# IMPROVING THE RIBOZYME TOOLBOX: FROM STRUCTURE-FUNCTION INSIGHTS TO SYNTHETIC BIOLOGY APPLICATIONS

by

Jessica Michelle Roberts

A dissertation
submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Biomolecular Sciences
Boise State University

August 2022

© 2022

Jessica Michelle Roberts

ALL RIGHTS RESERVED

BOISE STATE UNIVERSITY GRADUATE COLLEGE

## DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Jessica Michelle Roberts

Dissertation Title: The Ribozyme Toolbox: From Structure Function Insights to Synthetic Biology Applications

Date of Final Oral Examination: 20 May 2022

The following individuals read and discussed the dissertation submitted by student Jessica Michelle Roberts, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

| | |
|---|---|
| Eric J. Hayden, Ph.D. | Chair, Supervisory Committee |
| Allan R. Albig, Ph.D. | Member, Supervisory Committee |
| Matthew L. Ferguson, Ph.D. | Member, Supervisory Committee |

The final reading approval of the dissertation was granted by Eric J Hayden, Ph.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

# DEDICATION

This work, and the years of effort contributing to it, I proudly dedicate to my precious children. Wesley Ryan Soto and Theodore Danger Soto, you both inspire me to push myself to achieve great things, and I hope that I can open many doors for you both to have a bright future and a wonderful life. Because of you, my plate may be full, but my cup is overfloweth.

# ACKNOWLEDGMENTS

This endeavor has been a long and trying one, but also a time filled with so much joy, inspiration, and growth. This was no one man show, and I wouldn't have reached this point without the support, mentorship, and comradery of so many influential people that I am fortunate to have in my life.

First and foremost, I want to acknowledge that this would not have been possible without the unwavering support of my life partner and loving husband, Ryan Soto. Thank you for putting up with me during the stressful moments, for taking care of our children and home so I could focus on science, and for so much more. Importantly, thank you for believing in me when I didn't believe in myself, and encouraging me every step of the way. Truly, I can't express enough gratitude for you. Big thanks to my sons, Wesley and Theodore, for inspiring me and challenging me to grow. Thanks for bringing me joy every day, it is a pleasure to watch you both learn and grow into such wonderful humans, I am so proud to be your mother. I am truly blessed with an amazing and beautiful family!

On that note, I also want to give a big thank you to my mom, Tina Perkins. Thank you for giving me life, and for being such a hardworking, compassionate, and fun role model. Thank you for all your support my whole life, I wouldn't have accomplished this without your love and support. Angelica, thanks for being an amazingly supportive sister, and best friend. I appreciate you always being there for me, for your great advice, and for always being down for a fun time. I also want to acknowledge the support of my brother

Jonathan, and sister-in-law and friend, Megan. Grandma Launi, thanks for all your love and support my whole life. To my aunt Riana, thanks for all your advice and encouragement, thanks for helping me fill out my first FAFSA and college application, I wouldn’t be here without you helping me take that first step. I love and appreciate you all more than I can say!

I am so grateful for the wonderful mentorship I received from many along the way that has shaped me into the scientist I have become. Most notably, I want to express my sincere gratitude to my outstanding PhD advisor, Eric Hayden. Thanks, Eric, for your support in all its forms over these past years. You have so compassionately guided me through so many challenges and obstacles, and I have learned and grown in so many ways during my time in the Hayden lab. Thank you for all your thoughtful advice and mentorship, for providing me with so many opportunities, for allowing me to have a work-life balance with my family. Thanks for putting up with my maternity leave so my family could still grow despite being a woman pursuing a career in science. And so much more! I cannot say enough wonderful things about you and what an outstanding mentor you have been to me.

I also want to sincerely thank my supportive committee members, Allan Albig, and Matthew Ferguson. You have both been wonderful mentors along the way. Allan, I really enjoyed sharing lab space with you and your group. Thanks for showing me the ropes with tissue culture, and for all your advice with cloning and all my lentiviral woes, and so many other challenges that you provided advice and assistance on. Matt, thanks for your support and mentorship in microscopy and cell line development, and for providing opportunities to participate in such cutting-edge science. I appreciate the time and effort you all have put into my growth.

I'd also like to acknowledge the comradery and support of my peers and colleagues that I have gained along the way, you have all certainly made this a more colorful and enriching experience. Sarah Kobernat, your friendship means so much to me, thanks for your daily support and advice, it has been an honor to have been on this adventure together. Clémentine Gibard, you are such an inspiring woman and scientist, I am so grateful that you came into my life and have become such a great friend, I appreciate you so much. Stephanie Hudon, thank you for believing in me and encouraging me since I was an undergrad, you have been a trusted and valued mentor, colleague, and friend. Steven Burden, thanks for challenging me to dream bigger and for providing an example that it is possible to achieve anything you set your mind to with some hard work, confidence, and perseverance. Jim Beck, this definitely wouldn't have been possible without all your hard work. Thanks so much for your help with coding and data analysis, and for your daily comradery and support. It has been a pleasure to work with you. Devin Bendixsen, thanks for all your support and mentorship in the Hayden lab and for all your collaboration in our shared projects. Thanks to my other lab members that have made this experience more enriching, Gianluca Peri, Jeremy Herrera, and Michal Matyjasik. Thanks to my colleagues in the Albig lab that provided so much advice and support along the way - Jacob Crow, Mike Detweiler, Bryce Lafoya, and others. Big thanks to all of the support and encouragement from other friends and colleagues that I have gained. Marcelo Ayllon, Rosey Whiting, Giovan Cholico, Shivakumar Rayavara, Alex Soto, Simion Dinca, Stephanie Tuft, Elise Overgaard, and more! This has been such a blessed experience because of all of you that I share this path with.

I want to acknowledge the support of our awesome Biomolecular Sciences graduate program. To our program coordinator, Beth Gee, thanks for everything you have done to support me and all the rest of us. To our director, Denise, thanks for this opportunity and for your support along the way. I would also like to thank Rebecca Hermann for the many hours she helped me sort cells, and to Abir Rahman for training me on lentivirus protocols. To Sean Howard, thanks for your last-minute assistance with microscopy.

And last, but certainly not least, I want to thank several influential mentors that I had prior to entering this PhD program. If not for your support early in my academic career, I would not be here today. Henry Charlier (AKA Dr. Picklestein), thanks for seeing me and believing in me, thanks for providing me with opportunities to experience undergraduate research, and for putting in a good word to help me get into this program. To Nicole Frank, it was your introductory biology courses at the College of Western Idaho that woke me up to the intricate molecular wonders of life and inspired me to pursue a career in this field.

I am grateful for you all, what a long strange trip it’s been!

# ABSTRACT

Self-cleaving ribozymes are a naturally occurring class of catalytically active RNA molecules which cleave their own phosphate backbone. In nature, self-cleaving ribozymes are best known for their role in processing concatamers of viral genomes into monomers during viral replication in some RNA viruses, but to a lesser degree have also been implicated in mRNA regulation and processing in bacteria and eukaryotes. In addition to their biological relevance, these RNA enzymes have been harnessed as important biomolecular tools with a variety of applications in fields such as bioengineering. Self-cleaving ribozymes are relatively small and easy to generate in the lab using common molecular biology approaches, and have therefore been accessible and well exploited model systems used to interrogate RNA sequence-structure-function relationships. Furthermore, self-cleaving ribozymes are also being implemented as parts in the development of various biomolecular tools such as biosensors and gene regulatory elements. While much progress has been made in these areas, there are still challenges associated with the performance and implementation of such tools.

The work contained in this dissertation aims to address several of these challenges and improve the ribozyme toolbox in several diverse areas. Chapter one provides an introduction to pertinent background information for this dissertation. Chapter two aims to improve the ribozyme toolbox by providing and analyzing new high-throughput sequence-structure-function data sets on five different self-cleaving ribozymes, and identifying how trends in epistasis relate to distinct structural elements. Chapter three

uses such high-throughput data to train machine learning models that accurately predict the historically difficult to predict functional effects of higher order mutations in functional RNA's. Finally, in chapter four, I developed a biologically relevant platform to study the real time performance and kinetics of self-cleaving ribozyme-based gene regulatory elements directly at the site of transcription in mammalian cells.

TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 3D | Three dimensional |
| 5’-RACE | System for rapid amplification of cDNA ends |
| BSU | Boise State University |
| cDNA | Complementary DNA |
| CPEB3 | Cytoplasmic polyadenylation element binding protein 3 |
| CRISPR | Clustered regularly interspaced short palindromic repeats |
| Cryo-EM | Cryogenic electron microscopy |
| CTD | C-terminal domain |
| DMEM | Dulbecco’s modified eagle medium |
| DNA | Deoxyribonucleic acid |
| EDTA | Ethylenediamine tetraacetic acid |
| FACS | Fluorescence activated cell sorting |
| FCCS | Fluorescence cross correlation spectroscopy |
| FRT | Flippase recognition target |
| GC | Graduate College |
| GFP | Green fluorescent protein |
| HDV | Hepatitis D virus |
| LSTM | Long short-term memory |
| $Mg^{2+}$ | Magnesium ion |
| MSE | Mean squared error |

| | |
|---|---|
| mRNA | Messenger RNA |
| NMR | Nuclear magnetic resonance |
| NUPACK | Nucleic acid package |
| OH | Hydroxyl group |
| PCR | Polymerase chain reaction |
| RNA | Ribonucleic acid |
| RT-PCR | Reverse transcription PCR |
| SELEX | Systematic evolution of ligands by exponential enrichment |
| SHAPE | Selective 2' hydroxyl acylation analyzed by primer extension |
| TDC | Thesis and Dissertation Coordinator |
| Theo | Theophylline |
| tRNA | Transfer RNA |
| UTR | Untranslated region |

# CHAPTER ONE: DISSERTATION INTRODUCTION

Ribozymes are widespread and naturally occurring catalytic RNA molecules which have diverse applications in fields such as bioengineering, origins of life and evolution, and structural biology. Various types of ribozymes exist in nature, where they play important roles in the regulation and processing of genetic information. With the exception of the ribosome, naturally occurring ribozymes catalyze site specific cleavage and/or ligation of RNA phosphodiester backbones in processes such as splicing, viral genome processing, tRNA maturation, and gene regulation. While ongoing research continues to uncover new ribozymes and their biological functions, the work contained herein will focus on harnessing small self-cleaving ribozymes as tools in molecular biology. In this chapter, I will provide some foundational background information on the discovery and distribution of self-cleaving ribozymes, trends in their structure and catalysis, and demonstrate their utility as a model system to interrogate RNA structure and function. In addition, I will highlight some relevant ways that ribozymes have been utilized in bioengineering.

## Discovery and Distribution of Self-Cleaving Ribozymes

Small self-cleaving ribozymes were first discovered in the viral genomes of plant pathogens in the 1980’s, but have since been found to be widely distributed throughout all domains of life (Buzayan et al., 1986; Perreault et al., 2011; Prody et al., 1986; Roth et al., 2014; Webb et al., 2009). The self-cleavage activity of these early discovered ribozymes facilitates the production of genomic monomers from the multimeric concatamers that are produced during the replication of circular viral genomes (rolling circle replication). Shortly after the identification of the aforementioned hammerhead and hairpin ribozymes in the tobacco ringspot virus, a self-cleaving ribozyme was discovered in the RNA genome of the human hepatitis delta virus (HDV), a satellite virus to the hepatitis b virus, where it is also responsible for the cleavage of multimeric copies of the viral genome into individual monomers (Sharmeen et al., 1988). Efforts to identify the existence of self-cleaving ribozymes in the human genome employed in vitro selection of 150 basepair genomic fragments for those that exhibited cleavage activity. These efforts identified an HDV like ribozyme located in an intron of the cytoplasmic polyadenylation binding protein 3 (CPEB3), which has since been found to be highly conserved in mammals (Bendixsen et al., 2021; Salehi-Ashtiani et al., 2006, p. 3). Recently, bioinformatic approaches have been used to identify other novel classes of self-cleaving ribozymes (Roth et al., 2014; Salehi-Ashtiani et al., 2006; Weinberg et al., 2015).

