able to fold dependably and rapidly as this is critical for the RNA’s function. The distal tertiary interaction of L2-L3 in purine riboswitches is one such element, aiding in the pre-organization and formation of a competent binding pocket and allowing for the riboswitch to interrogate the proximal environment for ligand on a time frame relevant to transcriptional speed [188]. In the \( B. \ subtilis \ xpt-pbuX \) riboswitch, the interaction is extremely stable and capable of forming in the absence of either magnesium ions or ligand [189]. Strikingly, the L2-L3 interaction is observed to form outside the context of the three-way junction or helix 1, further emphasizing its intrinsic stability [130].

The binding pocket of purine riboswitches presents a fairly consistent framework for the recognition of ligands. At the junction of the helixes is a set of nearly invariant nucleotides (red Figure 3.1a) that serves to organize three critical bases that form a base quartet with the ligand (nucleotides 47, 51, and 74; Figure 3.1c). Discrimination between the recognized purine lies primarily in the identity of position 74, which forms a Watson-Crick pair with the ligand, being a cytosine for guanine recognition and uridine for adenine recognition. Position 51 is responsible for distinguishing between purine bases, where U51 recognizes the sugar edge via N3/N9, and 2’-deoxyguanosine, where C51 dissociates from residue 47 and shifts slightly to accommodate the ribose sugar (Section 1.3). The flexibility of the
Figure 3.1: **The purine riboswitch consensus and binding architecture** (a) A simplified consensus sequence projected upon the secondary structure of the purine family of riboswitches, derived from the Rfam database [80]. Red nucleotides indicate at least 90% conservation. Colored backgrounds represent paired regions corresponding to the overall three-dimensional structure. (b) Global architecture of a representative member of the purine family (*B. subtiles xpt-pbuX* PDB ID: 1U8D) bound to hypoxanthine (HX). (c) Hydrogen bonding pattern of the two residues that confer ligand specificity (*xpt*:hypoxanthine U51 and C74) and the critical organizing nucleotide U47.
purine binding pocket lies in contrast to that of lysine’s, where the pocket is rigid and non-cognate ligands capable of binding are a trivial solution that are required to adapt to the pre-established hydrogen bonding network. The helical junction of purine riboswitches can adapt to alternative ligands with little mutational input and can easily accommodate bulky functional groups like a ribose sugar [74, 76]. These initial observations of aptitude along with preliminary studies into its binding plasticity [36, 75, 190–195] lead us to choose the purine scaffold as a potentially facile and robust selection platform for novel RNA parts.

### 3.2 Library design and selection

To begin the investigation into the purine scaffold’s ability to host orthogonal compounds, we again employed a limited library localized to the junction and cognate binding pocket. A twenty-three nucleotide shell of residues surrounding the cognate ligand of the *B. subtilis xpt-pbuX* aptamer domain was randomized, resulting in approximately 33% of total sequence variation (Figure 3.2). Our library was designed to balance sequence diversity and preservation of structure, while the total possible solutions represented by this degree of randomization (∼7 x 10^{13} combinations) may be encapsulated by standard solid phase DNA synthesis methods, thus aiding in analysis and the sampling of all possible solutions. Again, the target for selection was the precursor of serotonin, the amino acid L-5-hydroxytryptophan (5htp Figure 2.4), which is the rate limiting metabolite in the synthesis of serotonin, a metabolite important in the regulation of human behavior and cognitive state [196]. 5htp was immobilized to a solid medium and 7 rounds of selection were carried out that included counterseleions against L-tryptophan and increasingly stringent washing procedures (Section 3.6). This resulted in observable enrichment by round four of the selection and a peak eluted fraction of 50% at round six.
Figure 3.2: An overview of the purine library selection. (a) Sequence and structural prediction of the *B. subtilis xpt* purine riboswitch used for the selection. The amplification cassette (maroon) and nucleotides selected for randomization (green) are highlighted. (b) Projection of the nucleotides selected for variation (green) onto the crystal structure of the aptamer (purple: the hypoxanthine ligand). (c) Chemical structure of 5htp with site of attachment to the solid phase medium shown. (d) Elution profile of the purine selection against 5htp. Counterselections were performed at rounds 5 and 6 against L-tryptophan and wash stringencies were incrementally increased after round 4 (Section 3.6).
The final pool of selection was highly convergent as revealed by preliminary sequencing. Therefore, we concluded the selection was complete and had settled upon its solutions to the partitioning pressure. In order to further interrogate the composition of the selection, every round was deep sequenced. This resulted in ~6 million reads with roughly equal coverage of every round and was processed according to the Materials and Methods (Appendix A). The majority (>80%) of round 7 sequences cluster into 3 groups, with the remaining clustering into small, insignificant populations that are difficult to interpret (Figure 3.3a). Two of the clusters, xA and xC, no longer contain the necessary residues essential in forming the loop-loop interaction present in the original scaffold. xC also obtained several additional point deletions during the selection, which appear to have reinforced a drastically alternative structure (Figure 3.3b). This alternative structure is in fact a member of a group of RNAs originally selected to specifically bind L-tryptophan and this fold presents an exact consensus sequence derived in those studies [197,198]. Over half of the core binding pocket of the L-Trp aptamer is composed of nucleotides from the designed constant regions of xpt and not the library, where select point deletions in the scaffold serve to rearrange the secondary structure and form part of a competent binding pocket with the aid of variable nucleotides from J2/3.

The xB cluster is one of the 10 most abundant clusters in the first round of sequencing, suggesting that it is a very robust solution from the initial pool. It is the only one of the three major clusters present in the final round of selection that is detectable in the first round of selection (Figure 3.4a). Coinciding with the observed elution profile, all observable families show a burst in enrichment in rounds 3 and 4 before leveling off for the final three rounds of selection. Counter selections and increasingly stringent washes seem to have minimal effect on the composition of the pool after round 5, implying that the RNA pool had settled on its solutions by this point. While the extent of mutagenesis was evident from observations
Figure 3.3: Deep sequencing analysis of the purine selection against 5htp. (a) Tree representation of the distance matrix generated by FastTree. It is clear that there are three main clusters (as independently colored). (b) Covariation model of the three observed clusters as generated by the bioinformatics pipeline (Appendix A) [199, 200]. xA (green) and xC (orange) no longer possess the sequence elements necessary to form the designed secondary and tertiary structures. xB maintained the primary sequence necessary for all designed structure, yet acquired a significant amount of mutational diversity during the selection like all sequences.
of the final round of selection, tracking the full length sequences through all rounds of selection reveals that there is a heavy pressure on minimization and destruction of structural elements from the very first round (Figure 3.4b). After only two rounds of selection and amplification, less than 25% of all sequences contain the desired constant elements. This number further falls until it drops below 1% at round four and finishes at roughly 0.25% for the final round. Cluster xB appears to have preserved sequence elements necessary for the designed structures, and it accomplished this in the face of huge mutational pressures. The diversity of xB is quite low in early rounds, but as the selection progresses additional diversity in the form of covarying and accepted mutations within the projected secondary structures are found, which reinforces the structural prediction (Figure 3.4c). The loop-loop interaction is no exception, obtaining point mutations constantly throughout the selection. However, the diversity that xB acquired in the loop-loop region mimics that found in phylogeny for the purine riboswitch (Figure 3.4d), further suggesting that this architectural motif was not only maintained, but critical to the aptamer’s recognition of 5htp. While the possibility that these mutations weren’t present in the initial pool of RNA molecules can’t be definitively ruled out, subsequent sequence analysis of all families suggests that there was in fact a constant influx of accumulated, mutational diversity throughout the selection across all sequences. While drift has been observed during selection previously [201], our strategy offers a unique view on the degree of near-neutral genetic drift present during selection that may eventually give rise to final functional sequences such as those from the xA and xC clusters.

The sequences obtained out of these clusters were highly convergent, with essentially a single sequence produced from each cluster and most of the sequence divergence stemming not from the randomized region, but mutational accumulation in the intended fixed sequence. We therefore chose the most abundant sequence from each cluster for further characterization.
Figure 3.4: Round by round sequencing analysis of the purine selection (a) Histogram showing the per round composition for the three major families of the RNA pool. xB (cyan) is observable after the first round, though with only 0.0075% of the total sequences. xC (orange) begins to dominate at round 2 while xA is not observable until round 4 of the selection, but once either of these families makes an appearance they begin to rise rapidly. By round 7 these three families comprise the majority of the selection pool. (b) The rate of accumulated mutations in the constant regions is staggering, with less than 1% of sequences maintaining all of the invariant nucleotides by round 4 and only 0.25% of sequences by round 7. (c) Covariation models of cluster xB for various rounds of the selection. After the first round of selection there is considerable diversity in the designed constant regions of the RNA, though the randomized junction region shows relatively low diversity. P2 seems to be a site of heavy mutational accumulation by round 2. Round 6 shows preservation of P2 and P3 as predicted by covariation and that the junction region of the RNA has become essentially fixed on a singular solution. The degeneracy of the L2-L3 interaction is fairly limited in round 1, yet gains a relatively steady influx of diversity as the rounds progress. (d) Interesting, the mutational diversity acquired by the xB cluster reinforces the designed loop-loop interaction as it recapitulates near wild-type levels of diversity found in the purine phylogeny (gray).
(denoted by numbered subscripts), with all of these sequences matching their respective cluster’s covariation model. Structure probing by SHAPE [158, 159] reveals that all three RNAs contain unambiguous ligand-dependent changes in their reactivity pattern. In both xA and xB, these were primarily localized to the three-way junction elements (Figure 3.5a red boxes), while xC shows changes outside of J2/3 in the fixed regions that correspond with the predicted loop that houses the L-trp binding site. The reactivity pattern is strikingly similar between the native purine xpt riboswitch and the selected xB_{17} aptamer, with the general level of modification in a given region being fairly similar. The same cannot be said for xA_{10} and xC_{3}, where regions of low modification are highly modified in the selected sequence (Figure 3.5a blue boxes) which suggests that structured and base-paired regions are now more flexible, presumably due to an alternative fold. The most striking preservation of the reactivity pattern occurs in L3 of the xpt riboswitch and the selected xB_{17} RNA. The native purine RNA contains two highly modified residues in this region whose signature is dependent upon formation of the loop-loop structure. xB_{17} is the only sequence of those tested that maintains this characteristic mark (Figure 3.5a yellow boxes), further suggesting that the new aptamer maintains the architecture designed into the selection. Not only does xB_{17} preserve this feature, but the addition of 5htp to solution actually increases the signature’s intensity, suggesting that binding reinforces the interaction (Figure 3.5b). This observation is encouraging, as it indicates that the junction region that houses the ligand is able to communicate to the peripheral elements, and likely *vice versa*. This is a trait critical to the function of the parental xpt riboswitch [75, 123] and other successful synthetic RNAs [131, 202]. Therefore, it seems likely that these dynamics would be useful in the application of the RNA.

To further elucidate the properties of the individual clusters, the selectivity and affinity
Figure 3.5: **Structural probing of select RNA sequences from the selection reinforces the sequencing analysis.** (a) All sequences show ligand dependent sequence modifications, suggesting they do in fact bind 5htp in a productive manner (red boxes). The xB17 shows strong protections and enhancements in the randomized junction regions, much like the wild-type xpt aptamer. Additionally, xB17 shows a strong signature in L3, a signature in the xpt aptamer indicative of tertiary structure formation (yellow box). This feature is not present in either of the other sequences tested. xA10 and xC3 show a different global reactivity pattern as compared to the xpt aptamer suggesting that they adopt an alternative structure (blue boxes). (b) The xB17 aptamer was subjected to further analysis using the software package SAFA [173]. The NMIA reactions in the absence and presence of ligand were integrated and the differential map is shown. The randomized regions show clear protections upon the addition of ligand.
of the three aptamers was assessed by isothermal titration calorimetry (Figure 3.6). All three sequences bind 5htp in the low micromolar range, with \( \text{xB}_{17} \) achieving the highest affinity and greatest selectivity (>70 fold over tryptophan). These affinities and levels of discrimination are comparable to previously selected synthetic and natural aptamers that bind amino acids [163,197,203,204]. \( \text{xA}_{10} \) and \( \text{xC}_{3} \) show a significantly greater discrimination between 5htp and serotonin, suggesting that the main-chain is critical for their recognition of the ligand, a trait already observed by the tryptophan aptamer [197]. The \( \text{xA} \) cluster does not contain any of the distinguishing features of the consensus tryptophan aptamer and its MFE structure is still predicted to contain the three helixes surround the junction, yet the recognition properties of \( \text{xA}_{10} \) and \( \text{xC}_{3} \) are highly similar suggesting they recognize the ligand in a similar manner even though their structures and conserved sequence could be substantially different. \( \text{xB}_{17} \) appears to recognize the ligand using a different approach, with the 5-hydroxy moiety of the indolic ring acting as a major recognition determinant. In fact, the hydroxyl group adds -2.1 kcal/mol of free energy to the binding interaction alone. Additionally, we could not detect any binding to 5htp when we abolished the L2-L3 interaction, further supporting the conclusions from the dynamics studies above.

These data strongly indicate that this selection yielded three distinct aptamers, with each with a unique fold and recognition properties. However, it appears that only the \( \text{xB}_{17} \) aptamer preserved the architecture and elements designed into the structure. Not only that, but the peripheral RNA elements actually have a beneficial impact, conferring the highest affinity and specificity out of all three candidates. This result supports our original hypothesis, that is, by adding in RNA elements with a desired trait, we have not restricted ourselves in functional space or as drastically in sequence space as we had thought since the RNA was readily able to explore alternative structures. In spite of this, the scaffold produced
Figure 3.6: The affinity and specificity are variable among the selected sequences. (a) Table of affinities for representative sequences from each cluster after the amplification cassette has been removed. xA10 and xC3 share a similar binding mode that does not readily distinguish between 5htp and L-trp. xB17 shows the highest affinity for 5htp and the greatest level of discrimination (>70 fold). (b) The removal of amplification cassette has minimal affect on the binding affinity for xB17.
an aptamer that retains the global fold which is a dynamic and functional aspect of the ligands recognition as evidenced by structural probing and the sequencing analysis. These realizations guided our subsequent characterization and studies towards the development of robust RNA parts.

3.3 Crystal structure of the xB_{17} aptamer

In order to further demonstrate the utility of a scaffolded selection, its ease of characterization, and to fully elucidate how xB_{17} recognizes its ligand within the context of a purine scaffold, we pursued the crystal structure. Since sequence analysis and biochemical characterization suggest that xB_{17} maintained a global architecture fairly similar to that of the parental xpt guanine riboswitch, we pursued an identical strategy to solve the crystal structure [73]. This simply entails modulation of the length of the P1 stem (between 4 and 7 base pairs) and screening constructs against a sparse matrix designed for RNA crystallization (Hampton Research, Section 3.6). The resulting crystals diffracted to 2.0 Å resolution in the presence of iridium hexamine. The unbiased electron density map was then calculated using single isomorphous replacement with anomalous scattering (SIRAS) in Phenix [170] and the structure modeled to the density in coot [171] resulting in an $R_{work}$ and $R_{free}$ of 22.6 and 25.0%, respectively (Table 3.1).

The final electron density map with the pertinent structure modeled reveals that all regions of the RNA are well resolved, especially in the junction region (Figure 3.7). As a test of the overall structure and ligand placement, a composite omit map was generated after the final round of refinement. This technique divides the unit cell into small blocks and serially omits the contents of these blocks generating a set of omit maps that are then combined to produce a single omit map for the entire contents of the unit cell. This yields an
unbiased map of the electron density for the entire unit cell without severely compromising phase quality and is a harsh test of the validity of the model. The composite omit map of xB17 retains density in the binding pocket that allows for the unambiguous placement and orientation of the ligand along with the surrounding nucleotides indicating that the model is indeed accurate (Figure 3.7b).

The overall fold of the RNA is also consistent with the biochemical analysis with all of the selected residues remaining localized to the junction region and not significantly altering the overall fold of the RNA architecture as compared to the xpt structure (compare Figure 3.8 a and b). The selected residues completely envelop 5htp and the constant architecture of the RNA plays no direct role in its recognition. The new junction forces helixes P2 and P3 about 6 Å further apart than in the parental xpt aptamer, yet the r.m.s.d. for the two overlaid structures remains reasonably low at 2.5 Å for all alignable atoms. While the P1 stem for both RNA aptamers is A-form, they position themselves relative to P2 and P3 at radically different angles (Figure 3.8c). If P1 is omitted from the structural alignment, the r.m.s.d. drops just below 2.0 Å and is likely a more accurate depiction of the similarity between the two RNAs. Additionally, the peripheral loop-loop interaction is consistent between the RNAs, with an r.m.s.d. of 0.54 Å for all atoms in the motif, which begins to approach the coordinate error of 0.22 Å for the structure. While structural probing had indicated that the P2 and P3 helices were intact and that the loop-loop interaction was formed, this degree of structural preservation is striking, and the structure definitively shows that the peripheral elements were maintained and dictate the fold of the new RNA.

The backbone trace through the three-way junction and randomized region is again grossly preserved, yet the composition is entirely unrelated to the parental RNA and the network of interactions is altogether unique. Most remarkable about the structure is the
# Extended Data Table 1

Data collection and refinement statistics

<table>
<thead>
<tr>
<th><strong>Data collection</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>C121</td>
</tr>
<tr>
<td><strong>Cell dimensions</strong></td>
<td></td>
</tr>
<tr>
<td>$a$, $b$, $c$ (Å)</td>
<td>127.5, 26.6, 63.4</td>
</tr>
<tr>
<td>$\alpha$, $\beta$, $\gamma$ (°)</td>
<td>90.0, 106.3, 90.0</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>19.9-2.0(2.07-2.00) *</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>0.08(0.19)</td>
</tr>
<tr>
<td>$I/\sigma I$</td>
<td>11.2(5.4)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>96.2(73.8)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>4.34(3.65)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Refinement</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>19.9-2.0(2.07-2.00)</td>
</tr>
<tr>
<td>No. reflections</td>
<td>59533</td>
</tr>
<tr>
<td>$R_{work}$/ $R_{free}$</td>
<td>21.9/26.2</td>
</tr>
<tr>
<td>No. atoms</td>
<td></td>
</tr>
<tr>
<td>RNA</td>
<td>1513</td>
</tr>
<tr>
<td>Ligand</td>
<td>16</td>
</tr>
<tr>
<td>Water</td>
<td>112</td>
</tr>
<tr>
<td>B-factors</td>
<td></td>
</tr>
<tr>
<td>RNA</td>
<td>29.3</td>
</tr>
<tr>
<td>Ligand</td>
<td>35.0</td>
</tr>
<tr>
<td>Water</td>
<td>23.5</td>
</tr>
<tr>
<td>R.m.s deviations</td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.007</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.31</td>
</tr>
<tr>
<td>Maximum likelihood coordinate error (Å)</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Number of crystals for each structure should be noted in footnote.