Unlike protein enzymes whose sequence and structure are fairly conserved across domains of life, most classes of ribozymes exhibit structural conservation but lack sequence conservation. Therefore, bioinformatic approaches aimed at identifying both novel and known ribozymes in genomic databases have employed searching for structural motifs and sequence co-variation rather than for ribozyme sequences. This approach has been relatively fruitful, and has added several classes of small self-cleaving ribozymes to the growing list of known naturally occurring self-cleaving ribozymes (i.e. twister, hatchet, pistol, twister sister) (Roth et al., 2014; Weinberg et al., 2015). Additionally, bioinformatics approaches have realized the widespread distribution of both HDV and hammerhead ribozyme motifs in the genomes of diverse organisms spanning across all domains of life (Hammann et al., 2012; Perreault et al., 2011; Webb et al., 2009; Webb & Lupták, 2011). The widespread distribution and biological importance of naturally occurring ribozymes has led to decades of research aimed at understanding RNA structure and catalysis, and the adoption of self-cleaving ribozymes as tools for various applications in bioengineering.

## Structure and Catalysis of Self-Cleaving Ribozymes

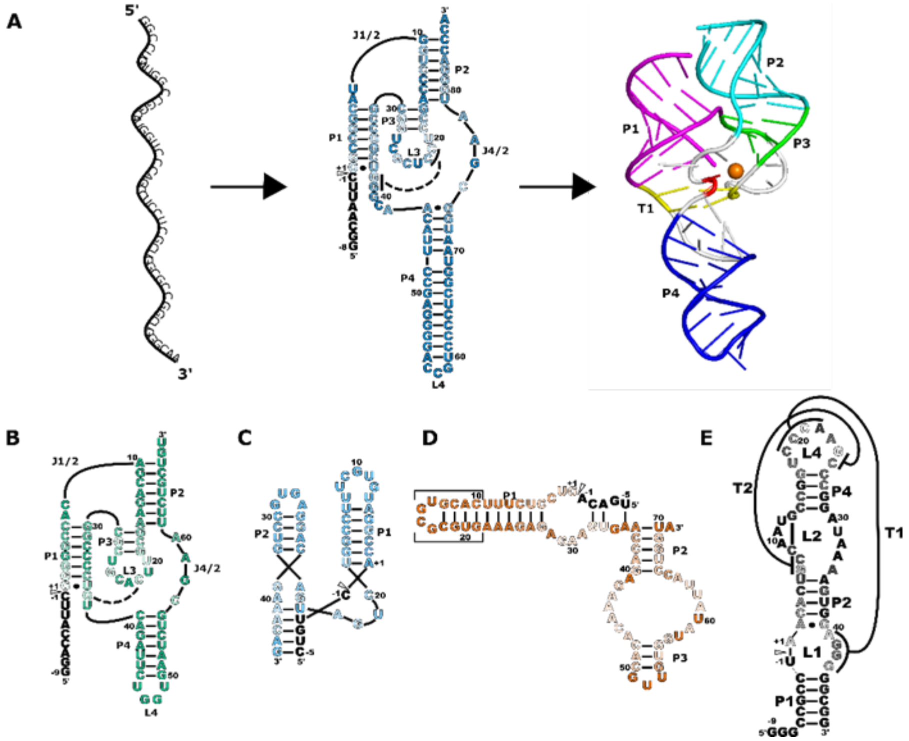


**Figure 1.1 Structures of self-cleaving ribozymes used in this work**
A) Primary (truncated), secondary, and tertiary structure of an HDV ribozyme. RNA primary sequence forms secondary structural elements such as basepaired helical regions and pseudoknots, which fold into a 3D tertiary structure composed of interactions such as helical stacking and tertiary contacts. B) Secondary structure of a CPEB3 ribozyme, C) a hammerhead ribozyme, D) a harpin ribozyme, and E) a twister ribozyme.

The self-cleaving activity of ribozymes is facilitated by the adaptation of a catalytically active secondary and tertiary structure which is dictated by the RNA sequence of the ribozyme (Figure 1). The secondary structure of RNA molecules is based on the ability of the RNA polymer to fold back on itself enabling anti-parallel base pairing interactions that form helical elements within a single molecule. Additionally, single stranded regions belonging to stem loops can participate in base-pairing interactions with other regions of the RNA molecule forming a structural element referred to as a pseudoknot. The tertiary conformation of RNA molecules is composed of coaxial stacking of helical regions, and tertiary contacts between non-basepaired regions of the RNA molecule driven by molecular interactions such as hydrogen bonding, base stacking, and the binding of metal ions, commonly $Mg^{2+}$ (Butcher & Pyle, 2011; Jimenez et al., 2015). The adoption of a catalytically active folded state of ribozymes facilitates the formation of an active site that orients nucleotides or co-factors in such a way that they can catalyze a site-specific transesterification reaction, leading to the cleavage of its own phosphodiester backbone.

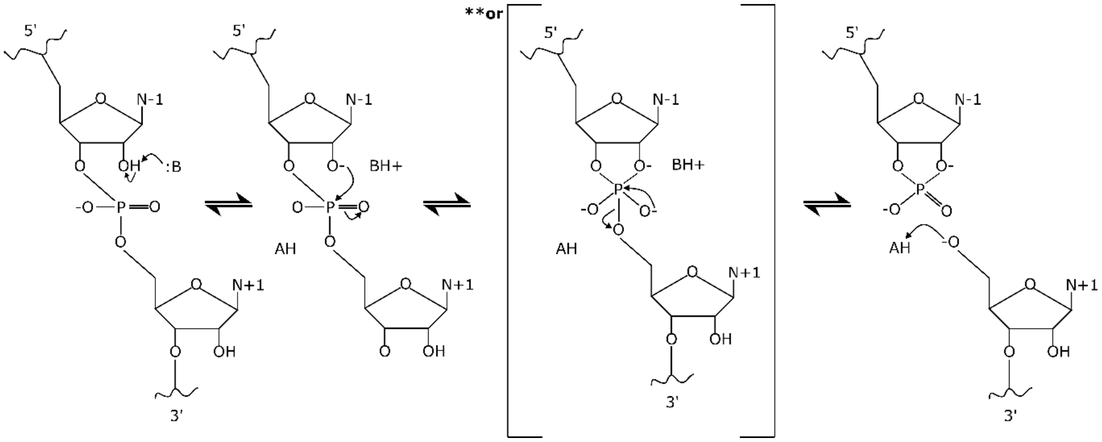


**Figure 1.2 Catalytic mechanism of self-cleaving ribozymes**

The mechanism of ribozyme self-cleavage proceeds through a trans-esterification reaction that is catalyzed by a concerted general acid-base mechanism. In this mechanism, a general base deprotonates the 2’ OH belonging to the ribose of the nucleophilic nucleotide, and a transesterification reaction proceeds, yielding a 2’-3’ cyclic phosphate on the upstream nucleotide, and an oxyanion on the 5' leaving group. A general acid then protonates the leaving group (Figure 2) (Chen et al., 2010; Fedor, 2000; Jimenez et al., 2015; Martick et al., 2008).

The ribozymes used in this dissertation

The work contained in this dissertation utilized five self-cleaving ribozymes - hammerhead, hairpin, HDV, CPEB3, and twister. While these self-cleaving ribozymes share the common catalytic mechanism described above, each achieves self-catalysis by folding into distinct three-dimensional RNA structures dictated by their unique sequences. This leads to interesting questions about sequence, structure, and function relationships that will be explored in the following chapters. Here, I will briefly summarize core features of the structure and catalysis of the ribozymes utilized in this work.

Perhaps the ribozyme whose structure and catalytic activity has been most well characterized is the hammerhead ribozyme. There are several naturally occurring permutations of the hammerhead ribozyme, referred to as type I, II, and III. The secondary structure contains three helical regions - stems I, II, and III, and are named according to which stem is the closing stem. The tertiary architecture features coaxial stacking between stems II and III, with a CUGA uridine turn orienting stem I in proximity to stem II, forming a tertiary contact that is important for efficient catalysis of the self-cleavage reaction. The three stems are branching from a 15 nucleotide catalytic core, where two guanines participate as a general acid and base. For the type III construct used in the research contained in this dissertation, G25 and G29 act as a general acid and base, respectively. The ionized N1 on G29 deprotonates the 2'OH of C-1, activating it as a nucleophile that can attack the adjacent scissile phosphate, and the 2’OH belonging to the ribose of G25 acts as a general acid, protonating the leaving group (Doudna, 1995; Martick et al., 2008; Scott et al., 2013).

The hairpin ribozyme that was discovered in the minus strand of the same tobacco ringspot virus as the hammerhead ribozyme also has several structural variants that have been characterized. Common features of the hairpin secondary structure are four helical regions, and two internal loop domains, loops A and B. In naturally occurring hairpin ribozymes, the four helices are connected via a four-way junction, whereas minimal hairpin structures that are frequently used in biochemical and structural analyses contain two stem loop domains connected via a two-way helical junction. In these minimal constructs, like the one utilized in the research contained herein, each of the two stem loop domains are bisected by the internal loops A and B, resulting in four total helical regions (Müller et al., 2012). Crystallographic, NMR, and mutational studies have elucidated critical non-Watson-Crick interactions both within loop A and loop B, and between the two loops. Hydrogen bonds between bases, as well as ribose-base hydrogen bonding interactions result in an intimate docking between loops A and B, which is critical in forming the active site of the ribozyme (Butcher et al., 1999; Cai & Tinoco, 1996; Fedor, 2000). Based on analogous structures, in the construct used in this study, a guanine at position 29 residing within loop A acts as a general base, whereas an adenine at position 59 within loop B acts as a general acid (Müller et al., 2012; Rupert & Ferré-D’Amaré, 2001).

The similarly structured HDV and CPEB3 ribozymes fold into a double nested pseudoknot conformation. The ribozymes contain five helical regions, (P1-P4, plus an additional pseudoknotted helical region, P1.1), with two coaxial stacks, where P1 and P1.1 stack on P4, and P2 stacks on P3. A single stranded region, J1/2 connects P1 and P2, and a second single stranded region, J4/2 connects P4 and P2. The active site of these ribozymes is located in the junction between P1, P1.1, and P3. Unlike hammerhead and hairpin discussed above where nucleobases act as both the general acid and base, HDV and CPEB3 both utilize a $Mg^{2+}$ ion as a general base that activates the 2'-hydroxyl of the nucleotide upstream from the cleavage site. A cytosine at position 75 and 57 (for HDV, and CPEB3, respectively) located in the single stranded J4/2 region acts as a general acid, donating a proton to stabilize the leaving group (Chen et al., 2010; Skilandat et al., 2016).

There are several permutations of the twister ribozyme that occur in nature (Eiler et al., 2014; Roth et al., 2014). The construct utilized in this work contains three stem regions (P1, P2, and P3), which are separated by two internal loops (L1 and L2). The stem loop to P4 forms two pseudoknotted tertiary contacts of opposite polarity with the two internal loops (L1 and L2). There is coaxial stacking between P1, T1, P2, and T2, with a helical twist between P1 and T1. The active site of the ribozyme is near the center of the ribozyme, along the major groove of the T1-P2 helix. Other variations of the twister ribozyme have two additional stem regions, P3, and P5, which branch out from L2, forming a four way junction (Liu et al., 2014). In the proposed mechanism for the twister ribozyme used in this work, G39 acts as the general base, deprotonating the 2'-hydroxyl of the nucleophilic oxygen. The N3 of the ribose of A1 is proposed to act as the general acid, donating a proton to the 5' oxyanion leaving group (Wilson et al., 2016).

## Self-Cleaving Ribozyme Insights and Applications

Self-cleaving ribozymes have been exploited as model systems to understand RNA structure and catalysis, and as tools for diverse applications in synthetic biology and engineering. Their small size and ease of generation via in vitro transcription using only basic molecular biology laboratory equipment makes self-cleaving ribozymes an accessible biomolecular system. As such, decades of research revolving around ribozymes has provided foundational insight into the structure and catalysis of RNA, and diverse and creative applications have emerged ranging from the development of ribozyme-based biosensors, synthetic genetic circuits and regulatory elements, with applications in medicine and biofuel production (Breaker, 2002; Ogawa & Maeda, 2008; Yokobayashi, 2019b). This section of the introduction will provide a brief background to highlight the utility of adopting self-cleaving ribozymes as model systems to gain insight into general RNA structure and biochemistry, as well as some ribozyme applications relevant to the work contained in this dissertation.