*Highest resolution shell is shown in parenthesis.
Figure 3.7: The crystal structure of xB\textsubscript{17}. (a) F\textsubscript{O}-F\textsubscript{C} difference map of the xB\textsubscript{17} structure contoured to 1 \(\sigma\). All regions of the RNA are well defined by the electron density, making placement of the residues and backbone unambiguous. (b) Composite omit (top) and fully refined (bottom) F\textsubscript{O}-F\textsubscript{C} difference map contoured to 1 \(\sigma\) reveal clear ligand density in the binding pocket even in the absence of model bias.
Figure 3.8: **The crystal structure of xB\textsubscript{17}**. (a) Cartoon representation of the 3D architecture of xB\textsubscript{17} with the variable region (cyan) and ligand (orange) highlighted. (b) 3D representation of the parental RNA in complex with its cognate ligand for comparison. (c) Overlay of the xB\textsubscript{17} aptamer and \textit{xpt}. The gross overall structure is very well preserved.
degree to which the variable region communicates and cooperates to form a robust binding pocket while leaving the invariant regions intact. Nearly every selected residue is involved in an interaction that aids in the formation of the junction, often contacting elements distal in sequence space (Figure 3.9). The P1 stem of xB17 involves many distinct interactions between J1/2 and J3/1 which serve to stabilize and orient critical nucleotides for the recognition of 5htp. For example, N2 of G20 makes a contact with the Hoogsteen face and phosphate of A76. This serves to break the A-form helix, and orient A21 such that its Hoogsteen face may be recognized by U51 of J2/3 (Figure 3.10a). This reinforces an interaction between N6 of A21 and the phosphate backbone of C75 in J3/1, a nucleotide that is critical in stabilizing the minimal ligand binding pocket described below. G23 appears to be a keystone for the aptamer as it makes several hydrogen bond contacts with G52 of J2/3, stabilizes the Watson-Crick face of U72 with its backbone in J3/1, coordinates an iridium hexamine molecule in the minor groove of P3, and mediates the coaxial stacking of the 3' side of P1 with P3 (Figure 3.10b). Interestingly, the wild-type B. subtilis xpt RNA has a miss matched C26–A44 pair at the base of P2 leading into the junction. xB17 compensates for this weak pair by pushing C26 out into solution, and the randomized position 25 was fixed on a uridine that now Watson-Crick pairs with A44. While these residues do not directly interact with the ligand, they form a compact and efficient structure that provides the basis for the RNA to recognize the ligand.

The core binding pocket of xB17 houses a single site for 5htp and is primarily formed by J2/3 using a common RNA structural module, the T-loop receptor or U-turn [206]. The first five nucleotides of the randomized J2/3 form a canonical T-loop structure, as originally observed in tRNA and commonly found in a number of other RNAs. This motif is defined by a reverse Hoogsteen U-A pair between positions 1 and 5 which orchestrate
Figure 3.9: The annotated secondary and tertiary structural interactions of xB17. The secondary structure is illustrated using the nomenclature of Leontis and Westhoff [205] where solid lines indicate canonical Watson-Crick pairing, dashed lines represent unique tertiary interactions such as single hydrogen bonds between residues, and additional interactions between residues are detailed in the inset table. This depiction of the structure shows the degree of interaction presented by the junction region and nearly every selected nucleotide (cyan) participates in the formation of the junction’s structure. Nucleotides responsible for direct interaction with 5htp are shown in red. The drawing is ignorant of the stacking interactions occurring within the junction, the L2-L3 interaction, and between P1 and P3.
Figure 3.10: **Residues from J1/2 mediate many interactions critical for the structure of the RNA** (a) Several nucleotides within J1/2 participate in the formation of the binding pocket and stability of the RNA. A$_{21}$ (red) interacts with the Watson-Crick face of U$_{51}$ of J2/3 (teal) whose backbone is critical to the specificity of the RNA for 5htp (orange). A$_{21}$ also stabilizing the backbone of C$_{75}$ (purple) which forms a tertiary interaction with J2/3. U$_{25}$ (blue) displaces the mismatched C$_{26}$ of P2 (pale green) to form a canonical Watson-Crick pair and serves to orient the junction region. (b) G$_{23}$ (green) is a keystone nucleotide within the junction region of the RNA. It serves to organize G$_{52}$ in J2/3 (teal), stabilizes the backbone transition between P1 (pale orange) and P3 (pink), mediates the coaxial stack between P1 and P3, and coordinates an iridium hexamine (gray) in the minor groove of P3. Nearly every hydrogen donor and acceptor of G$_{23}$ is utilized in the structure.
a three nucleotide loop (positions 2-4) [207]. In tRNA, position 3 of the T-loop forms a canonical Watson-Crick pair with G₁₉ of the D-loop which serves to orient G₁₈ such that it may stack between positions 4 and 5 of the T-loop. Both interactions are essential for tRNA function. In xB₁₇, positions 4 and 5 are A₄₈ and A₄₉ and instead of housing a nucleotide they recognize the 5htp ligand (Figure 3.11a). Additionally, position three of the xB₁₇ T-loop makes a tertiary Watson-Crick pair with C₇₅ of J₃/1, fully recapitulating the T-loop motif. The xB₁₇ aptamer’s T-loop overlaid with naturally derived variants from tRNA_Phe and the TPP riboswitch yields an r.m.s.d. of <1.0 Å for the backbone atoms even though the base composition varies significantly between the RNAs (Figure 3.11b). The orientation of the intercalating residue is also highly conserved and in the natural RNAs interacts with position 2 in a similar manner to U₄₆ of xB₁₇. The majority of direct contacts with the ligand occur through the T-loop motif, with the exception being the phosphate of U₅₁ which is responsible for the recognition of the hydroxyl group of the indolic ring. O₂ of U₄₆ (position 2) associates with the ε-nitrogen of the indolic ring, while A₄₉ (position 5) stacks with the ligand and recognizes the main chain atoms via its Watson-Crick face. As expected, the main chain atoms are accessible to the solvent since this was the site of coupling to the solid support medium. The only functional handle not recognized by the RNA was the one oxygen of the α-carboxyl group that was converted to an amide and poly-carbon linker on the affinity column. To account for the additional charge of the uncoupled carboxylate group on the ligand, the RNA recruits an iridium hexamine in the crystal structure, likely mimicking a hexahydrated magnesium under the selection conditions. Besides this, every polar functional group of the ligand is used productively by the RNA, thus supporting the specificity profile established by ITC.

The crystal structure also agrees very well with the structural probing experiments.
Figure 3.11: The binding pocket of xB_{17} is compact and recognizes the ligand via a T-loop motif. (a) All of the residues that directly recognize the ligand are located in J2/3 and were part of the library. A_{49} is responsible for stacking and the recognition of the main chain residues, while the phosphate of U_{51} and O2 of U_{46} mediates recognition of the side chain. (b) The T-loop motif consists of five residues (T1-T5) that for a sharp turn and house an intercalating residue between T4 and T5 with T3 often supported by a tertiary interaction. Overlay of the xB_{17} T-loop motif with that of the TPP riboswitch and tRNA_{Phe} shows that the backbone structures are essentially identical and all critical features of the motif are present.
Once the differential probing map from Figure 3.5b is projected upon the crystal structure, it becomes clear that the majority of the ligand dependent signal is localized around the varied regions constituting the binding pocket (Figure 3.12a). 5htp, like the intercalating adenosine in tRNA\textsubscript{Phe}, must be instrumental in the formation of the T-loop as all five RNA residues involved in the motif undergo a marked decrease in their modification upon ligand addition. Some interesting trends also become evident, such as nucleotide A\textsubscript{21}, which is weakly associated in the structure yet forms part of the hydrogen bonding network that organizes the ligand binding pocket. This position undergoes a ligand dependent decrease in its reactivity along with its binding partner U\textsubscript{51} showing a reinforcement of this interaction in the bound state (Figure 3.12b). C\textsubscript{26}, which is part of the C–A mismatch at the base of P2 in the \textit{B. subtilis} RNA, is heavily modified in the bound state where U\textsubscript{25} substitutes for this residue to aid in formation of the T-loop and initiates the A-form helix of P2. G\textsubscript{23} is again implicated as a critical linchpin in the structure via this analysis, where U\textsubscript{53} and U\textsubscript{72} show a large conformational change located around this residue. U\textsubscript{53} is oriented out into solution while U\textsubscript{72} is weakly organized in the bound structure with the probing indicating that U\textsubscript{53} undergoes a large increase in modification and U\textsubscript{72} becomes substantially protected, suggesting that these two nucleotides swap roles interacting with G\textsubscript{23} in a ligand dependent manner (Figure 3.12c). This analysis again demonstrates that the RNA does not simply recognize the ligand in a small, restricted space of the primary sequence and the entirety of the RNA is involved in the dynamics used to accommodate the ligand.

This synthetic RNA aptamer resembles natural riboswitch aptamers in many ways. It has a high specificity and affinity for its ligand, is able to communicate binding to peripheral sequence elements, and has a robust network of interactions that dictate its ligand binding properties. While the primary sequence of xB\textsubscript{17} gives no indication that the helical junction
Figure 3.12: Projection of the SHAPE reactivity pattern onto the crystal structure. (a) Overview of the projection shows that the majority of the ligand dependent reinforcement is located in the junction region proximal to the ligand. The distal signal in L3 is indicative of the tertiary formation and an enhancement of this signature shows that ligand binding is able to influence the global dynamics of the RNA. (b) Interesting patterns are revealed by this analysis such as the interplay between $A_{21}$ and $U_{51}$ which could not be interpreted without this combination of techniques. (c) The RNA localizes both $U_{53}$ and $U_{72}$ next to each other in 3D space. The probing pattern suggests that these two nucleotides have different conformations in the ligand bound and unbound states, likely centering around their interaction with $G_{23}$. 
would preserve its overall trends, it is uncannily so. While the G₄₇–C₇₅ interaction between J₂/₃ and P₁ is undoubtedly a function of the T-loop motif, it recapitulates a fold of the parental RNA where nucleotides 49 and 50 of J₂/₃ in purine riboswitches interact with the minor groove of the last two base-pairs of P₁ next to the ligand binding site. This interaction is essential to purine recognition and highly conserved (Figure 3.1a). Therefore, it seems likely that the implementation of the L₂-L₃ interaction as a scaffold has promoted a similar overall positioning of J₂/₃ and produced an RNA that possesses characteristic dynamics and a composition similar to that of natural riboswitch aptamer domains.

3.4 Investigations into xB₁₇ as an RNA sensor

Using the xpt purine scaffold as a host for a limited library has proven successful in the development of a novel robust aptamer, capable of achieving high specificity and affinity in the context of the peripheral motifs. Additionally, the architecture lends itself to rapid biophysical characterization should that be required for investigations into novel parts. These studies also point towards the aptamer behaving in a manner similar to that of natural riboswitches, with the junction region influencing global fold and dynamics upon ligand binding. Therefore, the aptamer should function well as an RNA sensor domain, hence I sought to couple the aptamer to various adaptors in order to test the ability of xB₁₇ to modulate an actuator in a ligand dependent fashion.

Our lab has recently demonstrated the modularity of select riboswitch adaptors (expression platforms) and their ability to translate their in vitro function into a cellular phenotype [79]. Various expression platforms from existing natural riboswitches that contained no sequence overlap with the recognition domain were engineered to host natural and synthetic RNA aptamers, resulting in a suite of modular RNA devices. Among these platforms was
a sequence derived from the promoter region of the metE gene in B. subtilis that naturally regulates the intracellular levels of S-adenosyl methionine (SAM). The metE adaptor can be engineered to host the parental guanine responsive xpt aptamer, and when coupled the novel device is able to attenuate transcription of the E. coli polymerase in vitro at a concentration similar to the K_D for guanine of xpt. While a dynamic range for this device of only 50% seems low, the maximal termination is near 100% making this an ideal switch for in vivo application with a low background expression of the actuator. Since natural pools of guanine exist intracellularly at levels that would trip the sensor regardless of ligand concentrations in the media, the C74U mutant was used to validate the switch in vivo. This mutant has a substantially reduced affinity (9 μM versus 0.025 μM) and the specificity is altered from guanine to adenine and 2-aminopurine (2AP) facilitating the device’s function in vivo. Indeed, once ported into a host E. coli strain the RNA device was able depress expression of a coupled fluorescent protein actuator by 12 fold, though a concentration of 730 μM 2AP was necessary to elicit a half-maximal response (a value 75x higher than the observed K_D). This level of reduced affinity when moving into in vivo applications is not uncommon in the field, with the likely cause being the cellular metabolism of the ligand. Additionally, the metE adaptor platform was able to host a synthetic theophylline responsive element with only a moderate 10-fold discrepancy between the K_D and concentration of ligand necessary to elicit a half-maximal response in vivo. This suggested that the metE platform is a good starting point for the implementation of a 5htp responsive RNA device.

A simple screening strategy was employed for coupling the xB_{17} aptamer to the metE expression platform. This involves the simple modulation of a small communication helix that serves to couple the two parts and variation of nucleotides on the 5’ side of metE P1 that serve to alter the stability between the two effector states of the switch (Figure 3.13a).
Initially, three constructs were tested that only varied the length and composition of the communication module (CM, green nucleotides: A-C). One of the constructs, xB\textsubscript{17}A/met\textsubscript{E}, resulted in a device capable of attenuating transcriptional read-through (RT) in the presence of 5htp (Figure 3.13b). The dynamic range of xB\textsubscript{17}A/met\textsubscript{E} is very narrow at \(\sim 15\%\) and the device is weak with a half-maximal response of \(>700\mu\text{M}\). However, the switch does approach near complete termination in the presence of high ligand concentrations and readily discriminates against tryptophan while showing no activity for the cognate ligand guanine (Figure 3.13c).

Given the initial results, the engineering strategy then became two fold: to reduce the amount of background by increasing the percentage terminated in the presence of ligand and to increase the dynamic range to a more reasonable value. The limited dynamic range is due to the inability of the aptamer to effectively communicate binding through the communication module to the met\textsubscript{E} adaptor. A likely reason is that either the adaptor’s stability lies outside of the effective range (i.e. being too stable) or that xB\textsubscript{17} is preformed to a large extent and unable to dynamically translate the energy of binding to the adaptor. Since structural studies indicate that ligand binding does in fact attribute energy to the global fold and we are limited in our ability to modify the aptamer as only a single functional sequence was obtained from the selection, I probed the stability of the communication module. By decreasing the stability of the CM (as in constructs o-o2 in Figure 3.13a) the P1 stem of met\textsubscript{E} should be preformed to a lesser extent leading to a higher degree of read-through product regardless of ligand concentration. To compensate for the loss in termination efficiency, variants that stabilized the P1 stem of met\textsubscript{E} (+1 and +2) were also introduced to influence the equilibrium between the terminated and read-through states. The low dynamic range could also be a result of the nucleotides present in P1 of met\textsubscript{E} preforming regardless
Figure 3.13: **Strategy for engineering the 5htp aptamer xB₁₇ onto the metE natural riboswitch platform.** (a) Sequence and secondary structure of the xB₁₇ 5htp aptamer (cyan) and the metE riboswitch expression platform adaptor (purple). The two RNA parts were coupled using a small communication module (CM; green) that was varied to screen for functional constructs. The 5' side of the metE P1 was also altered to screen for increases in dynamic range and % termination. The structure of metE is depicted in the bound and off state due to the formation of a rho independent terminator stem (PT). The switching sequence that is differentially sequestered in the on and off states is outlined in red along with its alternative state binding sequence. (b) One construct, xB₁₇A/metE, was able to attenuate transcription in a ligand concentration dependent manner. (c) A histogram comparing the dynamic range and specificity profiles of the parental xpt switch in metE to the new 5htp sensor.
of the state of the sensing domain, therefore mutations in P1 were introduced that simply destabilized the terminated state (-4 through -1) in order to tune the levels of expression. Many of these possible variants were screened in a matrix setup that surveyed dozens of possibilities, yet none of the constructs yielded a functional device to any extent, with the results being often ambiguous and following no clear trend.

While the initial xB_{17}A/metE device resisted further optimization and engineering, it is still a switch. Most impressive about the switch is the initial working construct was designed ignorant of the crystal structure, implying that the level of characterization I have achieved with xB_{17}, while beneficial and enlightening, is not necessary for a scaffolded RNA aptamer to be implemented as a functional device. The resistance to engineering may also be a function of the adaptor platform and not xB_{17}, as the delicacy of the metE platform has been observed before [79]. The platform only seems capable of functioning in a very narrow window of thermodynamic space where nearly any attempt to tune the expression platform via sequence alteration can ablate the switching behavior entirely.

The metE platform was chosen due to the parental scaffold’s success in its context, however the platform has not been a successful host of the xB_{17} aptamer. Fortunately, several additional riboswitch adapters are available that include the B. subtilis pbuE adenine and Dechloromonas aromatic metH S-adenosyl homocysteine responsive riboswitches [84]. Contrary to metE, these platforms are functional ‘on’ switches that promote the expression of an actuator in the presence of ligand and again have been shown to function in the context of natural and synthetic aptamers with minimal engineering, including xpt. The B. subtilis pbuE adenine riboswitch as a device, not just an aptamer domain, is also well understood with its function being the subject of numerous studies ranging from the single molecule level to intracellular characterization and optimization [85, 208, 209].
These two switches are not modular in their native capacity, with the switching sequence involving a substantial amount of overlap between the aptamer and the expression platform and \(\text{metH}\) is additionally constrained by a pseudoknot interaction. In both cases, additional sequence elements were added to create a conventional P1-type helix to which aptamer recognition domains could be appended. Both of these platforms also proved more amenable to sequence engineering than \(\text{metE}\) as their respective dynamic range and levels of read through are tunable with simple point mutations in the P1 stem. The results of these studies are reflected in the engineering scheme employed to create a 5htp transcriptional 'on' device (Figure 3.14a and b). All of these devices were initially cloned into a vector that placed the \(\text{xB}_{17}\) device upstream of the GFPuv fluorescent protein actuator and screened for their ability to attenuate levels of GFP fluorescence. The parental \(\text{xpt\ C74U 2AP}\) aptamer in the \(\text{metH}\) adaptor exhibits a 12 fold increase in fluorescence upon the addition of ligand, however the engineering strategy described in Figure 3.14 devices were always very dark with levels of fluorescence approaching background and the fluorescence could never easily be rescued, so this construct did not see any further development in this application.

The initial \(\text{pbuE}\) constructs (A-C) with a 6U terminator stretch showed a very high level of fluorescence in a rich, defined CSB media that lacked L-tryptophan and no ligand dependent effect was observed. Therefore, the strategy became to destabilize the communication module to allow for the energy of binding to be more effectively communicated to the adaptor sequence (o-o2 constructs) and to reduce the overall levels of expression by increasing the number of uridine residues after the rho independent terminator stem from 6 to 8 (8U). The result was a set of devices barely capable of inducing the expression of the actuator. Unfortunately, the system seemed crippled by the adverse effects of high concentrations of ligand on the cells. At the levels of ligand needed to induce a response the cells
Figure 3.14: **Strategy for coupling the 5htp aptamer xB17 onto the pbuE and metH engineered riboswitch platforms and in vivo screening.** (a and b) The design for engineering the 5htp aptamer xB17 onto the pbuE adaptor platform (orange) or metH (black). Again a small communication module (CM: green) was used to couple the two RNA parts and screened based on varying stability and length. Additional mutations at the base of P1 were inserted with the intention of tuning the response range of the RNA devices. The switching sequence and its alternative pairing sequence are again shown in red boxes (c) Table depicting the additional considerations taken while while screening potential devices. Various media conditions and actuators were attempted in order to mitigate the effects of the ligand concentrations need to elicit a response. (d) Various potential RNA devices screened in a rich chemically defined media containing 50 μM L-trp with a β-Galactosidase actuator. Inset shows that the devices do in fact switch *in vitro*, despite their poor performance *in vivo*.
exhibited a much reduced growth rate, usually beginning around 200-300 μM 5htp. This is within the realm of intracellular tryptophan pools [210] and likely begins a significant level of poisonous incorporation into critical protein systems, reducing the growth rate and inducing cellular stress responses. In attempts to remedy the effects of ligand on the cells, media conditions were screened that involved altering the amino acid source, the levels of tryptophan in the media, and the salts used for growth (Figure 3.14c). While increasing the levels of tryptophan in the media rescued cell growth and salt conditions could decrease background, there was never an observed ligand inducible signal with the fluorescent protein actuators. When the fluorescent actuator was swapped for the B-Galactosidase enzyme and the Miller assay used to quantify the levels of induction after the growth in a defined media containing 50 μM tryptophan, some constructs gave a small induction in signal at modest 5htp concentrations (1 mM). Additionally, the trends observed in the assay matched our intended effect, where stabilizing P1 increased the fluorescence, stronger terminator sequences decreased signal, and alterations to the communication module slightly tuned the dynamic range (Figure 3.14d). Several of the switches were shown to attenuate the transcription of host E. coli polymerase in vitro, yet the in vivo results were never statistically significant, likely an effect of the ligand induced stress upon the cells leading to a large degree of error and delicacy of the assay.

Even with the pbuE platform proving amenable to engineering and tuning efforts, the effects of the ligand in vivo could never be escaped and the pressures imposed by the reliance on host machinery proved too great in this context. The fluorescently coupled adaptors and actuators of the malachite green aptamer (MGA) and Spinach now become a tempting prospect, as they readily function in minimal in vitro assays in the complete absence of proteins and only indirectly rely on host systems intracellularly. Additionally, they can be
Figure 3.15: The engineered malachite green RNA based sensor is a robust in vitro reporter of 5htp concentration. (a) The sequence and secondary structure of the malachite green RNA aptamer appended in both orientations to the xB_{17} aptamer with the conditionally fluorescent malachite green small molecule shown (inset). (b) Cartoonized depiction of how the fluorescent RNA sensors work and validation of the 5htp:malachite green in vitro sensor. The malachite green aptamer appended in the B orientation shows much higher maximal brightness yet at the expense of dynamic range and affinity. Both constructs assayed at 1 mM MgCl_{2}. (c) Magnesium influences the maximal fluorescence, dynamic range (top), and the apparent K_{D} of the devices (bottom). Increases in magnesium result in higher levels of fluorescence and a tighter affinity, but at the expense of dynamic range leading to decreased signal-to-noise.
treated as inducible elements, where the levels of sensor are propagated in the absence of ligand and then signal induced at stationary growth phase, presumably eliminating the toxic effects of 5htp on the cell and sensor. A very simple initial scheme was developed based on the work of Stojanovic and Kolpashchikov [113], where the malachite green aptamer is appended to xB_{17} via a short communication module analogous to the previous transcriptional regulators (Figure 3.15a). An additional level of engineering is available for the malachite green aptamer in that its orientation relative to the sensor can be altered. The binding site is a simple defect within an otherwise consistent RNA helix, allowing for the sensory RNA element to be appended to either side of the binding pocket.

xB_{17} when appended to either side of the malachite green binding pocket (orientation A or B) exhibits a fluorescent signal whose intensity is dependent upon the concentration of 5htp (Figure 3.15b). Interestingly, the alternative orientations have nuanced differences in maximal intensity, affinity, and dynamic range suggesting that these devices are tunable for a desired application. These sensors are, again, not without their faults. At physiological concentrations of magnesium, the affinity for 5htp of the RNA sensors is significantly reduced, requiring millimolar levels of ligand in order to induce maximal response. The sensors also exhibit a strong magnesium dependence the concentration of which influences the dynamic range, levels of fluorescence, and overall affinity for 5htp (Figure 3.15c). At magnesium concentrations that are equivalent to the ITC data, the sensor is ~18x weaker than the xB_{17} aptamer alone, a value on par with sensors created previously [113].

The efforts with the malachite green aptamer produced robust functional devices in vitro, yet they are not applicable intracellularly as they exhibit severe cytotoxic effects [106]. However, this demonstrates the facile creation of a device whose sensing and function is entirely carried out by the RNA sequence. This lead us to pursue the implementation of
Spinach devices whose functional philosophy is identical to that of the malachite green aptamer, yet is amenable to in vivo application. The Spinach RNA aptamer is a mimic of the green fluorescent protein, selected to bind an extricated derivative of the GFP chromophore 3,5-difluoro-4-hydroxybenzylidene imidazolinone or DFHBI. This ligand exists exclusively in the phenolate form (the fluorescently competent state) at physiological pH and its fluorescence is conditionally dependent on its binding to the RNA. The ligand is cell permeable, exhibits low background fluorescence, and has no cytotoxic effects while the RNA is genetically encodable allowing the device to easily be ported and cloned in vivo [106]. Spinach has also been shown to function as a small molecule RNA biosensor of guanine in the context of the $\text{xB}_{17}$ parental aptamer $\text{xpt}$. The coupled $\text{xpt}$:Spinach RNA shows a nearly 20 fold induction of signal upon the addition of guanine and several other Spinach reporters were successfully used to monitor metabolite concentrations of SAM [117] and cyclic-di guanosine [118, 119] in vivo suggesting that this may be an adept platform for which to engineer a 5htp dependent RNA biosensor.

An identical engineering strategy to that employed in the malachite green aptamer and the previously established guanine sensor [117] was used to produce Spinach based devices, where the insertion of a small helical element couples the RNAs and transduce the energy of binding from one aptamer to the other. The intended design scheme is shown in Figure 3.16a, however this later proved to be incorrect as elucidate by the crystal structure [107]. As a more accurate representation of the constructs, the scheme and structure has been modified to produce a more real picture of the RNA devices in Figure 3.16b. An unintended consequence as a result of the misinformed structure is a bulged uridine immediately following the communicating helix. This likely played some role in alternative structures and an overall destabilization of the device, evidenced by the fact that the only two constructs capable of
Figure 3.16: The design of 5htp sensors that modulate DFHBI fluorescence. (a) The GFP chromophore (green) and the 3,5-difluoro-4-hydroxybenzylidene imidazolinone (DFHBI) conditional chromophore used in the subsequent studies. The initial devices and all small molecule binding sensors reported to date were built using the errant predicted structure shown in purple. The scheme to produce these structures was the same as for malachite green and previously described [117]. (b) A structurally informed depiction of our initial strategy to produce a 5htp dependent fluorescent sensor. The structure revealed a bulged uridine immediately following the designed communication module (green) which complicates the strategy. (c) The results of screening the various communication modules. Only constructs E and F, which were the most stable in design, were able to produce fluorescence and a ligand dependent change in intensity. (d) The xB_{17} F device possess a >5-fold induction of fluorescence upon addition of ligand and an apparent K_D of 160 μM.
inducing fluorescence contained the most stable helixes (E and F Figure 3.16c). Even so, the best sensor was able to induce a >5-fold increase in fluorescence upon addition of 5htp and an apparent $K_D$ of 160 μM (Figure 3.16d). The observed dissociation constant is 40-fold higher than the affinity of $xB_{17}$ alone, a likely consequence of the intrinsic destabilization of the communication helix.

The crystal structure revealed that the true secondary structure of the Spinach aptamer was drastically different from that of the predict structure initially published. It also suggested that the core binding motif of the Spinach RNA was in fact much smaller than initially thought and, as with so many synthetic aptamers, located between two A-form helixes. From an engineering standpoint this is beneficial as the Spinach adaptor may now be appended in both orientations to a small molecule sensing domain while the communication module can be designed more efficiently and predictably. The $xB_{17}$ 5htp aptamer was appended to both orientations (A and B) of the minimal Spinach aptamer (minSpinach) and using the knowledge gleaned from previous studies a suite of communication modules were used to develop a 5htp sensor (1-3; Figure 3.17a). With the more accurate information of the crystal structures many of the sequences tested showed a ligand dependent increase in fluorescence (Figure 3.17b). The dynamic range of these sensors ranged between 2 and 10 fold induction with varying levels of brightness. Interestingly, the aptamer alone, that is with $xB_{17}$ replaced by a stem loop, in the B orientation is only 60% as bright as in the A orientation. This is likely a consequence of the known folding problems inherent to spinach and is reflected in the overall levels of fluorescence seen with any device created using the B orientation.

All of the 5htp devices show a decreased fluorescence relative to the minSpinach aptamer alone with none being able to achieve even 1/10th the brightness of the minSpinach
Figure 3.17: **Construction and optimization of the minimal spinach platform.** (a) The x-ray crystal structure revealed a more minimal sequence for the spinach aptamer (purple) that recognized the ligand with a G-quadruplex (red nucleotides). This afforded the ability to alter the orientation of the minimized spinach aptamer to increase the degrees of engineering potential. Similar communication module strategies were employed as before (green). (b) Several of the initial constructs in both orientations produced a ligand dependent response, yet the overall fluorescence was significantly impaired. (c) A trend has now emerged in the relative stability of the CM necessary to transduce the binding signal independent of the CM length. The length of the CM can be increased in order to rescue some levels of fluorescence without adversely affecting the dynamic range or the affinity (d).
aptamer in the A orientation. This lead us to reconsider the design of the communication module in order to recoup more of the fluorescent signal of the parental RNA. Previous successful designs used a much longer communication stem to transduce the binding signal to Spinach, in part due to the structural misunderstanding. While the communication stem was much longer, it was significantly destabilized by the bulged uridine residue. The relative thermodynamic stability of all function communication modules lies in the predicted range of -5-8 kcal mol$^{-1}$, regardless of length. This trend holds true in the initial working Spinach and minSpin constructs, where Spinach CM F and the minSpinach CMs 1, 2, and 3 have predicted stabilities of -6.3, -4.9, -5.8, and -7.9 respectively. With this in mind an additional suite of constructs were made that contained an elongated 4 basepair stem, yet stayed within the thermodynamic realm known to effectively communicate a switching behavior. Of the five constructs tested in the A orientation, four of them displayed a robust switching behavior that increased overall brightness of the device to nearly 25% of the minSpinach aptamer and maintained 4-7 fold levels of induction. The apparent affinity of the devices is also comparable to the original Spinach construct, ranging from 110 to 215 $\mu$M (Figure 3.17d) which illustrates that the length and composition of the communication module can affect the properties of the actuator without significantly altering the sensor’s.

It’s important to note that despite the ease of their construction and predictable behavior, both the coupled Spinach and minSpinach constructs were unable to function in vivo. This is likely the culmination of several issues inherent to the Spinach aptamer. Thermodynamic stability, folding, and magnesium dependence of the Spinach aptamer have been problems realized since the initial publication, with only two additional publications on the matter of small molecule systems in the literature, both of which use the same cyclic-di-GMP riboswitch aptamer. Several studies have sought to improve the thermodynamic
stability and folding of Spinach, but the results are not readily amenable to small molecule sensors [104, 105]. It is also difficult to compare our aptameric sensor to others, as there is no published data on how coupled sensors relate in brightness to the fluorescent aptamer alone and only one case of observed affinity for these coupled devices (500 fold higher than the aptamer alone [118]), likely because they compare unfavorably. The Spinach aptamer alone is also difficult to observe intracellularly on its own, often with signal barely above the background of the unbound DFHBI making observation of a sensor such as those constructed above impossible. The compounding effects of inefficient folding, low intracellular magnesium concentrations, low signal to noise, and starting with a fluorescently impaired device have significantly hindered our attempts to develop a fully RNA based device that translates well into intracellular environments.

3.5 Discussion

The selective techniques described above have produced a competent aptamer against 5htp that is well folded and easily characterized. Despite the selective pressures placed upon the RNA pool, the xB cluster maintained its overall architecture acquiring mutations that were either accepted in or reinforced the designed fold. Not only was the fold preserved, but it seems to have had a beneficial impact on the recognition characteristics of the aptamer, as it produces the highest affinity and greatest discrimination for the ligand. The pressure to preserve the scaffold must have been enormous, as no other sequences belonging to any other pool contain the peripheral motif and were able to survive the partitioning pressures without it. The peripheral interactions are also able to inform on the binding state of the aptamer, as a global reinforcement of the structure is observed upon the recognition of ligand in structural probing experiments. This suggests that if the binding event can contribute energy to the
formation of the tertiary structure, it should be adept at communicating to adapter and actuator elements to create proficient devices. Thus our small and highly biased library is not only capable of producing a competent aptamer, but one that possesses characteristic traits associated with accomplished RNA sensor domains such as theophylline, neomycin, and many natural riboswitches.

It is interesting to note that nearly every varied nucleotide in the xB$_{17}$ aptamer, and therefore the xB cluster, was used to structure and form a binding pocket. This is fairly unique in RNA selections, where a typical library size of 60 to 100 nucleotides yields only a simple motif that can often be reduced to a small internal loop consisting of less than 20 nucleotides [33]. By introducing pre-selected and complex elements that attribute beneficial characteristics to the selection, we avoid the immense constraint of structural context normally placed on the selection allowing for more intricate solutions to be sampled. The strategy didn’t limit the pool to complex solutions however, as more than half of the final pool did not contain the peripheral motif and produced structurally simple solutions such as the tryptophan aptamer [197,198]. Therefore it seems that we have not restrained ourselves in sequence space, we have simply raised the ceiling to attain more complex and proficient aptamers.

It is clear that the xB$_{17}$ 5htp is a capable aptamer possessing complex behavior and structure, yet it’s ability to function as an *in vivo* reporter is complicated. The *in vitro* systems used to validate the switch clearly show that xB$_{17}$ is adept at communicating the binding event to the adaptor module, be it a pseudo-natural riboswitch platform or a synthetically derived fluorescent RNA molecule. While the observed affinity of these devices is significantly weaker than the aptamer alone, this is an attribute inherent and observed in all artificial RNA devices. This is not unexpected, as the affinity and many characteristics of
the device are a function of the aptamer domain, which was never pressured to undergo large conformational shifts upon ligand binding during the selection. It could be argued that the opposite is in fact true, where a preformed and stable RNA structure that rigidly displays a binding pocket for recognition is selected. Therefore, in order to elicit large conformational changes from the aptamer it must be destabilized, such that ligand binding may rescue the stability of the RNA which is then conferred to the appended adapter RNA sequence to evoke a response. The destabilization of the aptamer necessary for these devices has inherent consequence, such as a loss of affinity and, as is implicit in the riboswitch studies, reduced kinetics. Compounding these effects are the limitations of the adapter domains themselves, where limited examples exist and structural misinformation hinder their development. While the modular riboswitch platforms are tempting as they can produce huge responses with established actuators such as GFP and β-Galactosidase, they are subject to the state of the host machinery, making them susceptible to toxically high ligand concentrations. Additionally, more work on the intracellular front is needed in order to establish their in vivo characteristics and engineering potential. The fluorescent RNA reporters, while nearly independent of host machinery, are limited by their signal and intracellular stability, confounding their interpretation and limiting reproducibility. Recent studies have suggested improvements in this area, but they have not yet been thoroughly validated. Consequently, it makes identification of the culprit for the failed in vivo experiments difficult, and without a clear fault a remedy becomes arduous.

The scaffolded selection provides a toe-hold for understanding new aptamers, eliminating the need for truncation studies and identification of a minimal motif. This streamlines the rational development of functional devices in the absence of tedious characterization as evidenced by the development of the metE switch. However, a high level of characterization
in the form of a crystal structure or structural probing is rapidly attainable and comprehensible in this context should it prove necessary, a process that traditionally takes years and parallel efforts from different groups. The idea still suffers from one limitation in the selection process, which is to say that often a single robust solution is attained. This relegates engineering strategies to two dimensions, screening of the adaptors and actuators, with only a single sensor available. With many of the intricacies of the final device specified by the sensor domain, such as affinity, specificity, dynamic range, and maximal output, the option of multiple sensor domains becomes very attractive. Various solutions or even a single robust solution hosted within the context of a variety of scaffolds, each with their unique properties, would produce a three dimensional engineering matrix, where the sensing domain along with the adaptor and actuator may be screened such that all facets of the RNA are tunable in order to produce a fit device. These ideas lead us to pursue multiple RNA scaffolds, each belonging to the global three-way junction fold, with a limited selection in order to produce a suite of aptamers that could be used to produce a robust 5htp device.
3.6 Materials and Methods

Library Construction
In order to construct a library that balanced preserved the characteristics of the structure we wished to preserve (i.e. robust folding, ion independence, etc.) an 8 Å shell of residues surrounding hypoxanthine in the pbuX-xpt guanine aptamer was selected in structure viewing software and randomized on ultramers synthesized by IDT (IDTDNA.com). The synthesized ssDNA was made transcriptionally competent using standard Taq PCR conditions in which 1 x 10^{12} sequences were used in each 100 μL PCR and amplified for 15 cycles. Approximately 1 x 10^{14} sequences were constructed and transcribed by loading dsDNA sequences into a 12.5 mL transcription reaction containing 40 mM Tris-HCl, pH 8.0, 25 mM DTT, 2 mM spermidine, 0.01% Triton X-100, 4 mM each rNTP pH 8.0, 0.08 units inorganic phosphatase (Sigma-Aldrich, lyophilized powder), and 0.25 mg/mL T7 RNA polymerase and incubated at 37 °C for 4 hours. Transcription samples were then precipitated in 75% ethanol, cooled to -20 °C, centrifugally pelleted, reconstituted in a solution of 300 μL formamide, 3mL 8 M Urea, and 300 μL 0.5 M EDTA pH 8.0, and separated on a denaturing 8% polyacrylamide gel. The corresponding band was excised from the gel after UV shadowing and eluted in 0.3M NaOAc pH 5.0 before exchange and storage in 0.5x TE.