### Lessons in RNA biochemistry and structural biology

Small self-cleaving ribozymes were among some of the first RNA's whose structure was determined via X-ray crystallography, providing foundational examples of conserved RNA motifs and revealing many principles of RNA folding, ligand biding, and catalysis. Compared to proteins, RNA has unique challenges associated with crystallization. Electrostatic repulsion due to the repetitive negative charges located on their phosphate backbones interferes with crystal packing, and compared to the globular architecture in proteins, RNA tends to form more elongated structures that pack loosely into crystals. In addition, RNA molecules show structural dynamics that frequently results in misfolding, yielding non-homogenous samples (Holbrook & Kim, 1997; Ke & Doudna, 2004). These challenges discouraged initial efforts to crystallize large structured RNA's and RNA-protein complexes, making small self-cleaving ribozymes an approachable and accessible model to gain insight into features of RNA structure. Furthermore, they provided a platform for the development and refinement of methodology that has led to our ability to crystallize and resolve the structures of more complex RNA's (Ferré-D'Amaré, 2010, p. 1; Ferré-D'Amaré et al., 1998; Ferré-D'Amaré & Doudna, 2000). In addition to providing a foundational model system for RNA crystallization strategies and structure determination, self-cleaving ribozymes are being utilized as model systems to understand structural, functional, and evolutionary implications of sequence variation resulting from mutations, as well as in efforts to predict RNA structure computationally (Andronescu et al., 2005; Bendixsen et al., 2017, 2019; Miao et al., 2020)

Investigating self-cleaving ribozymes provided foundational insights into RNA catalysis. The mechanism that RNA ribozymes employ to catalyze the cleavage of phosphodiester bonds was historically controversial. Magnesium ($Mg^{2+}$) ions play promiscuous and fundamental roles in the formation and stability of RNA structures, therefore most ribozymes exhibit a magnesium dependence (Misra & Draper, 1998). This magnesium dependence understandably contributed to the prevailing view that RNA ribozymes are metalloenzymes (Dahm et al., 1993; Dahm & Uhlenbeck, 1991; Pontius et al., 1997; Pyle, 1993). Additionally, RNA biochemistry was influenced by the lessons gained from protein biochemistry. In protein enzymes, amino acids that play analogous roles in general acid-base catalysis have functional groups whose pKa's result in acidic or basic properties in the neutral environments they function in. The pKa's of the functional groups present in the nucleobases of RNA molecules indicated they should not carry a charge at neutral pH, and were therefore initially dismissed as potential participants in catalysis. Crystallization of the hammerhead ribozyme enabled the first atomic resolution view of RNA active sites (Doudna, 1995), and subsequent research confirmed that most ribozymes use general acid-base catalysis (Han & Burke, 2005; Martick et al., 2008). This demonstrates that adopting small self-cleaving ribozymes as model systems to study catalytic RNA can provide foundational knowledge on RNA structure and function.

### Self-cleaving ribozymes have been used as parts in the development of biosensors

In addition to providing a model system for understanding RNA structure and catalysis, ribozymes have been used to develop biosensors. RNA is capable of forming dynamic structures which can be modulated by the presence of other biomolecules and/or small molecules. Additionally, RNA has the potential for specific molecular recognition, and in vitro selections from randomized pools of RNA have generated a variety of RNA sequences called aptamers, which are capable of selectively binding to various compounds (Blind & Blank, 2015; Darmostuk et al., 2015). The potential for structural dynamics paired with the inherent enzymatic activity of ribozymes made them an attractive candidate for the development of allosteric biomolecular switches (Soukup & Breaker, 1999). Allostery in self-cleaving ribozymes is commonly achieved via the addition of aptamer domains in such a way that binding of the aptamer's ligand alters the overall conformation, and thus cleavage activity of the ribozyme. In this way, molecular recognition can be detected via the cleavage state of the ribozyme. Parallel and combined efforts utilizing rational design, in vitro selection, and computational design has yielded a variety of allosteric ribozymes that respond to and detect various metabolites (Breaker, 2002; Frauendorf & Jäschke, 2001; Koizumi et al., 1999; Kuwabara et al., 2000; Penchovsky, 2014). Such ribozyme-based biosensors have diverse applications in areas such as clinical diagnostics and intracellular metabolite detection.

### Applications of ribozymes in gene regulation and genetic circuit design

Another area that ribozymes have been successfully employed as tools in synthetic biology is in the development of gene regulatory elements called riboswitches. Advances in gene editing technology such as CRISPR-Cas9 have enabled the modification and engineering of biological systems with applications ranging from biofuel and drug production, to personalized gene therapies. However, gene regulatory technology lags behind gene editing technology, and there is a need for the development of robust and orthogonal avenues to externally control the expression of transgenes. To this end, synthetic RNA riboswitch-based gene regulation has become an attractive platform for the realization of protein-independent control over gene expression.

Riboswitches are naturally occurring metabolite responsive RNA regulatory elements that are typically located in the 5' or 3' untranslated regions of mRNA. They contain an aptamer domain that is capable of specifically binding to a ligand. Binding of the ligand causes a conformational shift that permeates through an adjacent expression platform in the riboswitch, which alters gene expression (Serganov & Patel, 2007; Zhang et al., 2010). There are various expression platforms, thus mechanisms that riboswitches implement in order to achieve alterations to gene expression. Common naturally occurring expression platforms fluctuate between structural conformations that either sequester or expose ribosomal binding sites, or shine-delgarno sequences. In addition, there are expression platforms where binding of the ligand causes a shifts between the presence of terminator and anti-terminator stems, or blocking of splice sites (Chang et al., 2012; Etzel & Mörl, 2017). Synthetic riboswitches have been developed using self-cleaving ribozyme expression platforms, and are commonly referred to as aptazymes. In this type of riboswitch, binding of the ligand causes a shift between a catalytically active and inactive conformation, or vice versa. Cleavage of the mRNA transcript disrupts mRNA processing and translation, resulting in a decrease in gene expression, whereas absence of self-cleavage allows downstream processing and translation to proceed largely uninhibited (Zhong et al., 2016).

Synthetic aptazyme based riboswitches are largely developed using similar rationale and approaches as aptazyme based biosensors which were discussed above. However, allosteric aptazmes selected in vitro often fail to recapitulate their activity in cellular environments, and so successful efforts have largely shifted to in vivo selection strategies (Desai & Gallivan, 2004; Michener & Smolke, 2012; Wieland et al., 2012). These efforts have yielded a plethora of synthetic riboswitches based on self-cleaving ribozymes that respond to a variety of ligands and utilize various strategies to alter gene regulation. Such synthetic riboswitches have been implemented in the external control of gene expression in bacteria, yeast, and mammalian cells (Ogawa & Maeda, 2008; Stifel et al., 2019; Townshend et al., 2015; Yokobayashi, 2019b; Zhong et al., 2016).

## Improving the Ribozyme Toolbox

Through decades of culminating research surrounding ribozyme structure, catalysis, biological function, and synthetic biology applications exciting new directions and corresponding challenges have emerged. Here, I will introduce recent advancements, current challenges, and how this dissertation work contributes to the improvement of self-cleaving ribozyme applications.

### Chapter two contributes valuable sequence-structure-function data sets

Advances in nucleic acid synthesis and next generation sequencing has opened doors for in depth exploration of the functional effects resulting from sequence variation within self-cleaving ribozymes. It is now possible to synthesize large mutagenized DNA libraries of small self-cleaving ribozymes, transcribe the libraries into RNA in-vitro, and measure the functional effects via next-generation sequencing. Efforts in this area have aimed at utilizing self-cleaving ribozymes as a model to understand the relationships between RNA sequence, structure, and function – with emphases in RNA evolution and innovation of function, as well as in the development and improvement of ribozymes with potential in various bioengineering applications (Andreasson et al., 2020; Bendixsen et al., 2019; Hayden, 2016; Kobori & Yokobayashi, 2016; Yokobayashi, 2019a).

Chapter two of this dissertation harnesses high-throughput analysis of self-cleaving ribozyme activity and trends in epistasis to elucidate structural information, and provides comprehensive double mutant data sets for exploitation in the development and refinement of computational approaches aimed at predicting the structural and functional effects of mutations to RNA. To do this, I comprehensively analyzed the self-cleavage activity of all possible single and double mutations to the hammerhead, hairpin, HDV, CPEB3, and twister ribozymes. The data is presented as heatmaps of relative activity and epistasis, and we show that features in the data correspond with known structural features, and identify trends in epistatic relationships in the context of common RNA structural elements. Furthermore, we openly share the resulting data for future use in the development and training of computational pipelines with long-term goals aimed at in situ prediction of RNA structure and function. The exploration, analysis, and sharing of these complete double mutant cycle data sets improves the ribozyme toolbox by providing open access and foundational data that can be easily exploited in a variety of applications.

### Chapter three pairs high-throughput experimental activity data with machine learning to predict the effects of higher order mutations

Efforts to harness self-cleaving ribozymes as tools such as in biosensors and riboswitches that are discussed above often rely on introducing changes to a RNA sequence in order to find a RNA molecule with the new desired function. However, the rational design of such functional molecules remains challenging because the effects of introducing multiple mutations to an RNA molecule are difficult to predict due to significant pairwise and higher-order epistasis that is observed in an RNA molecule. Recently, success in predicting the 3D structure of an RNA molecule, and even the functional effects of changes to an RNA sequence in ribozyme based gene regulatory elements has been facilitated by the use of machine learning approaches (Calonaci et al., 2020; Groher et al., 2019; Schmidt & Smolke, 2021). However, previous approaches have relied on crystallographic structural data, chemical probing data, thermodynamic calculations, and other types of data that are not often available for an RNA sequence of interest. Therefore, new types of easy to obtain training data containing information about the functional effects of mutations is expected to facilitate the use of machine learning to aid rational design of dynamic and functional RNA molecules.

Chapter three demonstrates the feasibility of such approach by using high-throughput self-cleavage activity data of mutants to a CPEB3 ribozyme to accurately predict the functional effects of higher-order mutations. Two different machine learning architectures (LSTM and Random Forest) were implemented, and their performance was compared. In addition, we explored the effects of incrementally increasing or reducing the type and size of our data sets on the accuracy of our predictions. The work presented in this chapter provides foundational proof of concept evidence that this type of high-throughput data can in fact be used alone as training data to successfully predict the effects of higher-order mutants of a self-cleaving ribozyme. Additionally, it provides important guiding insight into the amount and types of data necessary to achieve acceptably accurate predictions. This is expected to be a foundational paper in this newly emerging application of high-throughput sequencing data of RNA ribozyme reactions.

### Chapter four develops a live-cell platform to measure co-transcriptional synthetic aptazyme kinetics

Despite the exciting potential and existence of many synthetic ribozyme-based riboswitches, there have been challenges associated with utilizing these gene regulatory tools in real-world applications. One major obstacle has been that many of the available constructs exhibit high background expression and apparent ‘leakiness’ of gene expression when the switch should be in the off state. Furthermore, insufficient dynamic ranges in response to the ligand are common, and a more robust activation or repression of gene expression is still desired. Finally, synthetic riboswitches often fail to recapitulate performance when ported to an organism different from where they were selected to perform, even within the same domain (i.e. from one bacterial species to another, or yeast to mammalian). The added complexity of co-transcriptional mRNA processing (such as splicing, capping, and polyadenylation) present in higher order organisms likely contributes to these observed decreases in performance. Slow progress has been made in addressing these challenges, in large part because riboswitch performance is typically measured via bulk protein expression measurements. This approach loses important co-transcriptional kinetic and mechanistic insight that undoubtedly has a major impact on the performance of synthetic riboswitches. In order to improve synthetic riboswitch performance, it is important to fully understand ‘how’ and ‘when’ self-cleavage activity maximally influences changes in gene expression, and to elucidate such differences between aptazymes with suboptimal dynamic range and basal levels with those that exhibit superior function.

The goal of the research in chapter four is to developed a platform to measure the real-time performance and cleavage kinetics of synthetic aptazyme riboswitches in live human cells. To achieve this, I generated several human cell lines (293T, human embryonic kidney) that will enable the real-time single molecule measurements of mRNA transcripts that contain synthetic riboswitch constructs. The mRNA transcripts are fluorescently labeled to allow fluorescent detection. Each cell line contains a unique variant of an aptazyme based on the hammerhead ribozyme coupled to a theophylline aptamer. The riboswitch variants were selected for in yeast, and exhibit variation in both their basal expression levels as well as their dynamic range of influence over protein expression in response to their ligand, theophylline. This will allow us to identify key differences in cleavage kinetics that contribute to both basal level expression and dynamic ranges of performance, guiding ongoing efforts to design and select improved synthetic riboswitches. Therefore, this work will aid in the improvement of synthetic ribozyme-based gene regulatory tools by providing an avenue to remove the black box surrounding gene regulatory mechanisms and kinetics associated with bulk protein read-out of synthetic riboswitch performance.