Randomized Oligomer Sequences
xpt-pbuX purine scaffold (bold denotes first transcribed nucleotide; underlined is reinforced p1 stem):

<table>
<thead>
<tr>
<th>Library</th>
<th>5’-GCGCGCGGAATTCTAATACGACTCACTATAG- TAGGACTTTCCGGTCCAAGCTAATGCACTCCTGN- NNNNNCGCTGGATATGGCACGCTCN- NNNNNGGGCACCAGTAATGCTCCTCNNNNNGGG- TGCATTAGCAAAATCGGGCTTGGTTCGGTTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>T7 Site Appending Primer</td>
<td>5’-GCGCGCGGAATTCTAATACGACTCACTATAG- TAGGACTTTCCGGTCCAAGCTAATGCACTCCTGN- NNNNNGGGCACCAGTAATGCTCCTCNNNNNGGG- TGCATTAGCAAAATCGGGCTTGGTTCGGTTC</td>
</tr>
<tr>
<td>RT-PCR Primer</td>
<td>5’-GAACCGGACCCGAAGCCCG</td>
</tr>
</tbody>
</table>

Column Synthesis
For the derivatized columns, 3mL bed volume of EAH Sepharose 4B (GE) was dehydrated with DMF. 10 μmoles of fmoc-5-hydroxy-L-tryptophan or fmoc-pamino-phenylalanine and 10 μmoles of PyBOP were dissolved in 1 mL of DMF and added to the dehydrated column with 20 μmoles DIPEA and mixed end over end for 2 hours at room temperature. The column was then drained and washed extensively with DMF. Unreacted sepharose amines were acetylated by adding 1 mmole of acetic anhydride and 1 mmole DIPEA in approximately 1 mL DMF and mixing at room temperature for 1 hour. The column was drained of the acetylating mixture and washed with DMF prior to Fmoc deprotection using 20% v/v piperidine/DMF. Amino acid concentration on the column was determined by measuring the concentration of fmoc in the deprotection fractions (A_{301} ε = 8000). This method generated approximately 0.5-1 mM deprotected amino acid on the column. For counter selection, EAH sepharose was
prepared exactly the same except omitting the ligand coupling step resulting in acetylated sepharose.

**Selection Protocol**

R1:
350 μL of acetylated sepharose was equilibrated in selection buffer (10 mM HEPES pH 7.0; 250 mM NaCl; 50 mM KCl; 10 mM MgCl₂; 0.1 mg/mL tRNA) and 1 nmol of library RNA in 350 μL of selection buffer was incubated at RT for 30 min. The column was then drained and washed once with 350 μL of selection buffer. The flow through (750 μL total) was added to pre-equilibrated 5htp derivatized sepharose 4B column and incubated for 45 minutes. The column was then drained and washed three times with selection buffer before elution with 10mM 5htp in selection buffer (two 1 hour incubations in 350 μL; total 700 μL eluted volume). The eluted fractions were then concentrated to 50 μL in a 0.5 mL Ultracel 10kMWCO filter (Millipore) and ethanol precipitated in 0.3 M sodium acetate (pH 5.0), 5 μg glycogen, and brought to a final concentration of 75% ethanol before storage at -70 °C for 30 minutes. The solution was then pelleted at 13000 x g at 4 °C, decanted, and dried under vacuum. The dried pellet was then reconstituted with 0.7 mM each dNTP, 7 μM RT-PCR primer, and to a total volume of 14 μL before heating to 65 °C for 5 minutes and incubation on ice for 10 minutes. The solution was then brought up to 1x SuperScript III FS buffer conditions with 5 mM DTT and 200 units SuperScript III (Life Technologies) in a total volume of 20 μL before a 15 minute extension at 54 °C. The entire 20 μL reverse transcription solution was then PCR amplified in a total volume of 500 μL using Taq polymerase conditions. The amplified pool was then transcribed by adding 100 μL of the PCR reaction to a 1 mL transcription reaction containing 40 mM Tris-HCl, pH 8.0, 25 mM DTT, 2 mM spermidine, 0.01% Triton X-100, 4 mM each rNTP pH 8.0, 0.08 units inorganic phosphatase (Sigma-Aldrich, lyophilized powder), and 0.25 mg/mL T7 RNA polymerase and incubated at 37 °C for 2 hours. A 100 μL transcription reaction for 32P labeled RNA was performed under similar condition with the exception that the rNTPs were lowered to 2 mM for UTP, CTP, and GTP while ATP was reduced to 200 μM and 100 μCi of ATP-α-32P was supplemented. Transcription samples were gel purified as described in the library construction section with the gel loading conditions scaled accordingly.

<table>
<thead>
<tr>
<th>Round</th>
<th>[RNA] pmol</th>
<th>Washes</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>400</td>
<td>6</td>
<td>Counter selection against AcO-sepharose</td>
</tr>
<tr>
<td>3</td>
<td>400</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>10</td>
<td>30 second CS with 100 μM L-tryptophan</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>10</td>
<td>30 second CS with 100 μM L-tryptophan</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**Sequencing**

Initial sanger sequencing was performed by amplifying the final round with primers appending EcoRI and HindIII and cloning into the corresponding sites in pUC19. The plasmids were then transformed into DH10B(New England Biolabs), plated and grown at 37 °C overnight. Single colonies were picked, grown up in liquid culture, plasmid DNA extracted, and sent for sequencing with M13 reverse primer (SeqWright or Genewiz).
NGS protocol
Standard PCR was conducted with the following primers before sequencing on a MiSeq (Illumina)
Forward primer for hybridization to Illumina chip
1: Illumina adapter 2: Primer pad 3: T7 promoter/library compliment
5'-AATGATACGGCGACCACCGAGATCTACACTATGGTAATTGTGC/GCG/CGAATTCTAATACGACTC/ACTATAG
Reverse primers containing barcodes used to append necessary sequence. Barcodes courtesy of Rob Knight’s group. 1: Illumina adapter 2: Golay barcode 3: Primer pad 4: library compliment

<table>
<thead>
<tr>
<th>Round</th>
<th>Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATTCCCTTGTCTCCAGTCAG-TGCCGAACCGGACCGAGAAGC/CG</td>
</tr>
<tr>
<td>2</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATACGAGACTGATTAGTCAG-TCA/GCCGAACCGGACCGAAGCC/CG</td>
</tr>
<tr>
<td>3</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATGCTGTACGGGATTAGTCAG-TCA/GCCGAACCGGACCGAAGCC/CG</td>
</tr>
<tr>
<td>4</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATAC/GCA/GTACGGATACACCAGGTACGGACTAGTCAG-TCA/GCCGAACCGGACCGAAGCC/CG</td>
</tr>
<tr>
<td>5</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATTTGGTCAACCGATAAGTCAG-TCA/GCCGAACCGGACCGAAGCC/CG</td>
</tr>
<tr>
<td>6</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATATCGCAGAC/GTATACGGGACCGAAGCC/CG</td>
</tr>
<tr>
<td>7</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATGTCCTAGCTAAGTCAG-TCA/GCCGAACCGGACCGAAGCC/CG</td>
</tr>
</tbody>
</table>

Forward sequencing primer:
5'-TATGGTAATTGTGC/GCG/CGAATTCTAATACGACTC/ACTATAG
Reverse sequencing primer:
5'-AGTCAGTCAGCCGAACCGGACCGAAGCC/CG
Indexing primer:
5'-CGGGCTTCCGGTCCGGTCTCGTACTGACT

Chemical Probing
RNA was prepared as described previously [168]. Structure cassettes flanking the 5’ and 3’ ends of the RNA were added to facilitate reverse transcription and NMIA modification was performed using the established protocols [159] at 25 °C. RNA was probed at 100 nM in 100 mM Na-HEPES [pH 8.0], 100 mM NaCl, and 6 mM MgCl$_2$. Ligand concentration was 500 µM where indicated. Gel images were analyzed by SAFA [173] and ImageJ (NIH).

Isothermal Titration Calorimetry
All RNAs tested were exchanged into the selection buffer (10 mM HEPES pH 7.0; 250 mM NaCl; 50 mM KCl; 10 mM MgCl$_2$) and washed three times in 10k MWCO filter (EMD Millipore). The ligand was brought up from a dry solid directly into the binding buffer and
concentration established on a NanoDrop 2000 (Thermo Scientific: $\varepsilon_{275} = 8000 \text{ mol}^{-1} \text{ cm}^{-1}$ for 5-hydroxy indol moiety). The RNA was diluted to between 50-100 $\mu$M and the ligand was diluted to 10 fold that of the RNA. Titrations were performed at 25 °C using an iTC-200 micro calorimeter (Microcal, Inc.; GE). Data was analyzed and fitting was performed with the Origin 5.0 software suite (Origin Laboratories).

**Crystallographic data collection and processing**

The crystal construct DNA sequence that contained a 6bp P1 stem:

1: **T7 promoter**
2: **P1**

5'-GCGCGCCGAATTCTAAATACGACTCTACCTATAAGACACACTCTGTGATGATCGCGT-GGATATGGCAGCGATTGTTGGACACCGTAAATGTCTCTAACAACGTTGACAA

Additional P1 lengths tested:

<table>
<thead>
<tr>
<th>P1 Length</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5'-GGAUC-Aptamer-GGUCC</td>
</tr>
<tr>
<td>7</td>
<td>5'-GGACAC-Aptamer-GGUGUCC</td>
</tr>
<tr>
<td>8</td>
<td>5'-GGACACAC-Aptamer-GUGUGUCC</td>
</tr>
</tbody>
</table>

The sequences were PCR amplified, *in vitro* transcribed, and purified on an 8% denaturing acrylamide gel as previously described [168]. The RNA was then concentrated in an Amicon Ultra 15 10k MWCO filter (EMD Millipore, Inc.) and exchanged into 0.5x Tris-EDTA. The standardized hanging-drop method was used for crystallization with various buffers from the Nucleic acid mini and Natrix screens (Hampton Research). Trays were set up using protocols established by the its with 2 mM 5htp and 400 $\mu$M RNA. The conditions of the Nucleic acid mini screen sample 3 were further optimized for the construct containing a P1 stem of 6bp yielding the best crystals in 8-14% MPD, 40 mM sodium cacodylate pH 5.5, 4 mM MgCl$_2$, 12 mM NaCl, 80 mM KCl, and 4-9 mM cobalt hexamine for the mother liquor. Diffraction quality crystals were obtained by mixing 2 $\mu$L RNA:Ligand (1:1) and 3.5 $\mu$L mother liquor, micro-seeding, and incubation at 22 °C for 1-3 days. The crystals needed no further cryoprotection, were isolated in .2-.3 $\mu$m nylon loops, and flash frozen before data collection.

Data was collected using the CuK$_\alpha$ wavelength (1.5418 Å) radiation on an R-AXIS IV++ home source (Riguaku MSC), and the data was indexed and scaled using D*TREK. The crystals were also readily formed when the cobalt hexamine was replaced with 1-11 mM iridium hexamine allowing for the collection of anomalous data on the home source. Phases were determined by single isomorphous replacement with anomalous scattering (SIRAS). AutoSol [170] was used to find 12 iridium atoms that were then used to calculate phases. The resulting experimental density map displayed unambiguous features of the RNA backbone and helixes and was used for building the model.

The initial model was iteratively built without the ligand in Coot [171] between rounds of refinement in PHENIX [170]. The RNA model was brought through several rounds of refinement and simulated annealing before 5htp was built into the model. At this point of building, there was clear ligand density in the binding pocket that allowed for the sure placement and orientation of the ligand. The placement of the ligand and bases was validated by a composite omit map, a harsh validity test of the model that still retains density for the ligand. Water placement was automated in final rounds of refinement after ligand placement based on peak size in the F$_O$-F$_C$ differential map.
**In vitro transcription assays**

For brevity, only the sequences that show some level of transcriptional attenuation are provided.

<table>
<thead>
<tr>
<th>xB$_{17}$/metE</th>
<th>TTATCAAAAAAGATTTAGCTTTAAGTCTAACC-TATAGGATCTTACACGGCCAAAAAATTAATAACATTTTTCTTTAATGATGACGGATGATGACGCCACGCA-TGAATTTGGGAAACCGTAAATGCTTAC-ACGTTAGAGAGACGACTTACGTAGAAAAGCCTC-TTTCTCTCATGGGAAAGAGGCTTTTTGGTTGTTGAG-AAACCTCTTACGAGCTTGAACGCTGCGGGTGAGAAGAGAGTTTTTACATATAAACAGAGAAAAGAATG</th>
</tr>
</thead>
<tbody>
<tr>
<td>xB$<em>{17}$/o$</em>{pbu}$</td>
<td>AATATTTGAAATTCTTTATCAAAAAAGAGATTTAGCTTTAAGTCTAACC-TATAGGATCTTACACGGCCAAAAAATTAATAACATTTTTCTTTAATGATGACGGATGATGACGCCACGCA-TGAATTTGGGAAACCGTAAATGCTTAC-ACGTTAGAGAGACGACTTACGTAGAAAAGCCTC-TTTCTCTCATGGGAAAGAGGCTTTTTGGTTGTTGAG-AAACCTCTTACGAGCTTGAACGCTGCGGGTGAGAAGAGAGTTTTTACATATAAACAGAGAAAAGAATG</td>
</tr>
<tr>
<td>xB$<em>{17}$/o$</em>{1}$/pbu</td>
<td>AATATTTGAAATTCTTTATCAAAAAAGAGATTTAGCTTTAAGTCTAACC-TATAGGATCTTACACGGCCAAAAAATTAATAACATTTTTCTTTAATGATGACGGATGATGACGCCACGCA-TGAATTTGGGAAACCGTAAATGCTTAC-ACGTTAGAGAGACGACTTACGTAGAAAAGCCTC-TTTCTCTCATGGGAAAGAGGCTTTTTGGTTGTTGAG-AAACCTCTTACGAGCTTGAACGCTGCGGGTGAGAAGAGAGTTTTTACATATAAACAGAGAAAAGAATG</td>
</tr>
<tr>
<td>xB$<em>{17}$/o$</em>{2}$/pbu</td>
<td>AATATTTGAAATTCTTTATCAAAAAAGAGATTTAGCTTTAAGTCTAACC-TATAGGATCTTACACGGCCAAAAAATTAATAACATTTTTCTTTAATGATGACGGATGATGACGCCACGCA-TGAATTTGGGAAACCGTAAATGCTTAC-ACGTTAGAGAGACGACTTACGTAGAAAAGCCTC-TTTCTCTCATGGGAAAGAGGCTTTTTGGTTGTTGAG-AAACCTCTTACGAGCTTGAACGCTGCGGGTGAGAAGAGAGTTTTTACATATAAACAGAGAAAAGAATG</td>
</tr>
</tbody>
</table>

dsDNA templates from above were transcribed as previously described [82]. In brief, 50 ng of DNA template were incubated at 37 °C for 10 minutes in 12.5 μL of 2x transcription buffer (140 mM Tris-HCl, pH 8.0; 140 mM NaCl; 0.2 mM EDTA; 28 mM β-mercaptoethanol and 70 mg/mL BSA), 2.5 μL 50 mM MgCl$_2$, 100-200 μCi of ATP-$α$-$^{32}$P, and 0.25 units of *E. coli* RNA polymerase σ70 holoenzyme (Epicentre Biotechnologies) per reaction were brought
to 23 μL. The equilibrated reactions were then initiated with the addition of 7.5 μ reaction buffer (165 μM each rNTP, 0.2 mg/mL heparin, and the desired ligand concentration) and incubated for 15 minutes at 37 °C before quenching in Urea. The reactions were then separated on an 8% denaturing PAGE, dried and exposed on a phosphor imager screen. Quantitation of the radioactive counts were then carried out in ImageJ (NIH) and the data fit to a two-state model where applicable.

**In vivo riboswitch assays**

The assay used established vectors and protocols previously published reports except where indicated [79,84]. Briefly, the pBR322 vector was engineered to host the GFPuv construct under control of the proD promoter (pRR5) and chimeric riboswitches were inserted between the promoter and actuator. The sequence for all constructs tested in vitro were the same as used in in vivo. The resultant vectors were transformed into Keio parents cells (strain BW25113) with single colonies picked and grown overnight in the respective media with 100 μg ampicillin. The culture was then used to inoculate media containing a desired amount of 5htp at a 1:1000 ratio, and grown to early log phase (OD$_{600}$ = .2-.8). 300 μL of the culture was then used to take an OD$_{600}$ and fluorescence reading in a plate reader (Tecan). The fluorescence was normalized to the cell density to obtain the histograms shown.

Many variations were taken upon this approach in an attempt to overcome the effects of the ligand. The list is very long, including two types of media (CSB and M9), temperature (20, 30, and 37), amino acid source (CAS [0.01, 0.025, 0.05, 0.1, 0.2 %] and defined amino acids [50 μM each]) with and without tryptophan (1-500 μM screened). Additionally several different actuators were used, including GFPuv, mCherry, and β-Galactosidase with the Miller assay giving the most consistent and reproducible results. Many of these were grid screened against each other, and while they could rescue certain traits of the cells they were never robust sensors of 5htp media concentrations.

**Malachite green assays**

**T7 Promoter**

<table>
<thead>
<tr>
<th>xB$<em>{17}$:MGA$</em>{A}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'-GCGCGCGGAATTCTAATACGACTCACTATAGGATCCCGAC-TGATGATCGCGTGGATATGGCACCCGT-AAATGTCCCTAACGAAGCTCCTAGGATGGA</td>
</tr>
<tr>
<td>xB$<em>{17}$:MGA$</em>{B}$</td>
</tr>
<tr>
<td>5'-GCGCGCGGAATTCTAATACGACTCACTATAG-GAGAGGTAACCAATGTGGATCGCGTGGATATGGCACCACCTAG- AATTGTTGGACCCGTAAGTGTCCCTAACAAACCGACTC</td>
</tr>
</tbody>
</table>

The assay buffer for the malachite green experiments was 20 mM Tris-HCl, pH=7.4, 140 mM KCl, and 5mM NaCl with variable magnesium concentrations. The concentrations of RNA and malachite green were optimized at 1μM and 10 μM respectively, values similar to that previously established [113]. The RNA sensor, malachite green, MgCl$_2$, and desired concentration of ligand were combined and brought to 100 μL in 1x binding buffer and allowed to equilibrate for 15 minutes at room temperature. The reactions were then assayed on a plate reader with excitation at 615 nm and collecting and averaging emission from
650-660 nm. The resulting fluorescence data was plotted against ligand concentration and fit to a two state binding mode to achieve the observed \( K_D \).