Taken together, this dissertation contributes to the improvement of the ribozyme toolbox by providing comprehensive sequence-structure-function data sets on five distinct self-cleaving ribozymes, and by providing a platform to measure synthetic riboswitch performance in live cells. Together, these efforts will contribute to our ability to computationally predict the structural and functional effects of mutations to RNA in general, as well as provide a platform to gain guiding insight to improve the performance of ribozyme-based tools in bioengineering. The work contained herein lays a necessary foundation that will undoubtedly lead to grander realizations of machine learning facilitated prediction of RNA structure and function, and the improvement and implementation of synthetic regulatory RNA in applications ranging from biofuel production, environmental remediation, and personalized medicine.

## References

Andreasson, J. O. L., Savinov, A., Block, S. M., & Greenleaf, W. J. (2020). Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-15540-1

Andronescu, M., Zhang, Z. C., & Condon, A. (2005). Secondary Structure Prediction of Interacting RNA Molecules. *Journal of Molecular Biology*, *345*(5), 987–1001. https://doi.org/10.1016/j.jmb.2004.10.082

Bendixsen, D. P., Collet, J., Østman, B., & Hayden, E. J. (2019). Genotype network intersections promote evolutionary innovation. *PLOS Biology*, *17*(5), e3000300. https://doi.org/10.1371/journal.pbio.3000300

Bendixsen, D. P., Østman, B., & Hayden, E. J. (2017). Negative Epistasis in Experimental RNA Fitness Landscapes. *Journal of Molecular Evolution*, *85*(5), 159–168. https://doi.org/10.1007/s00239-017-9817-5

Bendixsen, D. P., Pollock, T. B., Peri, G., & Hayden, E. J. (2021). Experimental Resurrection of Ancestral Mammalian CPEB3 Ribozymes Reveals Deep Functional Conservation. *Molecular Biology and Evolution*, *38*(7), 2843–2853. https://doi.org/10.1093/molbev/msab074

Blind, M., & Blank, M. (2015). Aptamer Selection Technology and Recent Advances. *Molecular Therapy - Nucleic Acids*, *4*, e223. https://doi.org/10.1038/mtna.2014.74

Breaker, R. R. (2002). Engineered allosteric ribozymes as biosensor components. *Current Opinion in Biotechnology*, *13*(1), 31–39. https://doi.org/10.1016/S0958-1669(02)00281-1

Butcher, S. E., Allain, F. H.-T., & Feigon, J. (1999). Solution structure of the loop B domain from the hairpin ribozyme. *Nature Structural Biology*, *6*(3), 212.

Butcher, S. E., & Pyle, A. M. (2011). The Molecular Interactions That Stabilize RNA Tertiary Structure: RNA Motifs, Patterns, and Networks. *Accounts of Chemical Research*, *44*(12), 1302–1311. https://doi.org/10.1021/ar200098t

Buzayan, J. M., Gerlach, W. L., & Bruening, G. (1986). Non-enzymatic cleavage and ligation of RNAs complementary to a plant virus satellite RNA. *Nature*, *323*(6086), 349–353. https://doi.org/10.1038/323349a0

Cai, Z., & Tinoco, I. (1996). Solution Structure of Loop A from the Hairpin Ribozyme from Tobacco Ringspot Virus Satellite. *Biochemistry*, *35*(19), 6026–6036. https://doi.org/10.1021/bi952985g

Calonaci, N., Jones, A., Cuturello, F., Sattler, M., & Bussi, G. (2020). Machine learning a model for RNA structure prediction. *NAR Genomics and Bioinformatics*, *2*(4). https://doi.org/10.1093/nargab/lqaa090

Chang, A. L., Wolf, J. J., & Smolke, C. D. (2012). Synthetic RNA switches as a tool for temporal and spatial control over gene expression. *Current Opinion in Biotechnology*, *23*(5), 679–688. https://doi.org/10.1016/j.copbio.2012.01.005

Chen, J.-H., Yajima, R., Chadalavada, D. M., Chase, E., Bevilacqua, P. C., & Golden, B. L. (2010). A 1.9 Å Crystal Structure of the HDV Ribozyme Precleavage Suggests both Lewis Acid and General Acid Mechanisms Contribute to Phosphodiester Cleavage. *Biochemistry*, *49*(31), 6508–6518. https://doi.org/10.1021/bi100670p

Dahm, S. C., Derrick, W. B., & Uhlenbeck, O. C. (1993). Evidence for the role of solvated metal hydroxide in the hammerhead cleavage mechanism. *Biochemistry*, *32*(48), 13040–13045. https://doi.org/10.1021/bi00211a013

Dahm, S. C., & Uhlenbeck, O. C. (1991). Role of divalent metal ions in the hammerhead RNA cleavage reaction. *Biochemistry*, *30*(39), 9464–9469. https://doi.org/10.1021/bi00103a011

Darmostuk, M., Rimpelova, S., Gbelcova, H., & Ruml, T. (2015). Current approaches in SELEX: An update to aptamer selection technology. *Biotechnology Advances*, *33*(6, Part 2), 1141–1161. https://doi.org/10.1016/j.biotechadv.2015.02.008

Desai, S. K., & Gallivan, J. P. (2004). Genetic Screens and Selections for Small Molecules Based on a Synthetic Riboswitch That Activates Protein Translation. *Journal of the American Chemical Society*, *126*(41), 13247–13254. https://doi.org/10.1021/ja048634j

Doudna, J. A. (1995). Hammerhead ribozyme structure: U-turn for RNA structural biology. *Structure*, *3*(8), 747–750. https://doi.org/10.1016/S0969-2126(01)00208-8

Eiler, D., Wang, J., & Steitz, T. A. (2014). Structural basis for the fast self-cleavage reaction catalyzed by the twister ribozyme. *Proceedings of the National Academy of Sciences*, *111*(36), 13028–13033. https://doi.org/10.1073/pnas.1414571111

Etzel, M., & Mörl, M. (2017). Synthetic Riboswitches: From Plug and Pray toward Plug and Play. *Biochemistry*, *56*(9), 1181–1198. https://doi.org/10.1021/acs.biochem.6b01218

Fedor, M. J. (2000). Structure and function of the hairpin ribozyme. *Journal of Molecular Biology*, *297*(2), 269–291. https://doi.org/10.1006/jmbi.2000.3560

Ferré-D'Amaré, A. R. (2010). Use of the spliceosomal protein U1A to facilitate crystallization and structure determination of complex RNAs. *Methods*, *52*(2), 159–167. https://doi.org/10.1016/j.ymeth.2010.06.008

Ferré-D'Amaré, A. R., & Doudna, J. A. (2000). Crystallization and structure determination of a hepatitis delta virus ribozyme: Use of the RNA-binding protein U1A as a crystallization module11Edited by D. C. Rees. *Journal of Molecular Biology*, *295*(3), 541–556. https://doi.org/10.1006/jmbi.1999.3398

Ferré-D'Amaré, A. R., Zhou, K., & Doudna, J. A. (1998). A general module for RNA crystallization. *Journal of Molecular Biology*, *279*(3), 621–631. https://doi.org/10.1006/jmbi.1998.1789

Frauendorf, C., & Jäschke, A. (2001). Detection of small organic analytes by fluorescing molecular switches. *Bioorganic & Medicinal Chemistry*, *9*(10), 2521–2524. https://doi.org/10.1016/S0968-0896(01)00027-X

Groher, A.-C., Jager, S., Schneider, C., Groher, F., Hamacher, K., & Suess, B. (2019). Tuning the Performance of Synthetic Riboswitches using Machine Learning. *ACS Synthetic Biology*, *8*(1), 34–44. https://doi.org/10.1021/acssynbio.8b00207

Hammann, C., Luptak, A., Perreault, J., & Peña, M. de la. (2012). The ubiquitous hammerhead ribozyme. *RNA*, *18*(5), 871–885. https://doi.org/10.1261/rna.031401.111

Han, J., & Burke, J. M. (2005). Model for General Acid−Base Catalysis by the Hammerhead Ribozyme: PH−Activity Relationships of G8 and G12 Variants at the Putative Active Site. *Biochemistry*, *44*(21), 7864–7870. https://doi.org/10.1021/bi047941z

Hayden, E. J. (2016). Empirical analysis of RNA robustness and evolution using high-throughput sequencing of ribozyme reactions. *Methods*, *106*, 97–104. https://doi.org/10.1016/j.ymeth.2016.05.014

Holbrook, S. R., & Kim, S.-H. (1997). RNA crystallography. *Biopolymers*, *44*(1), 3–21. https://doi.org/10.1002/(SICI)1097-0282(1997)44:1<3::AID-BIP2>3.0.CO;2-Z

Jimenez, R. M., Polanco, J. A., & Lupták, A. (2015). Chemistry and Biology of Self-Cleaving Ribozymes. *Trends in Biochemical Sciences*, *40*(11), 648–661. https://doi.org/10.1016/j.tibs.2015.09.001

Ke, A., & Doudna, J. A. (2004). Crystallization of RNA and RNA–protein complexes. *Methods*, *34*(3), 408–414. https://doi.org/10.1016/j.ymeth.2004.03.027

Kobori, S., & Yokobayashi, Y. (2016). High-Throughput Mutational Analysis of a Twister Ribozyme. *Angewandte Chemie - International Edition*, *55*(35), 10354–10357. https://doi.org/10.1002/anie.201605470

Koizumi, M., Soukup, G. A., Kerr, J. N. Q., & Breaker, R. R. (1999). Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. *Nature Structural Biology*, *6*(11), 1062.

Kuwabara, T., Warashina, M., & Taira, K. (2000). Allosterically controllable ribozymes with biosensor functions. *Current Opinion in Chemical Biology*, *4*(6), 669–677. https://doi.org/10.1016/S1367-5931(00)00150-2

Liu, Y., Wilson, T. J., McPhee, S. A., & Lilley, D. M. J. (2014). Crystal structure and mechanistic investigation of the twister ribozyme. *Nature Chemical Biology*, *10*(9), 739–744. https://doi.org/10.1038/nchembio.1587

Martick, M., Lee, T.-S., York, D. M., & Scott, W. G. (2008). Solvent Structure and Hammerhead Ribozyme Catalysis. *Chemistry & Biology*, *15*(4), 332–342. https://doi.org/10.1016/j.chembiol.2008.03.010

Miao, Z., Adamiak, R. W., Antczak, M., Boniecki, M. J., Bujnicki, J., Chen, S.-J., Cheng, C. Y., Cheng, Y., Chou, F.-C., Das, R., Dokholyan, N. V., Ding, F., Geniesse, C., Jiang, Y., Joshi, A., Krokhotin, A., Magnus, M., Mailhot, O., Major, F., … Westhof, E. (2020). RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA*, *26*(8), 982–995. https://doi.org/10.1261/rna.075341.120

Michener, J. K., & Smolke, C. D. (2012). High-throughput enzyme evolution in Saccharomyces cerevisiae using a synthetic RNA switch. *Metabolic Engineering*, *14*(4), 306–316. https://doi.org/10.1016/j.ymben.2012.04.004

Misra, V. K., & Draper, D. E. (1998). On the role of magnesium ions in RNA stability. *Biopolymers*, *48*(2–3), 113–135. https://doi.org/10.1002/(SICI)1097-0282(1998)48:2<113::AID-BIP3>3.0.CO;2-Y

Müller, S., Appel, B., Krellenberg, T., & Petkovic, S. (2012). The many faces of the hairpin ribozyme: Structural and functional variants of a small catalytic rna. *IUBMB Life*, *64*(1), 36–47. https://doi.org/10.1002/iub.575

Ogawa, A., & Maeda, M. (2008). An Artificial Aptazyme-Based Riboswitch and its Cascading System in E. coli. *ChemBioChem*, *9*(2), 206–209. https://doi.org/10.1002/cbic.200700478

Penchovsky, R. (2014). Computational design of allosteric ribozymes as molecular biosensors. *Biotechnology Advances*, *32*(5), 1015–1027. https://doi.org/10.1016/j.biotechadv.2014.05.005

Perreault, J., Weinberg, Z., Roth, A., Popescu, O., Chartrand, P., Ferbeyre, G., & Breaker, R. R. (2011). Identification of Hammerhead Ribozymes in All Domains of Life Reveals Novel Structural Variations. *PLOS Computational Biology*, *7*(5), e1002031. https://doi.org/10.1371/journal.pcbi.1002031

Pontius, B. W., Lott, W. B., & von Hippel, P. H. (1997). Observations on catalysis by hammerhead ribozymes are consistent with a two-divalent-metal-ion mechanism. *Proceedings of the National Academy of Sciences*, *94*(6), 2290–2294. https://doi.org/10.1073/pnas.94.6.2290

Prody, G. A., Bakos, J. T., Buzayan, J. M., Schneider, I. R., & Bruening, G. (1986). Autolytic Processing of Dimeric Plant Virus Satellite RNA. *Science*, *231*(4745), 1577–1580.