**Spinach and minSpinach constructs**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5'-GACGCGACTGAATGAAATGGTGAAAGGACGCGGTCCATGATGATCGCGTGATATGGCGACGCATTGAATTGTCGTC-ACAAATTGTTGAGTAGAGTGTAGCTCCGTAACCTAGTCGCGTC</td>
</tr>
<tr>
<td>B</td>
<td>5'-GACGCGACTGAATGAAATGGTGAAAGGACGCGGTCCACTGATGATCGCGTGATATGGCGACGCATTGAATTGTCGTC-ACACCTTTGAGTAGAGTGTAGCTCCGTAACCTAGTCGCGTC</td>
</tr>
<tr>
<td>C</td>
<td>5'-GACGCGACTGAATGAAATGGTGAAAGGACGCGGTCCACTGATGATCGCGTGATATGGCGACGCATTGAATTGTCGTC-ACATTTGTTGAGTAGAGTGTAGCTCCGTAACCTAGTCGCGTC</td>
</tr>
<tr>
<td>D</td>
<td>5'-GACGCGACTGAATGAAATGGTGAAAGGACGCGGTCCATGATGATCGCGTGATATGGCGACGCATTGAATTGTCGTC-ACCTTTGTTGAGTAGAGTGTAGCTCCGTAACCTAGTCGCGTC</td>
</tr>
<tr>
<td>E</td>
<td>5'-GACGCGACTGAATGAAATGGTGAAAGGACGCGGTCCACTGATGATCGCGTGATATGGCGACGCATTGAATTGTCGTC-ACATTGTTGAGTAGAGTGTAGCTCCGTAACCTAGTCGCGTC</td>
</tr>
<tr>
<td>F</td>
<td>5'-GACGCGACTGAATGAAATGGTGAAAGGACGCGGTCCACTGATGATCGCGTGATATGGCGACGCATTGAATTGTCGTC-ACATTTGTTGAGTAGAGTGTAGCTCCGTAACCTAGTCGCGTC</td>
</tr>
</tbody>
</table>

minSpinach
<table>
<thead>
<tr>
<th></th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>GGACGAAGGACGGGTCCGGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>A2</td>
<td>GGACGAAGGACGGGTCCCTGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>A3</td>
<td>GGACGAAGGACGGGTCCGGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>B1</td>
<td>GGACGTTGAGTAGAGTGTGAGCTGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>B2</td>
<td>GGACGTTGAGTAGAGTGTGAGCTGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>B3</td>
<td>GGACGTTGAGTAGAGTGTGAGCTGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>A4</td>
<td>GGACGAAGGACGGGTCCCTTTGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>A5</td>
<td>GGACGAAGGACGGGTCCATATGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>A6</td>
<td>GGACGAAGGACGGGTCCATATGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>A7</td>
<td>GGACGAAGGACGGGTCCATATGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>A8</td>
<td>GGACGAAGGACGGGTCCATATGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>minSpin_A</td>
<td>GGACGAAGGACGGGTCCCTTTGATGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
<tr>
<td>minSpin_B</td>
<td>GGACGTTGAGTAGAGTGTGAGCTGATCGGCTGGATATGGCAACGTATTGTTGGACACCCTAATGCCTAAACATG-TTGAGTAGAGTGTGAGCGTCC</td>
</tr>
</tbody>
</table>

The spinach constructs were assayed under the same buffer conditions as described above for the malachite green system under 5 mM MgCl$_2$ and an excitation of 460 nm and collected and averaged emission from 503 to 511. The constructs often required equilibration for greater than 30 minutes or a heat denaturation and refolding in the presence of ligand in order to reach maximum fluorescence. Even still the greatest fluorescence was had when roughly equimolar amounts of raw transcription was added in place of the purified RNA, suggesting that the constructs were ill behaved.
Chapter 4

Various Three-Way Junction Architectures as Host Scaffolds for 5-hydroxytryptophan

4.1 A need for screenable sensor domains

The selection of novel aptamers using limited libraries hosted in the purine scaffold proved successful, resulting in the development of a robust 5htp aptamer. However, even though the aptamer is highly structured and possesses characteristics well associated with functional sensory domains, its implementation in vivo proved difficult. The inability to function in vivo is frustrating given that the devices can work in their in vitro validations. The culprit is likely multi-faceted, produced by an incomplete understanding of the adapter sequences in vivo, poor signal-to-noise, and ligand effects on the cell. The ability to improve upon these systems is limited with the sensor domain being a fixed, sole part forcing the adaptor platforms to adapt to the sensor. SELEX has often been viewed as a technique capable of producing competent sensory domains for use in various biotechnical applications, therefore most of the efforts into RNA biosensor development have focused upon the adaptor or actuator platforms. Despite the supposed ease of aptamer development, very few synthetic aptamers have functionality in vivo, likely because so few meet the criteria of being bio-orthogonal, well understood, and dynamically capable of transducing the binding event. Furthermore, our study suggests that even engineering well folding architectural elements
into the selection is not sufficient to produce a functional *in vivo* RNA device.

From an engineering standpoint it would be beneficial to screen all RNA parts (actuators, adaptors, and sensors) to expedite the development of an RNA device. This is currently limited by selection techniques, where often a single high affinity solution is obtained from the process and diversity from the selection is had at the expense of affinity or at the cost of time. Two remedies exist to the limited solution space: either a selection based on function is needed which would preclude the need for tedious screening, or various parallel selections would be needed to produce a diverse set of aptamers such that a more irrational and un-biased approach could easily yield many functional device candidates. While a functional selection is attractive, it is limited in practicality. The isolation of high-affinity aptamers *in vivo* is constrained by the transformation efficiency of $\sim 10^9$ in *E. coli* and much less in higher organisms such as yeast [211]. This is problematic when the frequency of observing an RNA aptamer in sequence space ranges from $\sim 1$ in $10^9$ for very low affinity aptamers to 1 in $10^{20}$ for very high affinity aptamers without the added pressures of functionality [212]. Traditional selection offers a much greater sampling of sequence space, often $> 10^{15}$ sequences. Unfortunately, traditional selection produces the simplest structural solution to a given problem, i.e. if a selection is carried out multiple times, a highly similar result will be produced even if more complex solutions exist and are sampled [212].

With that in mind, the purine selection that yielded xB$_{17}$ demonstrates one of two possibilities: either that the selection was not limited in sequence space (xB$_{17}$ is the best solution) or that localizing to a specific region of sequence space does not adversely effect the final result and many regions of sequence space, even when restricted from sampling the more traditional global minimum, can produce very competent aptamers (xB$_{17}$ is one of many good solutions not sampled in traditional selections). Even with enhanced rates
of mutagenesis during the selection, the scaffold is certainly limited in sequence space with >60% fixed. Hence the idea that we have sampled and retained the best solution is most certainly false. It is likely, however, that xB_{17} is one of many very good solutions that can be sampled at distant regions of sequences space. The practicality of this being that if these distant competent solutions can be sampled, they would provide the aptamer diversity that alleviates a pressure on regulatory RNA development while retaining high affinity and specificity. Multiple competent scaffolds could produce just this effect, where each localizes to an area of sequence space and produces a functionally similar, but contextually different aptamer. This lead us to introduce limited libraries into additional scaffolds to survey their ability to produce high affinity and specific aptamers to aid in the development of RNA devices.

4.2 A new reverse transcriptase for library amplification

During the previous lysine and purine selections, several different types of errors were accumulated throughout the selection. The first observation was in the lysine selections, where two separate sequence truncations of 8 and 13 nucleotides are observed that include or immediately succeed poly-adenosine stretches. This event occurs with single nucleotide fidelity, repeatedly deleting the same sequence in multiple templates as evidence by heterogeneous variable regions. The source of these errors could be at any of three levels; the transcription, the PCR, or the reverse transcription. We focused on the reverse transcription step, as this form of mutation strongly resembles template switching, a process inherent to retroviral reverse transcriptases (RTs). The process of template switching is usually observed near the 5’ end of direct sequence repeats, where the nascent DNA dissociates from the RNA template and re-anneals at a separate region of high homology to continue the synthesis of
cDNA [213]. The lysine template positions two stretches of adenosines in close proximity, displaying an ideal site for the virally derived RT used in these studies to switch templates. The frequency of template switching for a single cycle of viral replication can be as high as 0.1-1% in the native context of viral RTs, a function often critical to the survival and rapid adaptation of the virus [214–216]. While the high rates of recombination and template switching are beneficial to viral fitness, if the process is occurring during the selection it would be a severe hinderance to the maintenance of the complex RNA architecture.

In addition to the truncation events that perturbed the structure in the lysine riboswitch selections, the purine scaffolded selections acquired more non-specific point mutations and deletions (Figure 4.1). The deep sequencing analysis clearly shows that the point mutations and deletions were localized to areas critical for structure and to a poly-adenosine stretch on the 3' end of the RNA. At several positions the error rate exceeds 90% suggesting that this was another recurrent process that affects all sequences and not the effect of a single occurrence that was propagated. While the preciseness of mutation that occurred in the lysine selection is not present in the purine scaffold, the toll of mutational accumulation is severe with roughly 70% of the sequences containing errors that perturb the structure. While the source of the mutations can be at any step of the amplification cycle, the RT is again not ideal, with the error rate for Superscript III at 55 °C of 4.5 x 10^{-5} [217]. While this level of mutagenesis is not an issue in single cycle experiments, iterative rounds of amplification that require reverse transcription can result in a large amount of accumulated mutation. It’s interesting to note that this error is not readily observable in traditional selections, where mutations in the priming region are selected against during PCR and those occurring within the library are indistinguishable from the normal sequence variation.

The selective technique we wish to employ requires the preservation of structured RNA
Figure 4.1: A new group II intron derived RT to remedy the crippling effects of accumulated error. (a) The error accumulated during the purine selection approaches 90% for positions on the 3′ side of P1. These mutations weaken the amplification cassettes stem, presumably to ease the replicative pressures. (b) Comparison of the thermal stability and processivity of SuperScript III with two RTs derived from group II introns (TelI4c-MRF and GsI-IIC-MRF). GsI-IIC-MRF, the RT used in these studies, maintains activity at temperatures greater than 65 °C (top). Additionally, GsI-IIC-MRF is more processive through highly structured RNAs such as the group II intron RNA (bottom) showing a much lower rate of early truncated products and greater overall yield. (c) Mutation frequency was assessed with an in vivo lacZ assay where errors in the RT process leads to β-Galactosidase deficiency and white colonies. GsI-IIC-MRF shows a mutational frequency and error rate of less than half that for SuperScript III. These comparisons make it clear that GsI-IIC-MRF is a more capable and efficient RT for selection, especially when maintenance of structure is critical. Panels (b) and (c) taken from [218].
and is hindered by the inherently low fidelity of viral RTs. Therefore, we sought an alternative RT that would retain the processivity of the highly engineered viral Superscript III, but improve on its fidelity. The ability to carry out the reverse transcription reaction at elevated temperature is advantageous as higher-order RNA structure that would normally impede the RT is denatured or weakened and the specificity can be improved by the discrimination of mispaired primers. Only a few RTs capable of functioning at high temperatures are available, and many have relatively high error rates such as Superscript III [218]. However, several non-viral retroelements exist such as non-LTR-retrotransposons and mobile group II introns. Whereas infectious viruses must evade host responses and thus benefit from RTs with high error rates and recombination that introduces and propagates variation, these alternative retroelements require the accurate cDNA synthesis of long and structured RNA sequences for their fitness. Additionally, the mobile group II introns are found in thermophilic bacteria providing access to naturally derived RTs capable of functioning at high temperatures.

Following this logic, Mohr et al. identified and optimized the RT domain of a group II intron from the thermophile Geobacillus stearothermophilus (GsI-IIC-MRF) for its use in biotechnology [218, 219]. In its natural context, the RT must efficiently synthesize an accurate cDNA copy of intronic RNA that is >2-kb long and highly structured. This RT retains activity up to 70 °C with no observable loss in full length product up to 65 °C (Figure 4.1). As compared to Superscript III, GsI-IIC-MRF has a much higher specific activity (1300 vs. 220 units/μg), a reduced error rate (0.64 x 10^{-5} vs. 1.5 x 10^{-5}), and equal or greater processivity. These attributes make the thermostable GsI-IIC-MRF an improved RT for the amplification of our stable and highly structured libraries.
4.3 Additional scaffolds for RNA sensor diversity

In order to generate multiple solutions in parallel selections, additional scaffolds that are well characterized, folded, and structurally understood were sampled from the greater three-way junction fold. A class I cyclic di-guanosine monophosphate (c-di-GMP) riboswitch aptamer termed Vc2 is found upstream of the tfoX-like gene in *Vibrio cholerae* and was chosen on the basis that it is a capable sensor in the spinach system *in vivo* with the ability to adapt its specificity towards the orthogonal compounds cyclic AMP-GMP and cyclic di-AMP with limited mutational input [119, 220]. The structure shows that the cognate c-di-GMP ligand binds in the junction of the three helixes independent of direct interactions with P2 and P3. The guanine groups of c-di-GMP serve to mediate the stacking between P1 and P2 with the L2 loop supported by an interaction with the minor groove of P3, forming the distal interaction of the canonical three-way junction fold. The third representative, a type-III fast hammerhead ribozyme from *Schistosoma mansoni*, was selected as it is a model catalytic RNA system and has seen previous success as an RNA adaptor in synthetic biology [94, 221]. This fast hammerhead RNA is easily permutated allowing for the terminal ends to terminate in either stem I, III, or both in a bifurcated system. The RNA used in this selection is a closed type-III RNA, where the terminal ends are located in stem III forming P1. In this orientation, terminal loop of stem II (now P3) docks into stem III (P2) forming the global three way junction fold. The Vc2, fast hammerhead, and purine RNAs are each a unique solution to the three-way junction fold, possessing variable stem lengths and different peripheral interactions to stabilize the junction region which allows for a broad survey of structural space within the context of a pervasive scaffold. The limited libraries for fast hammerhead (Fhh) and Vc2 (cdG) were constructed as previously with the purine scaffold,
resulting in a 21 nucleotide library for each RNA localized to the junction of the three helixes with care taken to preserve the higher order structure (Figure 4.2). This rational again allows for the complete survey of sequence potential as it is less than the limit imposed by the solid phase synthesis of DNA.

In addition to surveying new scaffolds and using a more thermostable RT, the magnesium concentrations and platform sequence were adjusted in order to further refine the technique. The P1 stem of the selection cassette used previously was fairly thermostable, with a predict $T_M = 55 \, ^\circ C$ at 10 mM MgCl$_2$ and 300 mM monovalent ions. In an attempt to alleviate some of the amplification pressures of reverse transcription through a long and stable RNA structure, the P1 stem was weakened by altering the composition of basepairs while attempting to maintain the length. The new P1 stem has a predicted $T_M = 44 \, ^\circ C$ and is only a single basepair shorter than previously, sufficiently distancing the constant regions from the library. Additionally, at the base of the P1 stem on the 3’ side near the priming region for the reverse transcription reaction was a stretch of four consecutive adenosines. This stretch was the site of heavy mutagenesis in both the lysine and purine scaffolded selections, and while possibly not a problem with the new non-viral GsI-IIC-MRF RT, the linker residues were mutated to cytidine to assure no further issues with this region.

The magnesium concentration was also lowered from 10 mM to 3 mM in the selection. The simplest reason for this is to obtain aptamers that bind in conditions which more closely resemble the ionic conditions found in the cell. It has also been shown that higher affinity aptamers have a lesser dependence upon magnesium concentrations than lower affinity aptamers [156]. Many studies have noted that levels of magnesium above that present intracellularly are needed for high levels of RNA function [222-224], yet more structural information leading to higher levels of RNA complexity are able to compensate for the lack
of magnesium [156,212]. Since the peripheral motifs of the RNA scaffold were selected to function in an intracellular environment they should convey the necessary information to the limited library necessary for the production of high affinity aptamers without applying addition pressures to the selection. While a greater degree of structure, such as that provided by the scaffold, may obviate the need for magnesium it does not imply a more rigid and less capable aptamer as evidenced by the high degree of complexity witnessed in natural RNA elements such as riboswitches and ribozymes.

Seven rounds of selection against a 5htp derivatized sepharose column were carried out with a reduced washing stringency as compared to before. This allowed for the collection of nearly 10% of the total RNA pool through the first four rounds of selection with the intention of preserving diversity and reducing the pressure on the RT step by allowing for ample RNA in the reaction. The side effect of reducing the washing stringency in early rounds is that pool enrichment is hampered, with significant enrichment over background column retention not seen until round 5 of the selection when additional washing steps were added (Figure 4.2d). Counter selections against L-tryptophan were introduced in rounds six and seven with minimal effects on the total levels of RNA eluted from the column. The selection was halted at the end of the seventh round, where 15-17% of the total RNA could be competitively eluted. By design, this was much less total pool elution than previously as our sequencing results show that once enrichment is observed, the final solutions are present and subsequent rounds and pressures only serve to alter the relative distribution of the RNA pool. Initial Sanger sequencing after the cloning of individual sequences showed only a small amount of convergence for the sample sequences and that the length and composition of the template was well maintained, warranting the use of next generation sequencing for pool analysis.

Deep sequencing clearly shows that use of the new group II intronic RT and changes
Figure 4.2: Continued on the following page.
Figure 4.2: (From previous page) The design, cassette optimization, and selection of new scaffolds. (a) The *S. mansoni* fast hammerhead and Vc2 c-di-GMP riboswitch structures used to design the limited 21 nucleotide libraries (blue and red respectively). The C17 residue at which cleavage occurs in the fast hammerhead is shown in green and the c-di-GMP ligand for the Vc2 RNA is shown in teal. (b) Simplified secondary structures with the residues selected in the crystal structure correspondingly highlighted. (c) Each template contained a unique stem that linked the amplification cassette and the library. This served as a barcode with each scaffold amplified using a forward primer that could only efficiently amplify its corresponding template. The P1 stem was optimized to reduce the $T_M$ of the helix allowing for easier read through of the RT. Additional mutations are shown in red and serve to eliminate poly-adenosine stretches. (d) The percent RNA eluted and amplified per round of selection. From 5 to 15% of total RNA was collected in the first 4 rounds of selection in order to alleviate pressures on the amplification process and to preserve diversity. At round 5 the washing stringency was increase resulting in a much smaller elution percentage. The pool recovered quickly however, returning to greater than 10% competitive elution in the next round with the additional pressure of counterselection.

to the amplification platform reduced the overall error rates. Each template was self-aligned and the per residue error calculated and represented in Figure 4.3 as a histogram. All three templates showed a reduced rate of mutational accumulation through constant regions of the RNA, with the largest amount occurring at the base of P1 on the 3’ end. By this analysis, the rate of error accumulation reduces as the RT moves along the template, with rates at the 5’ end approaching 0%. While an exact calculation of the basal rate of error is complicated as it is a dynamic process that is dependent on distance from the 3’ primer and sequence composition, it ranges from less than 0.5% through the constant regions of the Fhh selection to 1-3% in the cdG and purine selections. On average this is 4-5 fold lower levels of accumulated error through seven rounds of selection than when using the SuperScript III RT.

The sequenced templates were then clustered and parsed by the bioinformatic pipeline as described and detailed in Appendix A revealing many well formed and convergent clades
Figure 4.3: The degree of accumulated error is reduced in the new selections. The original xpt selection is included for reference (top) as compared to the three new selections (bottom). The accumulation of point mutations and deletions has been minimized by adjusting the amplification cassette, the conditions, and using a more thermostable and processive RT. While there still exists a high degree of mutation at the 3’ side of P1, especially in the purine selection, the frequency of observed mutations is low, with no focused areas of intense mutagenesis as observed previously. The rates of mutagenesis through the structured regions necessary for preservation of the global fold appear to be at background levels for all templates when using the GsI-IIC RT.
within each pool. The process of identifying sequences for downstream applications in this context is different than in traditional selections as the pool was not over-selected to produce only a few, easily distinguishable aptamers. Instead, a high degree of diversity was intended and deep sequencing with clustering into distinct families used to isolate potential sequences. The cdG selection shows the lowest degree of convergence and a maximum likelihood tree does not clearly distinguish between clades and therefore potential solutions are not easily definable. Additionally, nearly half of the clusters resulting from the analysis do not appear to fold well, likely due to the increased complexity of cdG and maintenance of the invariant sequence which makes the identification of paired regions difficult in the absence of covariation. However, the output is still easy to triage, with cluster 17 comprising nearly 8% of the final sequence pool and predicted to fold into a structure similar to that of the parental riboswitch. One sequence clearly stood out as the most abundant in this cluster and was therefore chosen for further validation (cdG$_{17}$). The Fhh and purine selections had begun to converge to a higher degree than that of the cdG, with several easily distinguishable clades emerging in the pairwise identity trees (Figure 4.4). While the solutions may be more isolated, the process of selecting sequences for validation was identical to that in the cdG selection, with cluster 4 in the Fhh and cluster 1 in the purine being very abundant and predicted to fold into parental like structures.

Cluster 4 of the Fhh library comprises 30.3% of the final pool of RNA and folds into a predicted secondary structure that is highly similar to that of the parental fast hammerhead construct as supported by covariation in the helical regions (Figure 4.5b). The conservation model at the base of P2 appears degenerate, as no single nucleotide is present at greater than 75% frequency even though the position is well occupied. The dominant single sequence for the cluster (Fhh$_{4}$) is twice as abundant as any other, and it gives a clue as to the nature of
Figure 4.4: The three selections show variable degrees of intra-relatedness. (a) The cdG scaffold shows relatively little divergence and the clusters are not that distant in sequence space. (b) The second iteration of the purine scaffolded selection produced many distinct and isolated solutions with sequence clusters easily identifiable by eye. Interestingly, highly similar sequences to the xB₁₇ solution are observed at a low frequency. It should also be noted that even though cluster 1 and xB likely recognize the ligand with the same UUGAA motif, they are located at distant ends of sequence space. (c) The Fhh cluster also contains several highly related clusters. The sequence in gray is highly divergent and likely background sequences that have persisted throughout the selection. These sequences are not part of any cluster and would not be observed at this frequency in an over selected pool, suggesting that we are observing a high level of diversity at this stage of the selection.
the degeneracy in J2/3. The first five nucleotides present in the junction are UUGAA, the same composition as the T-loop found in J2/3 of xB17 in an orthogonal placement. This is intriguing as it raises the possibility that the Fhh three-way junction has produced the same functional solution as another architecture, suggesting that the 5htp binding motif is in fact transposable into different contexts that may confer a greater degree of functionality. What’s interesting about the degeneracy in the covariation model, which is computed with de-replicated sequences such that a single abundant sequence does not overly inform the model, is that when the nucleotide distribution is broken down by percent identity at each position, it comes into agreement with the natural phylogeny of the T-loop motif (Figure 4.5c). There are discrepancies between the two motifs, most notably in position 1 and 2, likely because one is host to a nucleobase and the other, possibly, a planar amino acid. Position 5 of the proposed Fhh T-loop is essentially fixed on an adenosine which, in xB17, is critical to the recognition of 5htp as it identifies the main chain atoms of the amino acid and coaxially stacks with the ligand. This result demonstrates the power of deep sequencing and preservation of the diversity in the final pool, which has allowed for the likely identification of the 5htp T-loop receptor motif under its functional nucleotide diversity, which may aid in the down stream application of the sequence.

Cluster 1 of purine re-selection comprises 24% of the final pool and also contains the UUGAA subsequence found in the original purine selection and Fhh4. Comparison to the xB17 from the previous selection shows that the motif is offset by a single nucleotide while J1/2 and J3/1 no longer identifiably contain the nucleotides responsible for forming the complex binding pocket interactions such as G23 or C75. However, this is not to suggest that the RNA binding pocket is more disordered, more likely is that the library accommodated and structured the T-loop in a different context. The 5’ side of J1/2 is expected to pair
Figure 4.5: The clusters of candidate sequences and sequence analysis. (a) Covariation model of cluster 17 of the cdG selection. The pairwise tree showed no clear identification of convergence or that the pool had settled on a solution. The clustering and analysis make identification of candidate clusters trivial, with cluster 17 showing both a high abundance and conserved sequence elements. (b) Cluster 4 of the Fhh selection shows strong covariation and accepted mutations in a predict and slightly altered P2 stem. (c) J2/3 of the Fhh selection appears to show no strong convergence in the covariation model. However, the most abundant sequence contains the UUGAA motif leading to the possibility that it is recognizing the ligand with a T-loop. Comparison of the natural T-loop phylogeny with that of the selected Fhh4 cluster shows a strong correlation. Positions 3 and 2 are the most discrepant, though position 3 allows for most nucleotides and xB17 contained a U at position 2 that is not widely found in phylogeny. (d) While P1 is the main focus of this work, two additional sequences from the abundant clusters 4 and 51 were chosen to further validate the selection scheme.
with residues in J3/1 yet there is not enough covariation or acceptable mutations to affirm the interactions. Interestingly, while the Fhh4 family of RNAs showed a distribution similar to that of the natural T-loop motif, J2/3 of the purine selection is again invariant. It’s not clear what this implies, but it could be postulated that the purine scaffold is either only able to host one functional sequence of the T-loop or that their is one form of the T-loop that is able to form a unique set of interactions with the purine scaffold that confer a greater fitness for the aptamer. A small clade of sequences contains a J2/3 that is identical to xB17, yet it is not one of the ten most abundant clusters and comprises only 2.3% of the final pool. The J3/1 of this small cluster is identical to that of xB17, yet J1/2 has a much greater degree of heterogeneity. Position 19, the first randomized position in J1/2, that was a U–C mismatch in the xB17 aptamer is now more often observed as a canonical G–C pair. This feature was elucidated by the crystal structure since the mismatch was still in the A-form conformation, but it is yet another example of the power of deep sequencing and large amounts of diversity in the final pool.

4.4 Validation of the aptamer suite.

The selection and subsequent sequence analysis yielded several promising new candidates for aptamer development. Before their development into sensory devices, a baseline characterization of their affinities and structure are need to ensure that the selection was successful and that the overall fold was maintained. The sequences as detailed above were subjected to ITC and SHAPE analysis as detailed in the previous chapter. ITC binding clearly shows that the selection was successful, with all sequences tested having affinities in the low μM range (table). As compared to xB17 the aptamers selected in lower magnesium concentrations have an altered specificity profile while maintaining their affinity for 5htp.
The new aptamers share a preference for serotonin, the decarboxylated cation and biologically relevant form of 5htp, and the methyl amide form of 5htp (5htp-NHMe) that most closely resembles the molecule presented on the selection column.

The Fhh$_4$ and purine cluster 1 (P$_1$) aptamers share a common sequence motif predicted to bind the ligand and the ITC data further supports this as they share a near identical affinity and specificity profile. The two aptamers do show some nuanced differences though, with Fhh$_4$ preferentially binding serotonin (1.2 μM) over 5htp-NHMe (2.5 μM) while P$_1$ prefers the methyl amide (4.7 vs. 1.3 μM respectively). While the cdG$_{17}$ aptamer contains no discernible T-loop motif, it also shares this specificity profile and most closely resembles that of Fhh$_4$. This suggests either that cdG$_{17}$ RNA has a highly degenerate T-loop motif that cannot be identified in the sequence, or that it is in some way compensating for the lack of a T-loop which allows the RNA to form an identical ligand binding pocket to that of Fhh$_4$ and P$_1$ but with different sequence elements.

To further verify the method of sequence selection of the aptamers, two additional purine sequences from the abundant clusters 4 and 51 were analyzed by ITC. These two clusters share little sequence identity with cluster 1, with cluster 4 showing no evidence of a T-loop and 51 only a very weak adherence to the profile. The most abundant sequence from cluster 4 (P$_4$) is able to maintain relatively similar affinity for 5htp (11 μM) as compared to P$_1$, yet its ability to discriminate against tryptophan is impaired, showing only a 15 fold selectivity as opposed >60 fold for P$_1$. The cluster 51 sequence (P$_{51}$) also shows a preference for the 5-hydroxy indol moiety, but at an ∼4-fold reduction in affinity for all ligands except serotonin. These results demonstrate that this method of sequence selection produced by the sequencing analysis does in fact yield the most competent aptamer.

Structural probing of the backbone ribose for the new aptamers reveals that the global
fold was preserved through the selection. The parental Vc2 riboswitch exhibits strong c-di-GMP dependent changes in its reactivity pattern in J1/2 and J3/1, while the overall probing pattern is not well defined with many regions of the RNA showing high reactivity. The cdG$_{17}$ aptamer mainly shows protections in J1/2 upon the addition of 5htp with an additional slight protection in J2/3 (Figure 4.6). This resembles the covariation model for the aptamer, where residues in J1/2 are more highly conserved than those in J2/3. In the Vc2 structure, L2 docks into the minor groove of P3 via two A-minor triple interactions to form the distal tertiary interaction that supports the three-way junction, and the probing reveals a ligand dependent change in the P3 docking site. While the probing pattern is different for this interaction between Vc2 and cdG$_{17}$, the Vc2 RNA also shows ligand dependent changes in the P3 docking site and L2, suggesting that binding event in the junction is communicated to the peripheral motif and that this phenomena may have been preserved.

The reactivity pattern of Fhh$_4$ and the parental type-III fast hammerhead are largely

<table>
<thead>
<tr>
<th></th>
<th>5htp</th>
<th>L-Trp</th>
<th>Serotonin</th>
<th>5htp-NHMe</th>
</tr>
</thead>
<tbody>
<tr>
<td>cdG$_{17}$</td>
<td>9.3±0.3 μM</td>
<td>n.d.</td>
<td>1.2±0.1 μM</td>
<td>2.1±0.4 μM</td>
</tr>
<tr>
<td>Fhh$_4$</td>
<td>7.3±2.8 μM</td>
<td>n.d.</td>
<td>1.2±0.2 μM</td>
<td>2.5±0.5 μM</td>
</tr>
<tr>
<td>P$_1$</td>
<td>8.8±1.5 μM</td>
<td>520±90 μM</td>
<td>4.7±0.3 μM</td>
<td>1.3±0.1 μM</td>
</tr>
<tr>
<td>P$_4$</td>
<td>11±1.1 μM</td>
<td>165±15 μM</td>
<td>15.7±4.3 μM</td>
<td>6.6±0.4 μM</td>
</tr>
<tr>
<td>P$_{51}$</td>
<td>60±15 μM</td>
<td>n.d.</td>
<td>15.8±1.5 μM</td>
<td>25±8 μM</td>
</tr>
<tr>
<td>xB$_{17}$</td>
<td>3.9±0.1 μM</td>
<td>280±30 μM</td>
<td>38±8 μM</td>
<td>26±9 μM</td>
</tr>
</tbody>
</table>

Table 4.1: The affinity and specificity of aptamers containing different folds.

All measurements taken at 25 °C in 10 mM MgCl$_2$. n.d. = not determinable under these experimental conditions.
Figure 4.6: cdG_{17} shows ligand dependent modifications. The raw gel (right) shows clear ligand dependent modifications in J1/2 and J2/3. The parental Vc2 RNA shows a ligand dependent protection in P3 at the tetra-loop docking site, while the cdG_{17} shows a ligand dependent modification on the opposite side of the helix. Additionally neither RNA shows any modifications in the presence of the irrespective ligand. Integration of the bands as a function of distance down the gel is shown at the bottom. After normalization and assignment, the ligand dependent changes are still evident (colored asterisks), especially in J1/2.
the same, with areas of low and high reactivity aligning well (Figure 4.7). In order to compare the structures, position G₈ (G₄₇ in the Fhh library) of the fast hammerhead was mutated to a cytidine. This mutant is still functional at a very low level [225], yet it is slow enough to allow for structural probing allowing for comparison of the global reactivity pattern with Fhh₄. The cleaved band of the fast hammerhead ribozyme is clearly present in the gel, yet full length product is visible and a major product. Therefore, the probed structure of the RNA can be assumed to be that of the natural and folded form. The overall pattern between the RNAs is similar, with the helical region of P2 and P3 being protected in both RNAs. The L2 loop is disordered in both RNAs to a similar extent while the L3 loop, which docks into the P2 stem, appears more structured than in the parental fast hammerhead. This may be due to a significant portion of the fast hammerhead RNA being cleaved in J1/2, leading to a more conformationally dynamic L3. In addition to sharing a similar reactivity pattern, Fhh₄ shows ligand dependent changes in J1/2, J2/3, the L3 docking site in P2 and a corresponding alteration in L3. While there is no corollary pattern to be observed in the fast hammerhead riboswitch, it appears again that the binding event is localized to the junction region and transduce that signal to the global fold of the RNA.

The P₁ RNA showed only a mild protection in J2/3, which is a much weaker response than the parental xpt or orthogonal xB₁₇ aptamers. In order to observe a response, the probing conditions needed to be adjusted to 1 mM MgCl₂ from 6 mM and the temperature raised to 30 °C. Even under these conditions the response to ligand is weak, surprising given that P₁ binds the ligand with high affinity and a reasonable enthalpic gain (Figure 4.8). The fold is well maintained in the P₁ RNA however, as the highly modified nucleotides in L3 that are indicative of the formation of the tertiary interaction in xpt are present in the new aptamer. Additionally, the two other 5htp verified purine sequences were subjected to the
Figure 4.7: The Fhh₄ shows ligand dependent modifications and shares a highly similar reactivity pattern to the *S. mansoni* G8C parental RNA. The raw gel (right) shows clear ligand dependent modifications in J1/2 and J2/3. The changes in J2/3 localize mainly to positions 3 and 4 of the predicted T-loop motif. Additionally, if the structure was maintained the terminal loop of P3 (L3) that docks into P2 in the parental RNA shows a ligand dependent protection. Integration of the bands as a function of distance down the gel is shown on the left. After normalization and assignment, the ligand dependent changes are still evident (colored asterisks).
Figure 4.8: The purine scaffolded RNAs exhibit weak ligand dependent modifications and all isolated sequences preserved the purine fold. Three isolated sequences from separate clusters identified by the sequencing analysis show different reactivity patterns and areas of ligand dependencies. The parental xpt riboswitch shows a strong modification in the presence of guanine, while P₁ shows a weak protection and P₄ a much stronger one (blue box). P₅₁ shows no ligand dependent changes even with verified binding by ITC. The L3 signature that is indicative of tertiary structure formation is present in all sequence (yellow box) clearly demonstrating that the fold was preserved in the selection, likely for all major clusters.
structural probing, with P4 showing the most definitive ligand dependent changes in J2/3 and J1/2 under these conditions. All three sequences isolated from the selection maintained the L3 signature, meaning that even aptamers with lower affinity and specificity are hosted within the scaffold and not an alternative fold as witnessed in the previous selection. Likely, many of the sequences in the final pool of the selection are well folded and contain the tertiary contact.

4.5 Aptameric screen for novel sensory devices.

With the isolated sequences from each scaffolded selection proving to be competent aptamers, we sought to investigate their potential to modulate transcription. While the aptamers had been selected for 5htp, they show a preference for serotonin. This is not unexpected given that serotonin is closer to the presentation on the column, but it should not be overlooked as it may provide a convenient, non-toxic alternative to 5htp should high ligand concentrations be needed to elicit a maximal response. Additionally, there is no high level understanding of the structure available to facilitate rational design and optimization. This makes aptameric screening strategies a harsh test of the design process, with a complete reliance on simple sequence modulation for the development of these sensors.

A simple screening strategy was again employed in the context of the pbuE modular riboswitch platform (Figure 4.9). The key difference in this iteration of development is that the adaptor platform and communication module are left fixed, and the potential sensor domain is the main variable in the scheme. Since several of the aptamers showed covariation at the base of P1, and xB17 possessed a weak base pair at this position possibly do to replicative pressures, the first two nucleotides of J1/2 and J3/1 were screened to provide some additional diversity. This is likely not needed, however, as the pbuE platform is tolerant
Figure 4.9: **The ability to screen for sensor domains against the same ligand.** With the scaffolded selections proving successful at producing high affinity and specific aptamers in various contexts, we are now able to screen the aptamer domain for functionality. The lead candidate from each selection was appended to the pbuE platform in the absence of a communication module. Additionally, the first and last residues of the library in cdG$_{17}$ and Ffh$_4$ are predicted to form a canonical Watson-Crick base pair, so they were removed to provide some additional diversity in an otherwise very limited screen. The P$_1$ aptamer is predicted by the data analysis to have a bulged cytidine as the last residue of the library in J3/1. While the folding prediction shows an elongated P1, this is possibly untrue given there is no covariation to support the secondary structure in this region. Therefore the aptamer was appended with and without the C. The pbuE adaptor platform is ideal for very limited screens as it is tolerant of mutation and functional over a wide range of P1 stabilities.
to a wide variety of P1 lengths and stabilities as seen previously in with xB17 and other studies [85].

In this very limited scheme which tests only the aptamer variable, the P1 RNA was able to attenuate transcription and act as a functional sensor domain. With no further optimization, the P1/pbuE device possessed a dynamic range of 54-66% with basal levels of read through product at 10% (Figure 4.10). With a robust switch, the ligand concentration necessary for half-maximal response of the device (T<sub>50</sub>) was able to be determined for all three ligands examined in the ITC experiments. The device shows an apparent T<sub>50</sub> of 190 μM for 5htp and 125 μM for the methyl amide which resembles the previously determined specificity profile. These values are ∼20 and ∼100 fold higher than the K<sub>D</sub>, respectively, which is a significant reduction in sensitivity though not beyond the realm of what is normally seen [84]. Interestingly, P1/pbuE shows a T<sub>50</sub> = 30 μM for serotonin which is only an ∼6-fold discrepancy from the K<sub>D</sub>. While this low value deviates from the ITC specificity profile, it is not entirely unexpected as many riboswitches operate under a kinetic regime where the T<sub>50</sub> is more a property of the on rate than the equilibrium state [160,226].