Pyle, A. M. (1993). Ribozymes: A distinct class of metalloenzymes. *Science*, *261*(5122), 709.

Roth, A., Weinberg, Z., Chen, A. G. Y., Kim, P. B., Ames, T. D., & Breaker, R. R. (2014). A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nature Chemical Biology*, *10*(1), 56–60. https://doi.org/10.1038/nchembio.1386

Rupert, P. B., & Ferré-D'Amaré, A. R. (2001). Crystal structure of a hairpin ribozyme–inhibitor complex with implications for catalysis. *Nature*, *410*(6830), 780–786. https://doi.org/10.1038/35071009

Salehi-Ashtiani, K., Lupták, A., Litovchick, A., & Szostak, J. W. (2006). A Genomewide Search for Ribozymes Reveals an HDV-like Sequence in the Human CPEB3 Gene. *Science*, *313*(5794), 1788–1792.

Schmidt, C. M., & Smolke, C. D. (2021). A convolutional neural network for the prediction and forward design of ribozyme-based gene-control elements. *ELife*, *10*, 1–27. https://doi.org/10.7554/ELIFE.59697

Scott, W. G., Horan, L. H., & Martick, M. (2013). The Hammerhead Ribozyme: Structure, Catalysis and Gene Regulation. *Progress in Molecular Biology and Translational Science*, *120*, 1–23. https://doi.org/10.1016/B978-0-12-381286-5.00001-9

Serganov, A., & Patel, D. J. (2007). Ribozymes, riboswitches and beyond: Regulation of gene expression without proteins. *Nature Reviews. Genetics*, *8*(10), 776–790. https://doi.org/10.1038/nrg2172

Sharmeen, L., Kuo, M. Y., Dinter-Gottlieb, G., & Taylor, J. (1988). Antigenomic RNA of human hepatitis delta virus can undergo self-cleavage. *Journal of Virology*. https://doi.org/10.1128/jvi.62.8.2674-2679.1988

Skilandat, M., Rowinska-Zyrek, M., & Sigel, R. K. O. (2016). Secondary structure confirmation and localization of Mg2+ ions in the mammalian CPEB3 ribozyme. *Rna*, *22*(5), 750–763. https://doi.org/10.1261/rna.053843.115

Soukup, G. A., & Breaker, R. R. (1999). Nucleic acid molecular switches. *Trends in Biotechnology*, *17*(12), 469–476. https://doi.org/10.1016/S0167-7799(99)01383-9

Stifel, J., Spöring, M., & Hartig, J. S. (2019). Expanding the toolbox of synthetic riboswitches with guanine-dependent aptazymes. *Synthetic Biology*, *4*(1). https://doi.org/10.1093/synbio/ysy022

Townshend, B., Kennedy, A. B., Xiang, J. S., & Smolke, C. D. (2015). High-throughput cellular RNA device engineering. *Nature Methods*, *advance online publication*. https://doi.org/10.1038/nmeth.3486

Webb, C.-H. T., & Lupták, A. (2011). HDV-like self-cleaving ribozymes. *RNA Biology*, *8*(5), 719–727. https://doi.org/10.4161/rna.8.5.16226

Webb, C.-H. T., Riccitelli, N. J., Ruminski, D. J., & Lupták, A. (2009). Widespread Occurrence of Self-Cleaving Ribozymes. *Science*, *326*(5955), 953–953. https://doi.org/10.1126/science.1178084

Weinberg, Z., Kim, P. B., Chen, T. H., Li, S., Harris, K. A., Lünse, C. E., & Breaker, R. R. (2015). New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nature Chemical Biology*, *11*(8), 606–610. https://doi.org/10.1038/nchembio.1846

Wieland, M., Ausländer, D., & Fussenegger, M. (2012). Engineering of ribozyme-based riboswitches for mammalian cells. *Methods*, *56*(3), 351–357. https://doi.org/10.1016/j.ymeth.2012.01.005

Wilson, T. J., Liu, Y., Domnick, C., Kath-Schorr, S., & Lilley, D. M. J. (2016). The Novel Chemical Mechanism of the Twister Ribozyme. *Journal of the American Chemical Society*, *138*(19), 6151–6162. https://doi.org/10.1021/jacs.5b11791

Yokobayashi, Y. (2019a). Applications of high-throughput sequencing to analyze and engineer ribozymes. *Methods*, *161*, 41–45. https://doi.org/10.1016/j.ymeth.2019.02.001

Yokobayashi, Y. (2019b). Aptamer-based and aptazyme-based riboswitches in mammalian cells. *Current Opinion in Chemical Biology*, *52*, 72–78. https://doi.org/10.1016/j.cbpa.2019.05.018

Zhang, J., Lau, M. W., & Ferré-D'Amaré, A. R. (2010). Ribozymes and Riboswitches: Modulation of RNA Function by Small Molecules. *Biochemistry*, *49*(43), 9123–9131. https://doi.org/10.1021/bi1012645

Zhong, G., Wang, H., Bailey, C. C., Gao, G., & Farzan, M. (2016). Rational design of aptazyme riboswitches for efficient control of gene expression in mammalian cells. *ELife*, *5*, e18858. https://doi.org/10.7554/eLife.18858

# CHAPTER TWO: RNA SEQUENCE TO STRUCTURE ANALYSIS FROM COMPREHENSIVE PARWISE MUTAGENESIS OF MULTIPLE SELF-CLEAVING RIBOZYMES

Jessica M. Roberts[1], Jim Beck[2], Tanner B. Pollock[3], Devin P. Bendixsen[1*] and Eric J. Hayden[1,2,3]

[1]Biomolecular Sciences Graduate Programs, Boise State University, Boise, ID, USA.
[2]Computing PhD Program, Boise State University, Boise, ID, USA.
[3]Department of Biological Science, Boise State University, Boise, ID, USA.
*Current address: Department of Zoology, Stockholm University, Sweden

## Abstract

Self-cleaving ribozymes are RNA molecules that catalyze the cleavage of their own phosphodiester backbones. These ribozymes are found in all domains of life and are also a tool for biotechnical and synthetic biology applications. Self-cleaving ribozymes are also an important model of sequence to function relationships for RNA because their small size simplifies synthesis of genetic variants and self-cleaving activity is an accessible readout of the functional consequence of the mutation. Here we used a high-throughput experimental approach to determine the relative activity for every possible single and double mutant of five self-cleaving ribozymes. From this data, we comprehensively identified non-additive effects between pairs of mutations (epistasis) for all five ribozymes. We analyzed how changes in activity and trends in epistasis map to

the ribozyme structures. The variety of structures studied provided opportunities to observe several examples of common structural elements, and the data was collected under identical experimental conditions to enable direct comparison. Heat-map based visualization of the data revealed patterns indicating structural features of the ribozymes including paired regions, unpaired loops, non-canonical structures and tertiary structural contacts. The data also revealed signatures of functionally critical nucleotides involved in catalysis. The results demonstrate that the data sets provide structural information similar to chemical or enzymatic probing experiments, but with additional quantitative functional information. The large-scale data sets can be used for models predicting structure and function and for efforts to engineer self-cleaving ribozymes.

## Introduction

Challenges with predicting the functional effects of changing an RNA sequence continues to limit the study and design of RNA molecules. Recently, machine learning approaches have made considerable advancements in predicting an RNA structure from a sequence. However, these approaches rely heavily on crystal structures of RNA molecules and sequence conservation of homologs, both of which are limited for RNA molecules compared to proteins (Calonaci et al., 2020; Townshend et al., 2021). In addition, describing an RNA molecule as a single structure can be inaccurate, and regulatory elements such as riboswitches demonstrate the importance of an ensemble of structures for an RNA function. It is unclear that predictions based on individual structures alone will be able to predict functional effects of mutations with the precision needed for many biotechnical and synthetic biology applications, or to predict disease-associated mutations in RNA molecules (Halvorsen et al., 2010). This suggests that new experimental data types might be important for understanding, designing, and manipulating the transcriptome.

Self-cleaving ribozymes provide a useful model to study sequence-structure-function relationships in RNA molecules. Self-cleaving ribozymes are catalytic RNA molecules that cleave their own phosphodiester backbone. They were first discovered in viruses and viroids, but numerous families of self-cleaving ribozymes have since been discovered in all domains of life (Prody et al., 1986). The CPEB3 ribozyme, for example, was discovered in the human genome and found to be highly conserved in mammals (Bendixsen et al., 2021, p. 3; Salehi-Ashtiani et al., 2006). Other self-cleaving ribozymes, such as the hammerhead and twister ribozymes, are found broadly distributed across eukaryotic and prokaryotic genomes (Perreault et al., 2011; Roth et al., 2014). The biological roles of ribozymes in different genomes and different genetic contexts remain an active area of investigation (Jimenez et al., 2015). In addition to being widespread across the tree of life, self-cleaving ribozymes have also been used for several bioengineering applications (Liang et al., 2011; Peng et al., 2021; Wei and Smolke, 2015; Zhong et al., 2016). For example, self-cleaving ribozymes are being combined with aptamers to develop synthetic gene regulatory devices, which have biotechnical and biomedical applications where ligand dependent control of gene expression is desired (Kobori et al., 2017, 2015; Stifel et al., 2019; Townshend et al., 2015).

The testing of mutational effects in ribozyme sequences has been accelerated by high-throughput experimental approaches. Most self-cleaving ribozymes are fairly small (<200 nt) and genetic variants can be made by chemical synthesis of a single DNA oligonucleotide that is then used as a template for in vitro transcription. The self-cleavage activity of the ribozyme requires a precise three-dimensional structure, and therefore activity can be used as a sensitive indirect readout of native structure. Mutations that disrupt the native structure are detected as reduced activity compared to the unmutated “wild-type” ribozyme. Several methods have been developed to enable the detection of ribozyme function by high-throughput sequencing of biochemical reactions (Bendixsen et al., 2019; Hayden, 2016; Kobori and Yokobayashi, 2016; Shen et al., 2021). For self-cleaving ribozymes, each read from the data reports both the mutations and whether or not that molecule was reacted (cleaved) or unreacted (uncleaved). Therefore, high-throughput sequencing allows numerous genetic variants to be pooled together and still observed hundreds to thousands of times in the data. This provides confidence in the fraction cleaved for each genetic variant in a given experiment, and genetic variants are compared to determine relative activity. Importantly, the data is internally controlled because both reacted and unreacted molecules are observed, which controls for differences in their abundance due to synthesis steps (chemical DNA synthesis, transcription, reverse-transcription, PCR).

A common approach to confirm structural interactions in RNA and proteins is through analysis of pairs of mutations (Dutheil et al., 2010; Olson et al., 2014). In this context, it can be useful to calculate pairwise epistasis, which measures deviations in the mutational effects of double mutants relative to the effects of each individual mutation (assuming an additive model of mutational effects). For example, in the case of a base-pair, each single mutation would disrupt the base-pairing interaction, destabilizing the catalytically active RNA structure and reducing activity. However, if two mutants together restore a base-pair, the relative activity of the double mutant would have much higher activity than expected from the additive effects of the individual mutations (positive epistasis). In contrast to paired nucleotides, double mutants at non-paired nucleotides tend to have a more reduced activity than expected from each individual mutation (negative epistasis) (Bendixsen et al., 2017; Li et al., 2016). In the case of two mutations that create a different base pair (i.e. G-C to A-U), it is known that the stacking with neighboring base pairs is also structurally important, and some base pair substitutions will not be equivalent in a given structural context. This creates a range of possible epistatic effects even for two mutations at paired nucleotide positions. In addition, some non-canonical base interactions within tertiary contacts may also show epistasis even when they do not involve Watson-Crick or GU wobble base pairing interactions. Nevertheless, the propensity for positive epistasis between physically interacting nucleotides suggests that a comprehensive evaluation of pairwise mutational effects should contain considerable structural information.