This is a drastic improvement over the previous xB17 devices, which showed much weaker T<sub>50</sub> values if they could even be established and low dynamic ranges of less than 15%. Previous work would suggest that even the level of 10% read through in the absence of ligand would make portage of the device in vivo difficult. While the device has not been thoroughly tested in vivo, initial studies do suggest that the amount of read through is enough to saturate the system, regardless of ligand concentrations. While the expression levels and dynamic range of the device may need tuning in order to elicit a response in vivo, these are engineerable properties of the adaptor platform.

While these studies were being performed, another iteration of the fluorogenic Spinach
Figure 4.10: P₁ successfully attenuates transcription and maintains specificity. Sample titrations are shown for each ligand tested (top). Serotonin shows the highest observed affinity (graph) yet the levels of transcription decrease rapidly at concentrations higher than 1 mM. While the specificity of the switch overall is maintained (inset), the variability in T₅₀ suggests that the device is under kinetic control, like many natural riboswitches.
RNA was published [104]. This adapter, termed Broccoli, began with the Spinach round 5 RNA pool from the selection and was cloned into *E. coli* for subsequent rounds of selection and polishing with the partitioning pressure on cellular fluorescence. The manuscript demonstrates that Broccoli is more thermostable, less magnesium dependent, and brighter than previous iterations and that it should be a more robust tool for *in vivo* application. Therefore, we decided to test the aptameric screening strategy in the context of an additional adaptor to further demonstrate the speed and viability of possessing multiple sensor domains for engineering devices.

It should be noted that the Broccoli was described ignorant of the crystal structure, even though the structure was published before the Broccoli manuscript was in review. Many of the mutations that were produced in the *in vivo* polishing steps are predicted to be neutral as they are either distal from the binding pocket in a paired region or projected into solution. Despite the authors claim that Broccoli is a novel aptamer, it is beyond a reasonable doubt that it is a variation of the pre-existing Spinach RNA with mildly improved characteristics. Nevertheless, when a simple aptameric screening strategy is again employed to survey for functional devices, many of them produce capable *in vitro* sensors (Figure 4.11).

The communication modules used for the Broccoli sensor development were two of the 4 basepair, destabilized designs used in the second iteration of the minSpinach constructs (A and U annotation here). These two communication modules possessed the greatest dynamic range and approached the highest levels of maximal fluorescence observed in those studies, so they were known to elicit the best response in our hands. Only the traditional orientation of Broccoli was surveyed in this work since the minSpinach_B construct showed a 40% decrease in fluorescence for unknown reasons. For comparison, the xB\textsubscript{17} aptamer was also included in these studies as a reference for improvements in the Broccoli aptamer itself, not just the
sensor domains.

The affinity for 5htp and serotonin as well as the magnesium dependence for all of the constructs were determined. While the Broccoli aptamer has reduced magnesium dependence on levels of maximum fluorescence, it does affect the maturation time of the device. Our constructs, including the Broccoli RNA in isolation, often required between 10 and 30 minutes to reach full fluorescence in 1-3 mM magnesium. At concentrations >5 mM the effect was minimal, with the constructs coming to equilibrium before the experiments could be assayed. Therefore, all constructs were allowed to incubate for 30 minutes prior to assaying in order to assure the device had come to full equilibrium, resulting in maximal fluorescence at 1 mM magnesium for the Broccoli construct. Interestingly, 1 mM serotonin reduces the fluorescence of Broccoli at 1 mM magnesium concentrations to half of that in the absence of ligand. As magnesium concentrations are increased the brightness is slightly rescued, showing 70% recovery at 10 and 20 mM magnesium.

Even with the ligand effects on the brightness of Broccoli, several P1 and Fhh4 devices that contained the full aptamer sequence showed an induction in fluorescence upon the addition of ligand (Figure 4.12 and Table 4.2). While the Broccoli aptamer alone does not exhibit a strong dependence on magnesium, the new devices show an effect. Under conditions containing 1 mM MgCl2, the response of the devices is suppressed, with the best sensors showing only 30% the brightness of Broccoli alone and induction factors of 3-4 fold. While this is a heavy reduction in the overall brightness, these sensors already show values that best those of the minSpinach constructs. At 3 mM MgCl2, the Fhh4 aptamer with the U communication helix (Fhh4U) shows a level of brightness equal to that of broccoli with a serotonin induction factor of 5.5. The xB17 sensor from the previous purine selection also functions well within the context of the Broccoli aptamer. Interestingly, the xB17 shows one
Figure 4.11: A simple strategy for the creation of robust fluorogenic sensors. The same aptameric modulation that was used to create the *pbuE* device was employed for Broccoli sensors. A communication module is required for the devices in order to transduce the binding event to the Broccoli adaptor/actuator. Two 4 base pair modules (A and U here) that were optimized in the previous minSpinach constructs were chosen based on their ability to produce high levels of fluorescence and sufficient dynamic range. The Broccoli RNA is a slight variation of Spinach, with the point mutations specifying the difference highlighted in green and residues that partake in formation of the G-quadruplex are in red.
(a) (b)
(c) (d)

Figure 4.12: The magnesium concentration has variable effects. The concentration of magnesium has a limited effect on the broccoli aptamer alone, yet all of the aptmeric sensors show some dependence. As the magnesium concentration is modulate from 1 to 10 mM (a-d) both the dynamic range and maximal levels of fluorescence are altered. All data points contained 0.5 μM RNA and 10 μM DFHBI with an external normalization.
Table 4.2: Broccoli Sensors Show Tunable Response Ranges and Affinities

<table>
<thead>
<tr>
<th>Construct</th>
<th>Ligand</th>
<th>[MgCl₂] mM</th>
<th>Fold Induction</th>
<th>Fₘₐₓ (%Broccoli)</th>
<th>F₅₀ μM</th>
</tr>
</thead>
<tbody>
<tr>
<td>xB₁₇A</td>
<td>5htp</td>
<td>1</td>
<td>3.5</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>6.3</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>7.3</td>
<td>122</td>
<td>190±30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>6.0</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>Serotonin</td>
<td></td>
<td>20</td>
<td>4.1</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1.3</td>
<td>6.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>3.1</td>
<td>34.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>3.7</td>
<td>62.7</td>
<td>n.d.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>3.9</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>3.0</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>xB₁₇U</td>
<td>5htp</td>
<td>1</td>
<td>1.9</td>
<td>21.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>3.1</td>
<td>72.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>3.3</td>
<td>105</td>
<td>180±20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>3.3</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>Serotonin</td>
<td></td>
<td>20</td>
<td>3.2</td>
<td>135</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.4</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.8</td>
<td>18.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>1.1</td>
<td>33.2</td>
<td>n.d.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1.3</td>
<td>52.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>1.6</td>
<td>66.4</td>
<td></td>
</tr>
<tr>
<td>P₁₄A</td>
<td>5htp</td>
<td>1</td>
<td>1.2</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1.3</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>1.4</td>
<td>3.1</td>
<td>n.d.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1.7</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>Serotonin</td>
<td></td>
<td>20</td>
<td>2.0</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1.0</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1.3</td>
<td>2.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>1.4</td>
<td>3.2</td>
<td>190±70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1.8</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2.2</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>P₁₁U</td>
<td>5htp</td>
<td>1</td>
<td>2.5</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>6.0</td>
<td>37.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8.2</td>
<td>71.2</td>
<td>240±20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>8.2</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>Serotonin</td>
<td></td>
<td>20</td>
<td>6.1</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>5.1</td>
<td>17.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>8.8</td>
<td>54.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>9.0</td>
<td>78.1</td>
<td>52±4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>7.2</td>
<td>98.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>5.1</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>Fhh₁₇A</td>
<td>5htp</td>
<td>1</td>
<td>1.1</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1.8</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>2.4</td>
<td>10.4</td>
<td>n.d.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>3.7</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Serotonin</td>
<td></td>
<td>20</td>
<td>5.1</td>
<td>26.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1.5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>4.4</td>
<td>17.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>7.2</td>
<td>31.4</td>
<td>790±190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>10.9</td>
<td>53.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>13</td>
<td>69.5</td>
<td></td>
</tr>
<tr>
<td>Fhh₁₇U</td>
<td>5htp</td>
<td>1</td>
<td>1.6</td>
<td>12.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>2.5</td>
<td>43.3</td>
<td>590±90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>2.8</td>
<td>69.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>3.1</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>Serotonin</td>
<td></td>
<td>20</td>
<td>3.0</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>3.8</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>5.5</td>
<td>93.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>4.8</td>
<td>116</td>
<td>260±50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>3.7</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>3.2</td>
<td>136</td>
<td></td>
</tr>
</tbody>
</table>
of the strongest dependencies on magnesium, not reaching full levels of fluorescence until conditions contain 10 mM magnesium. The magnesium dependence appears to be a trend with the purine scaffold, with the P₁ aptamer unable to produce maximal fluorescence with either communication module until 10 mM magnesium. The most interesting trend observed in these studies is the fact that several of the RNAs produce fluorescence greater than that of the Broccoli aptamer alone. The effect only occurs at moderate to high magnesium concentrations, but it suggests that the aptamers are capable of structuring Broccoli to a greater degree than on its own. This supports the original rational for scaffolded selection providing the context for robust folding, and with previous reports showing that Broccoli is only \(\sim 60\%\) folded \[104\] it makes this conclusion tempting, though further validation is required.

The communication module can have drastic effects on the RNA devices. The A and U communication modules in the context of xB₁₇ serve mainly to dictate the dynamic range, leaving levels of fluorescence and specificity a function of magnesium and the aptamer as observed before. In the context of the Fhh₄ and P₁ aptamers, however, the communication module can decide a very competent sensor from a completely ineffective one. This is most obvious for P₁, where the A module can never produce sufficient levels of fluorescence, yet the U module can rescue the near full fluorescence of broccoli and shows a very strong observed affinity (\(F_{50}\)) of 52 \(\mu\)M in 5 mM magnesium conditions. The communication module can also dictate the observed specificity as a function of magnesium in the new aptameric devices, where this appears a function of the aptamer in the context of the xB₁₇ RNA. Constructs with the xB₁₇ aptamer show a strong preference for 5htp over serotonin independent of magnesium concentration, while the P₁ and Fhh₄ constructs prefer serotonin until the magnesium concentration reaches a level that allows for equal recognition of both ligands,
likely indicating a higher affinity state of the aptamer for both ligands. Additionally, the Broccoli devices show no appreciable affinity for tryptophan, a property which must be true if these devices are to be ported into *in vivo* systems. The selectivity, affinity, and ability to produce at least equal levels of fluorescence at physiological conditions suggest that these devices should function well in a host organism, with those studies currently being pursued.

### 4.6 Discussion

The strategy of using multiple complex scaffolds as hosts for limited libraries has produced multiple competent aptamers with high affinity and specificity. Additionally, the scaffolds allow for the expedient analysis of sequencing data, as they limit the amount of variable sequence that needs to be analyzed making highly related clusters and likely functional candidates easily identifiable for further validation and downstream analysis. The limitation in sequence diversity does not come at the expense of structural complexity as this is presented by the variable scaffold allowing for the sampling of many highly complex and diverse sequences.

The production of multiple solutions allows for an additional variable to be screened in the quest for novel RNA devices. This was previously unattainable, with the majority of selections producing a single understandable solution leaving the aptamer’s utility as a sensory part dependent upon its hopeful ability to interface efficiently with one of the various RNA adaptors. This was readily demonstrated in the initial purine selection where the single scaffolded aptamer was unable to efficiently function as a sensory device in the context of transcriptional adaptors even thought the aptamer was proficient at specifically recognizing the ligand. By attaining multiple high affinity solutions, we are afforded a novel strategy; aptameric screening. The utility of aptameric screening is best demonstrated in the *pbuE*
transcriptional attenuation devices, where no further optimization of the adaptor platform was needed to produce a synthetic switch that rivals the response of natural riboswitches. While a simple port into a cellular host did not produce a capable \textit{in vivo} sensor it is likely not a property of the aptamer, but a tunable aspect of the riboswitch adaptor platform, the variables of which have only begun to be explored. Once the \textit{in vivo} properties of the modular riboswitch expression platforms are understood and produce a portable system, it seems likely that the aptameric screening strategy will produce an adept device.

The results of the aptameric screen with the coupled adaptor/actuator Broccoli have provided many tempting candidates for intracellular sensors. Several of the new devices seem capable of functioning at, or even above, the limit of Broccoli, with \(\text{xB}_{17}\) showing a relative fluorescence nearly 2-fold brighter than the uncoupled fluorescent RNA. The suite of sensors also allows for the possibility of sensing both 5htp and serotonin independently, as the various sensory domains provide tunable affinities for the two ligands. The ability of Broccoli to function intracellularly as a coupled small molecule sensor has yet to be demonstrated in any context, yet with Broccoli being an evolutionary successor to Spinach it seems likely that it is capable and these possibilities are currently being explored. If an intracellularly functional variant is had, this study could provide the basis for a detailed understanding of the attributes necessary to attain a functioning intracellular device, the basis for which is not well understood. Regardless of \textit{in vivo} results, the aptameric screening strategy is capable of producing devices that operate at the full functional potential of the coupled RNA adapter, relieving the pressure on the sensory domain’s performance and shifting the focus to the development of more capable \textit{in vivo} adaptors.
4.7 Materials and Methods

The library construction and column preparation were identical to the first iteration of selection but with the following notable exceptions: The buffer for selection contained reduced magnesium and more physiologically relevant monovalent cations (25mM HEPES, pH 7.0; 150 mM KCl; 50 mM NaCl; 3mM MgCl$_2$): GsI-IIC-MRF reverse transcriptase was used in place of SuperScript III. The precipitated RNA pellet was brought up in 1.25 mM dNTPs and 20 μM RT-PCR primer prior to denaturation at 65 °C, annealing at 4 °C, and equilibration at 60 °C. The solution was then brought up to 1x GsI-IIC-MRF buffer conditions (10 mM NaCl; 1 mM MgCl$_2$; 20 mM Tris, pH 7.5; 1 mM DTT) in 20 μL total volume and sufficient enzyme was added for extension at 60 °C. PCR was as previously described.

<table>
<thead>
<tr>
<th>Library</th>
<th>Tnx Start</th>
<th>P1 Stem</th>
</tr>
</thead>
<tbody>
<tr>
<td>xpt-pbuX</td>
<td>GCGGCAGATTCCTATACGACTCACTATAAGGACTT-CGGTCCTTGGATAGGACTCNNNNNNCGC-GTGGATATGGCAGCNNNNNNGGCGACCG-TAAATGTCNNNNNNGGTTCTATCCCAATCG-GGCTTCGGTCGGTTCC</td>
<td></td>
</tr>
<tr>
<td>Purine II</td>
<td>GCGGCAGATTCCTATACGACTCACTATAAGGACTT-CGGTCCTTGGATAGGACTCNNNNNNCAC-ACCATTCCGAAAGAGTGGGACGNNNNNCTCCGCCTAA-ACCGAAGTAGGTAGCAGGGNNNNNNTGTCCCTATCC-CCAAACGGCTTCGGTCGGTTCC</td>
<td></td>
</tr>
<tr>
<td>Vc2 cdG</td>
<td>GCGGCAGATTCCTATACGACTCACTATAAGGACTT-CGGTCCTTGGATAGGACTCNNNNNNCAC-AATGAAATGTACTACCCANNNNNNNGCCTCCTATCC-CCAAATAGGNNNNNNGCTCCTATCC-CCAAACGGCTTCGGTCGGTTCC</td>
<td></td>
</tr>
<tr>
<td>S. mansoni Fhh</td>
<td>GCGGCAGATTCCTATACGACTCACTATAAGGACTT-CGGTCCTTGGATAGGACTCNNNNNNCAC-AATGAAATGTACTACCCANNNNNNNGCCTCCTATCC-CCAAATAGGNNNNNNGCTCCTATCC-CCAAACGGCTTCGGTCGGTTCC</td>
<td></td>
</tr>
</tbody>
</table>

<p>| Selection Scheme |</p>
<table>
<thead>
<tr>
<th>Round</th>
<th>[RNA] pmol</th>
<th>Washes</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000</td>
<td>3</td>
<td>Counter selection against AcO-sepharose</td>
</tr>
<tr>
<td>2</td>
<td>400</td>
<td>3</td>
<td>Counter selection against AcO-sepharose</td>
</tr>
<tr>
<td>3</td>
<td>400</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>200</td>
<td>6</td>
<td>30 second CS with 100 μ L-Trp</td>
</tr>
<tr>
<td>7</td>
<td>200</td>
<td>10</td>
<td>30 second CS with 500 μ L-Trp</td>
</tr>
</tbody>
</table>

Sequencing and analysis
Reverse primers containing barcodes used to append necessary sequence. Barcodes courtesy of Rob Knight’s group. 1: Illumina adapter 2: Golay barcode 3: Primer pad 4: library compliment
### Chemical Probing

RNA was prepared as described previously [168]. Structure cassettes flanking the 5' and 3' ends of the RNA were added to facilitate reverse transcription (Table below) and NMIA modification was performed using the established protocols [159] at 25 °C except for the purine sequences which were probed at 25 °C. RNA was probed at 100 nM in 100 mM Na-HEPES [pH 8.0], 100 mM NaCl, and 6 mM MgCl₂ or 1 mM MgCl₂ for the purine isolates. Ligand concentration was 1 mM where indicated. Gel images were analyzed by SAFA [173] and ImageJ (NIH).

### Isothermal Titration Calorimetry

All RNAs tested were exchanged into the ITC binding buffer (10 mM HEPES pH 7.0; 250 mM NaCl; 50 mM KCl; 10 mM MgCl₂) and washed three times in 10k MWCO filter (EMD Millipore). The ligand was brought up from a dry solid directly into the binding buffer and concentration established on a NanoDrop 2000 (Thermo Scientific: ε₂₇₅ = 8000 mol⁻¹ cm⁻¹ for 5-hydroxy indol moiety). The RNA was diluted to between 50-100 μM and the ligand was titrated at 10 fold that of the RNA. Titrations were performed at 25 °C using an iTC-200 micro calorimeter (Microcal, Inc.; GE). Data was analyzed and fitting was performed with the Origin 5.0 software suite (Origin Laboratories).

### In vitro Transcription Assay

As described previously in Section 3.6 with no alterations.