Here, we report comprehensive analysis of mutational effects for all single and double mutants for five different self-cleaving ribozymes. Relative activity effects of all single and double mutations were determined by high-throughput sequencing of co-transcriptional self-cleavage reactions, and this data was used to calculate epistasis between pairs of mutations. The ribozymes studied include a mammalian CPEB3 ribozyme, a Hepatitis Delta Virus (HDV) ribozyme, a twister ribozyme from *Oryza sativa*, a hairpin ribozyme derived from the satellite RNA from tobacco ringspot virus, and a hammerhead ribozyme (Bendixsen et al., 2021; Burke and Greathouse, 2005; Chadalavada et al., 2007; Liu et al., 2014; Müller et al., 2012). For each reference ribozyme, a single DNA oligo template library was synthesized with 97% wild-type nucleotides at each position, and 1% of each of the three other nucleotides. This mutagenesis strategy was expected to produce all possible single and double mutants, as well as a random sampling of combinations of three or more mutations. The mutagenized templates were transcribed in vitro, all under identical conditions, where active ribozymes had the opportunity to self-cleave co-transcriptionally. All ribozyme constructs studied cleave near the 5'-end of the RNA, and a template switching reverse transcription protocol was used to append a common primer binding site to both cleaved and uncleaved molecules. Subsequently, low cycle PCR was used to add indexed Illumina adapters for high-throughput sequencing. Each mutagenized ribozyme template was transcribed separately and in triplicate, and amplified with unique indexes so that all replicates could be pooled and sequenced together on an Illumina sequencer. The sequencing data was then used to count the number of times each unique sequence was observed as cleaved or uncleaved, and this data was used to calculate the fraction cleaved. The fraction cleaved of single and double mutants was normalized to the unmutated reference sequence to determine relative activity. The relative activity values of the single and double mutants were used to calculate all possible pairwise epistatic interactions in all five ribozymes. We mapped epistasis values to each ribozyme structure to evaluate correlations between structural elements and patterns of pairwise epistasis values. The results indicated that structural features of the ribozymes are revealed in the data, suggesting that these data sets will be useful for developing models for predicting sequence-structure-function relationships in RNA molecules.

## Results and Discussion

### Evaluation of read depth and mutational coverage

The accuracy of our relative activity measurements depends on the number of reads we observe that map to each unique ribozyme sequence (read depth). Each reference ribozyme has a different nucleotide length resulting in different numbers of possible single and double mutants. In addition, the pooling of experimental replicates for sequencing does not result in equal mixtures of each replicate. In order to determine read depth, we mapped reads to the reference sequences and counted the number of reads that matched each ribozyme, while allowing for 1 or 2 mutations. We observed every single and double mutant for all ribozymes in each replicate, indicating 100% coverage of these mutant classes for all of our data sets. The distributions of observations for each single and double mutant of each ribozyme are shown in Supplementary Figure 1. The HDV data showed the lowest depth, possibly because it is a larger ribozyme (87 nt), and fewer reads mapped to the single and double mutants (Table 1). Nevertheless, from this analysis we conclude that the data contains complete coverage of all single and double mutants and ample read depth for all five ribozymes.

### Epistatic effects in paired nucleotide positions show stability-dependent signatures

In order to evaluate how the effects of mutations mapped to the ribozyme structures, we plotted the relative activity values as heat maps (Figures 1-5). We then used this data to calculate epistasis between pairs of mutations. We first inspected nucleotide positions known to be involved in base-paired regions of the secondary structure of each ribozyme. In this heatmap layout, many paired regions showed an anti-diagonal line of high activity double mutant variants with strong positive epistasis

(Figures 1-5, insets). In addition, pairs of mutations off the anti-diagonal tended to show negative or non-positive epistasis. Pseudoknot elements that involve Watson-Crick base pairs also showed this pattern, including the single base pair T1 element in CPEB3 (Figure 1) and the two base pair T1 element in HDV (Figure 2). The layout of mutations in the heatmap places paired nucleotide positions along the anti-diagonal and compensatory double mutants that change one Watson-Crick base pair to another are found on this anti-diagonal. Individual mutations that break a base pair will often reduce ribozyme activity, but the activity can be restored by a second compensatory mutation resulting in positive epistasis. In contrast, double mutants off-diagonal usually disrupt two base pairs (unless they result in a GU wobble base pair). It is expected that breaking two base pairs in the same paired region would be more deleterious to ribozyme activity than breaking one base pair, but it appears that two non-compensatory mutations in the same paired region are more deleterious than expected from an additive assumption, and frequently create negative epistasis off-diagonal within paired regions.

To quantify the observed difference in epistasis between nucleotide positions that form a base pair and two that do not, we plotted the distribution of epistasis values for double mutants on and off the anti-diagonal within the paired regions of each ribozyme. Statistical analysis indicated that the distributions were significantly different ($p<0.001$, Mann-Whitney U-test), and the epistasis values between paired nucleotide positions (on-diagonal) were consistently more positive than two mutations in positions that are not directly base paired (off-diagonal). This analysis was consistent for every individual paired region in each ribozyme (Figures 1-5, panel C). This pattern of epistasis in paired regions demonstrates the utility of comprehensive double-mutant activity data for identifying base paired regions in RNA structures.

It is interesting to note that the magnitude of the difference in the distributions of epistasis values for double mutants at paired and non-paired positions was different for different paired regions (Supplementary Figure 2). Specifically, short paired elements with fewer base pairs seemed to show large differences in the distributions of epistatic effects for paired and unpaired positions, while longer paired elements showed small differences in these distributions. For example, the short P3 (3 bp) in CPEB3 and HDV, and T1 (4 bp) in the twister ribozyme showed very large differences between the distributions of epistasis values at paired versus non-paired positions. These small regions are highly sensitive to mutations, and most pairs of mutations within this region result in almost no detectable activity except when they create a different Watson-Crick base pair (Figures 1-5). These structural elements have positive epistasis along the anti-diagonal, and negative epistasis off diagonal, resulting in large differences between the distributions of epistasis (Supplementary Figure 2). In contrast, the P4 stem in HDV has the most base pairs of any paired region in this data set (14), and losing one of these base pairs was not deleterious to riboyzme activity in our experiments (Figure 2). Because the single mutations had little effects on the self-cleavage activity, a compensatory mutation restoring a base pair did not result in positive epistasis (Figure 2). Futher, only weak negative epistasis is observed off-diagonal indicating that the loss of two base pairs in P4 was somewhat tollerated compared to shorter paired regions. The distributions for epistatis for paired and unpaired positions in P4 of HDV show only a small difference (Supplementary Figure 2). Together, the differences between epistasis in short and long base paired regions suggests that the thermodynamic stability of each paired region is important for the observed activity differences contributing to epistasis, which might ultimately affect the utility of this data for identifying paired regions in RNA structures.

In order to quantify the influence of thermodynamic stability on epistasis in different paired regions, we calculated the minimum free energy for each paired region and compared mutational effects. We split each paired region into two separate RNA sequences that contained only the base paired nucleotides and used nearest neighbor rules to calculate the minimum free energy of their interaction (NUPACK). This approach neglects thermodynamic contributions from terminal loops, but allowed for a consistent approach to compare internal and terminal paired regions. We found a significant negative correlation between the median deleterious effects of single mutations and the minimum free energy of the paired regions (Supplementary Figure 3). This analysis indicates that more stable structural elements may be harder to identify from epistatic effects. However, it is possible that more stable elements would show stronger epistasis under different experimental conditions, such as different temperatures or magnesium concentrations (Peri et al., 2022).

### Catalytic residues do not have any high-activity mutants, and do not exhibit epistasis

Self-cleaving ribozymes often utilize a concerted acid base catalysis mechanism where specific nucleobases act as proton donors (acid) or acceptors (base) (Jimenez et al., 2015), and mutations at these positions abolish activity. Analyzing the effects of individual mutations will not distinguish catalytic nucleotides from structurally important nucleotides. Comprehensive pairwise mutations, on the other hand, can potentially distinguish between structurally important nucleotides involved in paired regions that show positive epistasis from compensatory effects. The catalytic cytosines of the CPEB3 (C57) and HDV (C75) act as proton donors due to perturbed pKa values (Nakano, 2000; Skilandat et al., 2016). For the twister ribozyme (Figure 3) the guanosine at position G39 acts as a general base, and the adenosine at position A1 acts as a general acid (Wilson et al., 2016). The catalytic nucleotides for the Hammerhead ribozyme (Figure 5) are the Guanosines located at positions G25 and G39 (Scott et al., 2013). The hairpin ribozyme (Figure 4) contains catalytic nucleotides at positions G29 and A59 (Wilson, 2006). In the relative activity heat maps, the columns and rows associated with these nucleotides result in low activity values (Figures 1-5, Supplementary Figure 4). It is important to note that because there is complete coverage of all double mutants in this data set, we can be certain that there are no possible compensatory mutations. These results show how catalytic residues can be identified in the comprehensive pairwise mutagenesis data.

### Unpaired nucleotides show tertiary structure dependent mutational effects.

Mutations to nucleotides found in terminal loops that are not involved in tertiary structure elements showed high relative activity for most single and double mutants, and essentially no epistasis. This is not surprising if these loops reside on the periphery of the ribozyme and are not involved in structural contacts with other nucleotides. This is the case for L4 of the CPEB3 and HDV ribozymes (Figure 1, Figure 2), and L1 and L3 of the hairpin ribozyme (Figure 4). Two mutations within these loops do not reduce activity, and mutations in these loops do not rescue other deleterious mutations such as those that break a base pair (Figures 1, 2, and 4).

The internal loops (LA and LB) of the hairpin ribozyme are structurally important (Figure 4). Interactions between nucleotides within LB include six non-Watson-Crick base pairing interactions that are important for the formation of an active ribozyme structure (Fedor, 2000). Several non-canonical base-base and sugar-base hydrogen bonds between nucleotides within LA are also important for the formation of the active site (Fedor, 2000; Wilson, 2006). Docking between LA and LB is necessary for the formation of a catalytically active ribozyme and is facilitated by a Watson-Crick base pair between G1 and C46 in the version of the ribozyme used here (Rupert and Ferré-D'Amaré, 2001). In contrast to terminal loop regions, most single mutations within LA and LB resulted in low self-cleavage activity in our data (Figure 4). In addition, the double mutants within and between loop A and loop B show several instances of strong positive epistasis (Figure 4, Insets), and the distributions of epistasis within and between these loops are significantly different than the terminal loops that are not structurally important (Figure 4D). This positive epistasis indicates that many of the important structural contacts can be facilitated by other specific pairs of nucleotides. For example, the double mutant G1C and C46G shows strong epistasis suggesting that swapping a C-G base pair for the G-C base pair can restore activity by facilitating docking between the two loops. Several double mutants at positions that form non-canonical interactions in LB show positive epistasis. For example, mutation A41G shows positive epistasis when the interacting nucleotide C65 is mutated to a G or U. The non-canonical base pair G42:A64 shows positive epistasis for the mutations G42U A64G. The non-canonical A45:A59 interaction shows positive epistasis for several pairs of mutations (A45U A59C, A45C A59C, A45G A59U). Finally, the non-canonical base pair A47:G57 in LB, and C3:A28 in LA, both show positive epistasis for double mutants that result in an AU base pair. This analysis indicates that important structural contacts can be achieved with several different nucleotide combinations. The difference between terminal loops and loops with structural importance highlights how activity-based data can help identify non-canonical structures that are challenging to predict computationally, and that might be difficult to identify by other common approaches, such as chemical probing experiments (Walter et al., 2000).

Another example of structurally important unpaired regions can be found in the CUGA uridine turn (U-turn) motif in the hammerhead ribozyme (Figure 5). This CUGA turn forms the catalytic pocket and positions a catalytic cytosine (-1C) at the cleavage site (Doudna, 1995). A crystal structure of the sTRSV ribozyme showed a base pair between the nucleotides corresponding to C20 and G25 in the ribozyme construct used for our experiments (Chi et al., 2008). These two nucleotides showed strong positive epistasis for the mutations C20G and G25C, which substitutes a G:C base pair for the original C:G base pair. All other single and double mutants in this region showed low activity, and no instances of strong positive epistasis within or between this motif (Figure 5). The low activity resulting from mutations in this region confirms the functional importance of this motif, and indicates that this motif cannot be easily formed or rescued by sequences with up to two mutational differences, except for the G:C base pair swap.

Tertiary interactions between loops in the hammerhead ribozyme provide another example of structurally important loop regions. Type III hammerhead ribozymes, like the one used in this study, contain tertiary interactions between nucleotides in the loops of P1 and P2 that are implicated in structural organization of the catalytic core. A crystal structure of this loop-loop interaction showed a network of interhelical non-canonical base pairs and stacks, with several nucleobases in stem-loop I interacting with more than one nucleobase in stem-loop II (Chi et al., 2008; Martick and Scott, 2006). However, there are numerous different loop sequences in naturally occurring hammerhead ribozymes indicating that this loop-loop interaction can be formed by a variety of different sequences (Burke and Greathouse, 2005; Perreault et al., 2011). We therefore anticipated that the we would observe a significant level of positive epistasis between these two loops for double mutations that were capable of maintaining these tertiary interactions. Surprisingly, however, we found that most individual and double mutations do not reduce activity (Figure 5), and double mutants do not show positive epistasis (Supplementary Figure 5). This indicates that the multiple interactions between the loops compensate for mutations that break a single interaction. It is interesting to note that the mutational robustness of these loops has been exploited in bioengineering applications, where insertion of an aptamer into one of the loops and randomization of the other allowed for the selection of synthetic riboswitches (Townshend et al., 2015). The identification of robust structural elements though high-throughput mutational data could be useful for identifying better targets for aptamer integration in other ribozymes.

### Epistasis plots are an informative approach to visualizing high-throughput activity data.

Previous studies have reported comprehensive pairwise mutagenesis of ribozymes that provide interesting opportunities for comparison to the data presented here. For example, all pairwise mutations in a 42-nucleotide region of the same twister ribozyme were previously reported (Kobori and Yokobayashi, 2016). Compared to our experiments, these previous experiments used a later transcriptional time point (2h) and lower magnesium concentration (6mM). They did not calculate epistasis, and reported the Relative Activity of all double mutants using heatmaps similar to the figures presented here. The results were highly similar, and the authors were able to identify paired regions in the data. The similarity between the results illustrates the reliability of this sequencing-based approach, which is promising for future data sharing and meta-analysis efforts. In another prior work, all pairwise mutations in the glmS ribozyme were analyzed using a custom-built fluorescent RNA array (Andreasson et al., 2020). The power of this approach is that they were able to monitor self-cleavage over short and long time scales, which enables differentiating both very slow and very fast self-cleaving variants. While the authors did not calculate pairwise epistasis, they reported relative activity heatmaps and also "rescue effects" when the activity of a double mutant is sufficiently higher than the activity of a single mutant. This rescue analysis is very similar to positive epistasis, but only takes into account one mutation at a time. This analysis was also able to identify many of the know base-pair interactions and some tertiary contacts in the glmS ribozyme. In addition, they were able to observe some minor secondary structure rearrangement, where mutations in some nucleotides were able to rescue neighboring nucleotides by shifting the base-pairing slightly. The pairwise epistasis analysis presented here adds an additional approach to extract information from such high-throughput sequencing-based analysis of self-cleaving ribozymes. Unlike the rescue analysis, which can only identify positive interactions, the ability to detect negative epistatic interactions may help further identify structurally important regions for RNA sequence design and engineering efforts.

## Conclusion

We have determined the relative activity for all single and double mutants of five self-cleaving ribozymes and use this data to calculate epistasis for all possible pairs of nucleotides. The data was collected under identical co-transcriptional conditions, facilitating direct comparison of the data sets. The data revealed signatures of structural elements including paired regions and non-canonical structures. In addition, the comprehensiveness of the double mutants enabled identification of catalytic residues. Recently, there has been significant progress towards predicting RNA structures from sequence using machine learning approaches. The machine learning models are typically trained on structural biology data from x-ray crystallography, chemical probing (SHAPE), and natural sequence conservation. Self-cleaving ribozymes have been central to this effort. Our approach is similar to SHAPE in that it can be obtained with common lab equipment and commercially available reagents. The activity data presented provides information similar to natural sequence conservation, except that it provides quantitative effects of mutations, not just frequency. We hope that the activity-based data presented here will provide information not present in these other training data sets and help advance computational predictions.

## Materials and Methods

### Mutational library design and preparation of self-cleaving ribozymes

Single-stranded DNA molecules used as templates for in vitro transcription were synthesized with 97% of the base of the reference sequence and 1% of the three other remaining bases at each position (Keck Oligo Synthesis Resource, Yale). The ssDNA library was made double stranded to allow for T7 transcription via low cycle PCR using Taq DNA polymerase.

### Co-transcriptional self-cleavage assay

The co-transcriptional self-cleavage reactions were carried out in triplicate by combining 20 µL 10X T7 transcription buffer (500 µL 1M Tris pH 7.5, 50 µL 1M DTT, 20 µL 1M Spermidine, 150 µL 1M MgCl2, 280 µL RNase Free water), 4 µL rNTP

(25mM, NEB, Ipswich, Ma), 8 µL T7 RNA Polymerase-Plus enzyme mix (1,600 U, Invitrogen, Waltham, Ma), 160 µL nuclease free water, and 8 µL of double stranded DNA template (4 pmol, 0.5 µM PCR product) at 37°C for 30 minutes. The transcription and co-transcription self-cleavage reactions were quenched by adding 60 uL of 50 mM EDTA. The resulting RNA was purified and concentrated using Direct-zol RNA MicroPrep Kit with TRI-Reagent (Zymo Research, Irvine, Ca), and eluted in 7µL nuclease free water. Concentrations were determined via absorbance at 260 nm (ThermoFisher NanoDrop, Waltham, Ma), and normalized to 5µM. Reverse transcription reactions used 5 picomoles RNA and 20 picomoles of reverse transcription primer in a volume of 10 µL. RNA and primer were heated to 72 °C for 3 mins and cooled on ice. Reverse transcription was initiated by adding 4 µL SMARTScribe 5x First-Strand Buffer (TaKaRa, San Jose, Ca ), 2 µL dNTP (10 mM), 2 µL DTT (20 mM), 2 µL phased template switching oligo mix (10 µM), and 2 µL SMARTScribe Reverse Transcriptase (200 units, TaKaRa) (Bendixsen et al., 2020). The mixture was incubated at 42 °C for 90 mins and the reaction was stopped by heating to 72 °C for 15 mins. The resulting cDNA was purified on a silica-based column (DCC-5, Zymo Research) and eluted into 7 µL water. Illumina adapter sequences and indexes were added using high-fidelity PCR. A unique index combination was assigned to each ribozyme and for each replicate. The PCR reaction contained 3 µL purified cDNA, 12.5 µL KAPA HiFi HotStart ReadyMix (2X, KAPA Biosystems, Wilmington, Ma), 2.5 µL forward, 2.5 µL reverse primer (Illumina Nextera Index Kit) and 5 µL water. Several cycles of PCR were examined using gel electrophoresis and a PCR cycle was chosen that was still in logarithmic amplification, prior to saturation. Each PCR cycle consisted of 98 °C for 10 s, 63 °C for

30 s and 72 °C for 30 s. PCR DNA was purified on silica-based columns (DCC-5, Zymo Research) and eluted in 22.5 μL water. The final product was then verified using gel electrophoresis.

High-throughput sequencing

The indexed PCR products for all replicates were pooled together at equimolar concentrations based off of absorbance at 260 nm. Paired end sequencing reads were obtained for the pooled libraries using an Illumina HiSeq 4000 (Genomics and Cell Characterization Core Facility, University of Oregon).

Sequencing data analysis

Paired-end sequencing reads were joined using FLASh, allowing 'outies' due to overlapping reads. The joined sequencing reads were analyzed using custom Julia scripts that implement a sequence- length sliding window to screen for double mutant variants of a reference ribozyme. Nucleotide identities for each mutant were identified and then counted as either cleaved or uncleaved based on the presence or absence of the 5'-cleavage product sequence. The relative activity (RA) was calculated as previously described (Kobori and Yokobayashi, 2016). Briefly, a fraction cleaved (FC) was calculated for each genotype in each replicate as FC= $N_{clv}/(N_{clv} + N_{unclv})$. This value was normalized to the reference/wild type fraction cleaved as RA = $FC/FC_{wt}$. The RA values were averaged across the three replicates and then plotted as a heatmap. Epistasis interactions for each double mutant (i, j) were quantified as previously described (Bendixsen et al., 2017), where $Epistasis\ (\varepsilon) = \frac{\log(i,j)}{\log(i)\log(j)}$ . In order to eliminate false positive detection of epistasis interactions, values were filtered to eliminate instances where the difference between the double and any of the single mutants was less than 1-3σ of the overall distribution of differences between the single and double mutant relative activities. Values greater than 1 indicate positive epistasis, and values less than zero indicate negative epistasis. Mann-Whitney U test was used to determine the probability that epistasis or activity values of different structural elements were from the same distribution.

Correlation of thermodynamic stability of paired regions and observed mutational effects. Each base paired region was split into two separate RNA sequences containing only the nucleotides involved in base pairing, omitting nucleotides belonging to stem loops. Complex formation between each pair of strands at was analyzed in Nupack using Serrra and Turner RNA energy parameters in order to obtain minimum free energy values for each paired region (37°C, [1µM]). Using custom Julia scripts, the median relative activity for single mutations to each paired region was plotted as a function of the calculated free energy and a Pearson correlation coefficient was calculated.

## Acknowledgements

The authors acknowledge funding from the National Science Foundation (EH, grant number OIA-1738865, OIA-1826801), National Aeronautics and Space Administration (EH, grant number 80NSSC17K0738), and the Human Frontier Science Program (EH, Ref.-No: RGY0077/2019).

## Competing Interests Statement

The authors declare that no competing interests exist.

## Data Availability

Sequencing reads in FastQ format are available at ENA (PRJEB52899). Sequences, activity data, and computer code is available at GitLab ( https://gitlab.com/bsu/biocompute-public/mut_12).

## References

Andreasson JOL, Savinov A, Block SM, Greenleaf WJ. 2020. Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme. *Nature Communications* **11**:1–11. doi:10.1038/s41467-020-15540-1

Bendixsen DP, Collet J, Østman B, Hayden EJ. 2019. Genotype network intersections promote evolutionary innovation. *PLoS Biol* **17**:e3000300. doi:10.1371/journal.pbio.3000300

Bendixsen DP, Østman B, Hayden EJ. 2017. Negative Epistasis in Experimental RNA Fitness Landscapes. *J Mol Evol* **85**:159–168. doi:10.1007/s00239-017-9817-5

Bendixsen DP, Pollock TB, Peri G, Hayden EJ. 2021. Experimental Resurrection of Ancestral Mammalian CPEB3 Ribozymes Reveals Deep Functional Conservation. *Molecular Biology and Evolution* **38**:2843–2853. doi:10.1093/molbev/msab074

Bendixsen DP, Roberts JM, Townshend B, Hayden EJ. 2020. Phased nucleotide inserts for sequencing low-diversity RNA samples from in vitro selection experiments. *RNA* **26**:1060–1068. doi:10.1261/rna.072413.119

Burke DH, Greathouse ST. 2005. Low-magnesium, trans-cleavage activity by type III, tertiary stabilized hammerhead ribozymes with stem 1 discontinuities. *BMC Biochemistry* **6**:14. doi:10.1186/1471-2091-6-14

Calonaci N, Jones A, Cuturello F, Sattler M, Bussi G. 2020. Machine learning a model for RNA structure prediction. *NAR Genomics and Bioinformatics* **2**. doi:10.1093/nargab/lqaa090

Chadalavada DM, Cerrone-Szakal AL, Bevilacqua PC. 2007. Wild-type is the optimal sequence of the HDV ribozyme under cotranscriptional conditions. *RNA* **13**:2189–2201. doi:10.1261/rna.778107