### Broccoli Fluorescence Assays

RNA was prepared as described previously [168], with additional 0.5x TE buffer washes in a 10kMWCO Amicon Ultra (Millipore) to minimize the carry over of metal ions. All constructs were assayed at 0.5 μM RNA and 10 μM DFHBI in a buffer containing 80 mM Tris-HCl, pH 7.4, 150 mM KCl, and 50 mM NaCl. The buffer, ligand, magnesium, and DFHBI were mixed prior to the addition of RNA and allowed to incubate for 30 minutes at room temperature. The concentration of ligand in the magnesium screens was 1.5 mM and magnesium was 5 mM in the ligand affinity experiments. All assays were performed in Greiner 96-well flat bottom black fluorescence plates (Thermo Scientific) on a Tecan plate reader exciting at 460 nm and collecting and averaging 506-510 nm emission.
<table>
<thead>
<tr>
<th>Sequence Isolates with SHAPE and ITC Primers</th>
<th>Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>cdG$_{17}$</td>
<td>GGATAGGACATGTAATCTCCAAACCATTGCCGAAAGAGTGAGCCTAGACCTCCGGCTAAACCAGGAAGGTAGGTTAGCGGGGCTAGGATCTCCTATCC</td>
</tr>
<tr>
<td>Fhh$_4$</td>
<td>GGATAGGAGCTGTTTTGGTATCCAATGAAAATGACTACC-AACTTGAATCTCCAAATAGGCTAGGTAGTATCCTATCC</td>
</tr>
<tr>
<td>P$_1$</td>
<td>GGATAGGACCTCTCTGTTTGGCTAGATAGCCAGCGCAAT-TGAAGAATGGGCACCGTAAATGCTGTAGACGGGTTTCCCTATCC</td>
</tr>
<tr>
<td>P$_4$</td>
<td>GGATAGGACTCTCATCCGCGGCGGATGAGCACGCAGGAGATGTGTGTTGGACACCCTCCCTAGGACCCGGGTCTCTATCC</td>
</tr>
<tr>
<td>P$_{51}$</td>
<td>GGATAGGACTCAACCATCTCCTGCGGATGATGGCAGACGCAGCCCTCCAGTGGGCACCGTAAATGCTCCGTAGACCGGCTCTATCC</td>
</tr>
<tr>
<td><strong>F T7:SHAPE</strong></td>
<td>GCGCGCGAATTCTAATACGACTCACCTATAGGAACCTTCCG-TTCCTTGGATAGGAA</td>
</tr>
<tr>
<td>R SHAPE</td>
<td>GAACCGGAACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGATTGGGATAGGA</td>
</tr>
<tr>
<td><strong>F T7:cdG$_{17}$</strong> ITC**</td>
<td>GCAGCGCGAATTCTAATACGACTCACCTATAGGATAGGACATGTAATCTCC</td>
</tr>
<tr>
<td>R :cdG$_{17}$ ITC</td>
<td>GGATAGGACATACCTAGCCACG</td>
</tr>
<tr>
<td><strong>F T7:Fhh$_4$ ITC</strong></td>
<td>GCGCGCGAATTCTAATACGACTCACCTATAGGATAGGAGCCTGTTGCTATGCC</td>
</tr>
<tr>
<td>R Fhh$_4$ ITC</td>
<td>GGATAGGAGCTACCTAGCC</td>
</tr>
<tr>
<td><strong>F T7:Purine ITC</strong></td>
<td>GCGCGCGAATTCTAATACGACTCACCTATAGGATAGGAGCCTTGGATAGGACTTC</td>
</tr>
<tr>
<td>R Purine ITC</td>
<td>GAACCGGACCGGAAAGCCCG</td>
</tr>
</tbody>
</table>
### Sensors

<table>
<thead>
<tr>
<th>Sensor</th>
<th>A: Broccoli</th>
<th>U: Broccoli</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₁ₐ</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
</tr>
<tr>
<td>P₁ᵤ</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
</tr>
<tr>
<td>Fₜ₄ₐ</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
</tr>
<tr>
<td>Fₜ₄ᵤ</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
</tr>
<tr>
<td>xB₁₇ₐ</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
</tr>
<tr>
<td>xB₁₇ᵤ</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
<td>GGACGGAGACGGTGGTCATATTCTGGTTCCGCTAGAT-ATGGCAGCATTGAAGAATGGGCACCGTAAATGTCTGT-AGACATATTCCAGTAGAGTGTTGGGCTCCGGTCC</td>
</tr>
<tr>
<td>P₁/pbuEₗ</td>
<td>AATATTGAGCTGTGGACAATTAATCATCGGCTCGTATAATGTGTGGAATTAAATAGCTATTATCAGGATTTTTCTGTGTCGCAGTATATTGGCAGAATGGGCAAGTGAATGGTGACAGTTTTTGTTATGACATTTTTGTGAATCATCAGGATTTTTTTTTATATCAAAAAAAACACCTTAAAGGAGTTTGTGAT</td>
<td>AATATTGAGCTGTGGACAATTAATCATCGGCTCGTATAATGTGTGGAATTAAATAGCTATTATCAGGATTTTTCTGTGTCGCAGTATATTGGCAGAATGGGCAAGTGAATGGTGACAGTTTTTGTTATGACATTTTTGTGAATCATCAGGATTTTTTTTTATATCAAAAAAAACACCTTAAAGGAGTTTGTGAT</td>
</tr>
</tbody>
</table>
The case for RNA as a developmental tool in synthetic biology is strong and has been made repeatedly. With its minimal genetic footprint, decomposable domains, low toxicity and side effects, cis regulatory mechanisms, and the ability to exert its function in the absence and independently of host regulatory mechanisms makes RNA the ideal macromolecule for engineering. The major area of interest for synthetic RNA development has been in the small molecule recognition field, where synthetic riboswitches capable of inducing a biological response, such as gene expression or fluorescence output, in the presence of an external effector molecule are desired. The aggregate work to date has been inspired by the revelation of pervasive natural RNA regulatory systems and the advancement of in vitro selection. Using in vitro selection, hundreds of aptamers have been generated that bind a variety of ligands including nucleotides, antibiotics, amino acids, intracellular cofactors, and carbohydrates among many others. With the ability to generate such a wide variety of potential RNA recognition domains and the development of several successful sensors, such as theophylline, tetracycline, and neomycin, it has generally been assumed that traditional selections can meet the needs of synthetic RNA biology. However, there exists only a handful of examples where aptameric devices are capable of functioning intracellularly, and considerably fewer in mammalian systems, possibly due to this assumption (Table 5). The initial hinderance
in the development of these devices was a lack of a robust and modular adaptor platform capable of mediating the binding event of the aptamer to a greater cellular response. As there was no clear starting point for an adapter, development has focused heavily on these systems and they now include fluorescent RNA, transcriptional attenuation cassettes, and allostERIC ribozymes among others.

Every developed adapter has only been validated in vivo with either naturally occurring riboswitches or the previously mentioned synthetic aptamers, greatly limiting their potential. The dearth of synthetic aptamers capable of transitioning into functional systems is troubling, with recent insights suggesting that the high affinity and thermal stability of aptamers produced by in vitro selection are beneficial to switching, but that those properties alone are not sufficient for function. In the case of the neomycin selection, a destabilized unbound state that allows for greater conformational rearrangement was found to be the determining factor in functionality [126, 127]. These properties are not readily selectable in current techniques, at least not in the absence of selections where the partitioning pressure is a functional output. Additionally, it is often necessary to engineer this destabilized basal state into the RNA, a process that inherently weakens the affinity of the aptamer, often to beyond the limits of practicality.

The folds of natural riboswitch aptamer domains possess both a high affinity binding state and a large degree of conformational flexibility, a trend that now seems to be emerging in functional synthetic aptamers. Moreover, riboswitches are under natural selective pressures that incorporate the need to fold efficiently and robustly in the absence of high magnesium concentrations. Small catalytic RNAs are under similar pressures, but with the fold serving to position critical catalytic residues instead of a ligand binding pocket. Additionally, many of these RNAs belong to a greater fold of the three-way junction supported by
Table 5.1: **The in vivo RNA devices** The lack of competent synthetic aptamers able to function intracellularly in any context is troublingly low. Adapted from [227]

<table>
<thead>
<tr>
<th>Expression platform</th>
<th>Aptamer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theophylline</td>
</tr>
<tr>
<td>Translational ON</td>
<td>B</td>
</tr>
<tr>
<td>Translational OFF</td>
<td>B Y</td>
</tr>
<tr>
<td>Transcriptional ON</td>
<td>B</td>
</tr>
<tr>
<td>Transcriptional OFF</td>
<td>B</td>
</tr>
<tr>
<td>Ribozyme</td>
<td>B Y M Y M Y</td>
</tr>
<tr>
<td>Splicing</td>
<td>Y</td>
</tr>
<tr>
<td>Trans-acting switch</td>
<td>B</td>
</tr>
</tbody>
</table>

A combinatorial matrix of sensory aptamers, the type of expression platform (adaptor), and the system in which it has been demonstrated (B-bacteria; Y-yeast; M-mammalian cell culture)

... a distal tertiary interaction, which hosts many diverse functions and is pervasive throughout biology. It therefore stands to reason that the three-way junction motif provides the context for function, regardless of what that function may be. We can then remove the cognate function of well understood representatives of the three-way junction that possess desired properties, such as robust folding and multiple dynamic states, and program in a desired one, such as the recognition of 5htp. The new aptamers contain many of the properties of the parental scaffold, such as magnesium dependence and communication of the binding event to the global fold of the RNA in an analogous manner to the cognate ligand. With many features of the novel RNAs attributed by the the scaffold, and the scaffolds being understood well themselves, it eases the process of validating and comprehending the new RNAs on a structural level.

The ability to engineer well folding motifs that were evolved to function *in vivo* into the selection is powerful, but the ability of scaffolds extends beyond this. From a sequence analysis standpoint, the presence of fixed regions is beneficial, allowing for high quality alignments
that enables the observation of minimal small motifs responsible for ligand recognition. Mo-
tifs as degenerate as the T-loop would be difficult to identify in traditional selections without
adjacent highly conserved residues. The fixed sequences presented by the scaffold allow for
small motifs such as the T-loop to be identified, characterized, and validated easily when it
might otherwise be missed. The scaffold also limits the sampled sequence space and restricts
the solution to a local minimum, likely not the global. However, this does not inhibit the iso-
lation of diverse and competent aptamers. Obtaining multiple solutions in multiple contexts
has a practical benefit, the ability to screen aptamers as sensors against adaptor platforms.
Traditional selective techniques produce the least complex solution, often resulting in only
a single identifiable sequence that constitutes the majority of the final pool. This limits the
engineering potential of RNA devices to a one dimensional screen. If the adaptor platform
is additionally constrained, by host choice or application per se, development of the device
becomes even more difficult and often impossible to implement. Completely fixing the adap-
tor platform, as in the pbuE and Broccoli studies, is an incredibly harsh test of the idea of
aptameric screening. Yet one of the aptamers produces a result that rivals some of the best
known switches in either the natural or synthetic world, proving that scaffolded selections
are a useful part in a growing RNA synthetic biology toolkit.

While the in vitro analysis of the devices shows the competency of aptameric screening,
the looming question is their capability in vivo. While these experiments are currently
underway, they are not likely a direct test of the sensor domains developed in these studies.
Many facets of Broccoli remain unknown, especially given the caveats associated with the
Spinach predecessor and the degree of similarity between the two constructs. Time and
additional studies will determine if the issues with Spinach are truly resolved. If the Broccoli
aptamer is capable of functioning as an in vivo small molecule sensor, it seems reasonable
to assume that the scaffolded devices, with their equal or greater levels of fluorescence, will also be functional.

With the initial studies of scaffolded selections coming to a close, focus now shifts towards the development of additional functional sensors for biotechnical applications. Scaffolded selections provide a medium between traditional selections with subsequent difficult in vivo screening and small molecule screening strategies, such as drug discovery efforts against riboswitches, that do not allow for control of the ligand produced. In consideration of new effector molecules, the optimal ligand must meet several criteria such that it is not limited in downstream applications. The ligand must be inexpensive to produce, be non-pharmacological and non-toxic in both bacteria and eukaryotes, have no interfering intracellular compounds, be small and cell permeable, and allow for the isolation of high affinity aptamers. Containing all of these attributes makes the list sparse, however, this is not a new field and several existing compounds, e.g. 2,4-dinitrotoluene, have been investigated and largely abandoned due to their inability to effectively regulate. Compounds of this nature could benefit from a modular aptamer which would enable their broad application and interface with the existing toolbox. The scaffolded selection techniques detailed above could enable this, substantially alleviating the pressure on a tool in the synthetic biology toolbox.
Bibliography


Simon Blouin and Daniel A Lafontaine. A loop–loop interaction and a K-turn motif located in the lysine aptamer domain are important for the riboswitch gene regulation control. RNA, 13(8):1256–1267, August 2007.


Appendix A

Deep Sequencing Analysis Methods

The intent of this appendix is to give some insight into the workings of the clustering, alignment, and analysis of sequencing data that results in the selection of maximum-likelihood, single sequences from a pool of hundreds of thousands. The result of these efforts is a script (not a program), that invokes published and open source code from many sources. I refer to the script as Archimedes throughout, though that name is ambiguous to its function and workings.

Several things must be done prior to evoking the script. If barcoding was performed (i.e. multiple libraries were sequenced on the same chip) they must first be dereplicated. For this I used scripts from the QIIME package [228], the first of which was split_libraries_fasta.py, which performs quality filtering by truncating the sequence at a user defined Phred score (Phred ≥ 20 in this study corresponding to a 1% chance that nucleotide is incorrectly identified) and removes any sequences with ambiguous base calls. It then uses the two files generated by the Illumina sequencer, the actual sequencing file and the barcode file both in .fastq format, along with a user generated mapping file to assign barcode I.D.s to each sequence (e.g. purine library or flh library; Above). The resulting
.fasta file that contains all demultiplexed sequences was then fed to `split_libraries.py` with the same mapping file to remove any sequence after the P1 stem on the 3' side of the sequence. A separate fasta file that contained all sequences for a given barcode was then generated with the `filter_fasta.py` script. These files were then trimmed to the base of P1 on the 5' side and numerically renamed using the FastX toolkit in the Galaxy web server. The resulting .fasta files were the input for the Archimedes script, though Archimedes will handle any .fasta file format should the quality filtering and barcode splitting be treated differently.

![Diagram](image_url)

**Figure A.2: Overview of sequencing analysis pipeline**

The first step for Archimedes uses the clustering algorithm within Usearch [166] as invoked through the QIIME script `pick_otus.py`. This method generates seed sequences that are then clustered based on a user defined percent identity. For these studies, they were clustered at 90% and any clusters containing only a single sequence were discarded. The result is a file that contains sequence identifiers for all sequences in a cluster. The top ten most populous clusters are then mapped back to their original sequence file and all sequences for a cluster are written to a file in .fasta format. These files are usually very large if there is any degree of convergence for the sequenced pool. This level of sequence information
is redundant, containing many copies of the same sequence and the number of sequences limits the alignment and motif identification employed later. Therefore, 250 sequences were taken as a representative of each cluster for further analysis. More sequences than this did not provide any greater levels of information and the time for analyzing 10 clusters of 250 sequences each was 30 minutes to an hour, so this seemed a reasonable compromise between the amount of information and computational expense. These parameters can be easily adjusted in the shell script should more or less sequences need to be analyzed.

From this point the analysis is split and aligned by two different methods. This was done because the results from the de novo workflow of CMfinder [199] were often difficult to interpret. The idea of CMfinder is powerful, where small candidate motifs for each sequence are selected on stability, then compared among sequences to inform on frequency, and highly abundant motifs used to generate local alignments. This alignment is then used to generate a covariation model and construct a secondary structure based on infernal [229]. This results in many motifs, that can then be combined to produce the structure of more complex RNAs. The method works well with natural RNA sequences such as riboswitches where there is an abundance of sequence information throughout the RNA including the helical domains, but the presence of our fixed regions which lack covariation in the structured stems poses a problem for the workflow. The final output is often incomplete structures and a loss of sequence information (Figure A.3). Many times a parental-like fold can still be found as one of the resulting motifs, but cases where it is not found are inconclusive given that alternative folding programs (such as Sfold) show a more expected fold.

As an alternative to the CMfinder's workflow, alignments generated from a separate process can be input into the final step that generates the covariation model and secondary structure with infernal. Muscle was chosen as an orthogonal alignment program based on
its wide acceptance and its claimed ability to achieve higher accuracy and speed. Muscle generates alignments based on the full sequence and not small motifs, thereby allowing the fixed sequences to anchor the more degenerate variable regions and provide a more structured context. The output is a single motif, but this predicted structure usually agrees well with the parental fold.

All structures and folds presented here are the output of the muscle aligned CMfinder model with the exception of cdG_{17}, whose structure represented in this work was forced to that of the parental fold based on the structural probing indicating that they are likely in agreement. The de novo CMfinder workflow produces candidate motifs with the parental fold for all sequences described here, except they are often lacking in information that slightly misrepresents the sequence. The iterative nature of building a complex structure from simple motifs in CMfinder leads to false regions of low occupancy in the degenerate regions of the junction, especially when there is not a high degree of convergence, like the T-loop. In theory, the output that is treated as a final structure here is meant as a starting point for additional rounds of refinement in CMfinder, but we sought a more straight forward pipeline and thus chose the muscle workflow for these studies as it produces reasonable results with no further work to be done. Both workflows are left implemented in the script, as either or both can inform on the most likely structure for a sequence and it is likely that one method will not be ideal for all analysis.

The final step in the script uses R2R [200] at its default settings to generate the final figures shown in this work. R2R uses the structurally informed alignment from either branch of the workflow and regenerates the infernal covariation model. The model is then printed onto the secondary structures as shown herein. This makes analysis of the covariation model easy, clearly showing whether a base pair covaries and reinforces the predicted structure or
has low confidence with many unaccepted configurations. The use of R2R here is simply to visualize the covariation model and structure predicted in the previous steps, the raw output of which is usually difficult to interpret by eye.

The clusters identified for further characterization and implementation into RNA devices were the most abundant cluster in the pool that was projected to have a parental like fold. Both of these criteria are often met by the most abundant cluster overall, the naming of which herein (i.e. 1, 4, or 17) is an artifact of the initial clustering and not necessarily informative of the frequency of the cluster. For individual sequences isolated from each selection, they represented the clear dominant sequence within the cluster. Therefore the sequences used as parts for RNA devices in this study were the most abundant sequence in the most abundant cluster, often resulting in the isolated sequence being one of the more highly represented sequences for the selection as a whole. However, identification of this sequence in the absence of this analysis is difficult or impossible given that the final pool is under selected with a greater degree of diversity than normally observed.
Figure A.3: Comparison of the de novo CMfinder workflow with the simplified muscle workflow. The traditional workflow involving CMfinder produces motifs that are similar to the parental fold. However, the lack of sequence information in the constant regions hinders the motif prediction and positions in the junction region that appear to have low occupancy are actually occupied at nearly 100% upon manual sequence inspection. Global alignment with muscle is constrained by the fixed sequence and only provides a single structural candidate for manual inspection. This candidate preserves the covariation, yet is more informant on the occupancy and nucleotide distribution.