Chi Y-I, Martick M, Lares M, Kim R, Scott WG, Kim S-H. 2008. Capturing hammerhead ribozyme structures in action by modulating general base catalysis. *PLoS Biol* **6**:e234. doi:10.1371/journal.pbio.0060234

Doudna JA. 1995. Hammerhead ribozyme structure: U-turn for RNA structural biology. *Structure* **3**:747–750. doi:10.1016/S0969-2126(01)00208-8

Dutheil JY, Jossinet F, Westhof E. 2010. Base Pairing Constraints Drive Structural Epistasis in Ribosomal RNA Sequences. *Molecular Biology and Evolution* **27**:1868–1876. doi:10.1093/molbev/msq069

Fedor MJ. 2000. Structure and function of the hairpin ribozyme. *Journal of Molecular Biology* **297**:269–291. doi:10.1006/jmbi.2000.3560

Halvorsen M, Martin JS, Broadaway S, Laederach A. 2010. Disease-Associated Mutations That Alter the RNA Structural Ensemble. *PLOS Genetics* **6**:e1001074. doi:10.1371/journal.pgen.1001074

Hayden EJ. 2016. Empirical analysis of RNA robustness and evolution using high-throughput sequencing of ribozyme reactions. *Methods*, In vitro selection and evolution **106**:97–104. doi:10.1016/j.ymeth.2016.05.014

Jimenez RM, Polanco JA, Lupták A. 2015. Chemistry and Biology of Self-Cleaving Ribozymes. *Trends in Biochemical Sciences* **40**:648–661. doi:10.1016/j.tibs.2015.09.001

Kobori S, Nomura Y, Miu A, Yokobayashi Y. 2015. High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Res* **43**:e85. doi:10.1093/nar/gkv265

Kobori S, Takahashi K, Yokobayashi Y. 2017. Deep Sequencing Analysis of Aptazyme Variants Based on a Pistol Ribozyme. *ACS Synth Biol* **6**:1283–1288. doi:10.1021/acssynbio.7b00057

Kobori S, Yokobayashi Y. 2016. High-Throughput Mutational Analysis of a Twister Ribozyme. *Angew Chem Int Ed Engl* **55**:10354–10357. doi:10.1002/anie.201605470

Li C, Qian W, Maclean CJ, Zhang J. 2016. The fitness landscape of a tRNA gene. *Science* **352**:837–840. doi:10.1126/science.aae0568

Liang JC, Bloom RJ, Smolke CD. 2011. Engineering Biological Systems with Synthetic RNA Molecules. *Molecular Cell* **43**:915–926. doi:10.1016/j.molcel.2011.08.023

Liu Y, Wilson TJ, Mcphee SA, Lilley DMJ. 2014. Crystal structure and mechanistic investigation of the twister ribozyme. *Nature Chemical Biology* **10**:739–44. doi:http://dx.doi.org/10.1038/nchembio.1587

Martick M, Scott WG. 2006. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* **126**:309–320. doi:10.1016/j.cell.2006.06.036

Müller S, Appel B, Krellenberg T, Petkovic S. 2012. The many faces of the hairpin ribozyme: Structural and functional variants of a small catalytic rna. *IUBMB Life* **64**:36–47. doi:10.1002/iub.575

Nakano S. 2000. General Acid-Base Catalysis in the Mechanism of a Hepatitis Delta Virus Ribozyme. *Science* **287**:1493–1497. doi:10.1126/science.287.5457.1493

Olson CA, Wu NC, Sun R. 2014. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology* **24**:2643–2651. doi:10.1016/j.cub.2014.09.072

Peng H, Latifi B, Müller S, Lupták A, A. Chen I. 2021. Self-cleaving ribozymes: substrate specificity and synthetic biology applications. *RSC Chemical Biology*. doi:10.1039/D0CB00207K

Peri G, Gibard C, Shults NH, Crossin K, Hayden EJ. 2022. Dynamic RNA fitness landscapes of a group I ribozyme during changes to the experimental environment. *Molecular Biology and Evolution* msab373. doi:10.1093/molbev/msab373

Perreault J, Weinberg Z, Roth A, Popescu O, Chartrand P, Ferbeyre G, Breaker RR. 2011. Identification of Hammerhead Ribozymes in All Domains of Life Reveals Novel Structural Variations. *PLOS Computational Biology* **7**:e1002031. doi:10.1371/journal.pcbi.1002031

Prody GA, Bakos JT, Buzayan JM, Schneider IR, Bruening G. 1986. Autolytic Processing of Dimeric Plant Virus Satellite RNA. *Science* **231**:1577–1580.

Roth A, Weinberg Z, Chen AGY, Kim PB, Ames TD, Breaker RR. 2014. A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat Chem Biol* **10**:56–60. doi:10.1038/nchembio.1386

Rupert PB, Ferré-D'Amaré AR. 2001. Crystal structure of a hairpin ribozyme–inhibitor complex with implications for catalysis. *Nature* **410**:780–786. doi:10.1038/35071009

Salehi-Ashtiani K, Lupták A, Litovchick A, Szostak JW. 2006. A Genomewide Search for Ribozymes Reveals an HDV-like Sequence in the Human CPEB3 Gene. *Science* **313**:1788–1792.

Scott WG, Horan LH, Martick M. 2013. The Hammerhead Ribozyme: Structure, Catalysis and Gene Regulation. *Prog Mol Biol Transl Sci* **120**:1–23. doi:10.1016/B978-0-12-381286-5.00001-9

Shen Y, Pressman A, Janzen E, Chen IA. 2021. Kinetic sequencing (k-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters. *Nucleic Acids Research* **49**:e67–e67. doi:10.1093/nar/gkab199

Skilandat M, Rowinska-Zyrek M, Sigel RKO. 2016. Secondary structure confirmation and localization of Mg2+ ions in the mammalian CPEB3 ribozyme. *RNA* **22**:750–763. doi:10.1261/rna.053843.115

Stifel J, Spöring M, Hartig JS. 2019. Expanding the toolbox of synthetic riboswitches with guanine-dependent aptazymes. *Synthetic Biology* **4**. doi:10.1093/synbio/ysy022

Townshend B, Kennedy AB, Xiang JS, Smolke CD. 2015. High-throughput cellular RNA device engineering. *Nat Meth* **advance online publication**. doi:10.1038/nmeth.3486

Townshend RJL, Eismann S, Watkins AM, Rangan R, Karelina M, Das R, Dror RO. 2021. Geometric deep learning of RNA structure. *Science* **373**:1047–1051. doi:10.1126/science.abe5650

Walter NG, Yang N, Burke JM. 2000. Probing non-selective cation binding in the hairpin ribozyme with Tb(III)11Edited by J. Doudna. *Journal of Molecular Biology* **298**:539–555. doi:10.1006/jmbi.2000.3691

Wei KY, Smolke CD. 2015. Engineering dynamic cell cycle control with synthetic small molecule-responsive RNA devices. *Journal of Biological Engineering* **9**:21. doi:10.1186/s13036-015-0019-7

Wilson TJ. 2006. Nucleobase catalysis in the hairpin ribozyme. *RNA* **12**:980–987. doi:10.1261/rna.11706

Wilson TJ, Liu Y, Domnick C, Kath-Schorr S, Lilley DMJ. 2016. The Novel Chemical Mechanism of the Twister Ribozyme. *J Am Chem Soc* **138**:6151–6162. doi:10.1021/jacs.5b11791

Zhong G, Wang H, Bailey CC, Gao G, Farzan M. 2016. Rational design of aptazyme riboswitches for efficient control of gene expression in mammalian cells. *eLife* **5**:e18858. doi:10.7554/eLife.18858

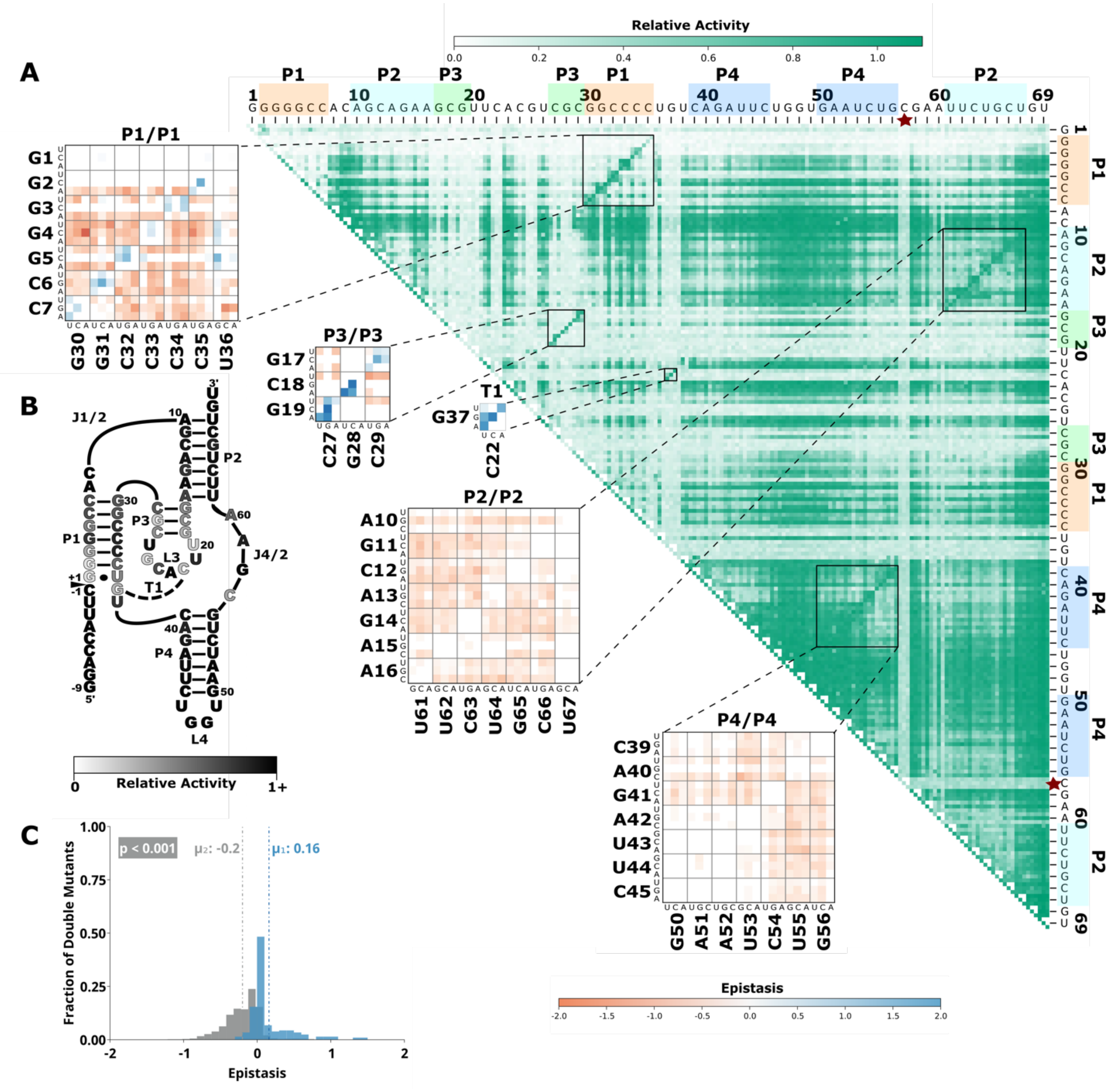


**Figure 2.1. Effects of mutations and pairwise epistasis in a CPEB3 ribozyme.**

A) Relative activity heatmap depicting all possible pairwise effects of mutations on the cleavage activity of a mammalian CPEB3 ribozyme. Base-paired regions P1, P2, P3, P4, and T1 are highlighted and color coordinated along the axes, and surrounded by black squares within the heatmap. Pairwise epistasis interactions observed for each paired regions are each shown as expanded insets for easy identification of the specific epistatic effects measured for each pair of mutations. Instances of positive epistasis are shaded blue, and negative epistasis is shaded red, with higher color intensity indicating a greater magnitude of epistasis. Catalytic residues are indicated by stars along the axes. B) Secondary structure of the CPEB3 ribozyme used in this study. Each nucleotide is shaded to indicate the average relative cleavage activity of all single mutations at that position. C) Histogram showing the distributions of epistasis in the paired regions of CPEB3. The distribution for double mutants within a paired region that are not involved in a base-pair is shown in grey, and the distribution for nucleotides involved in a base-pair is shown in blue.

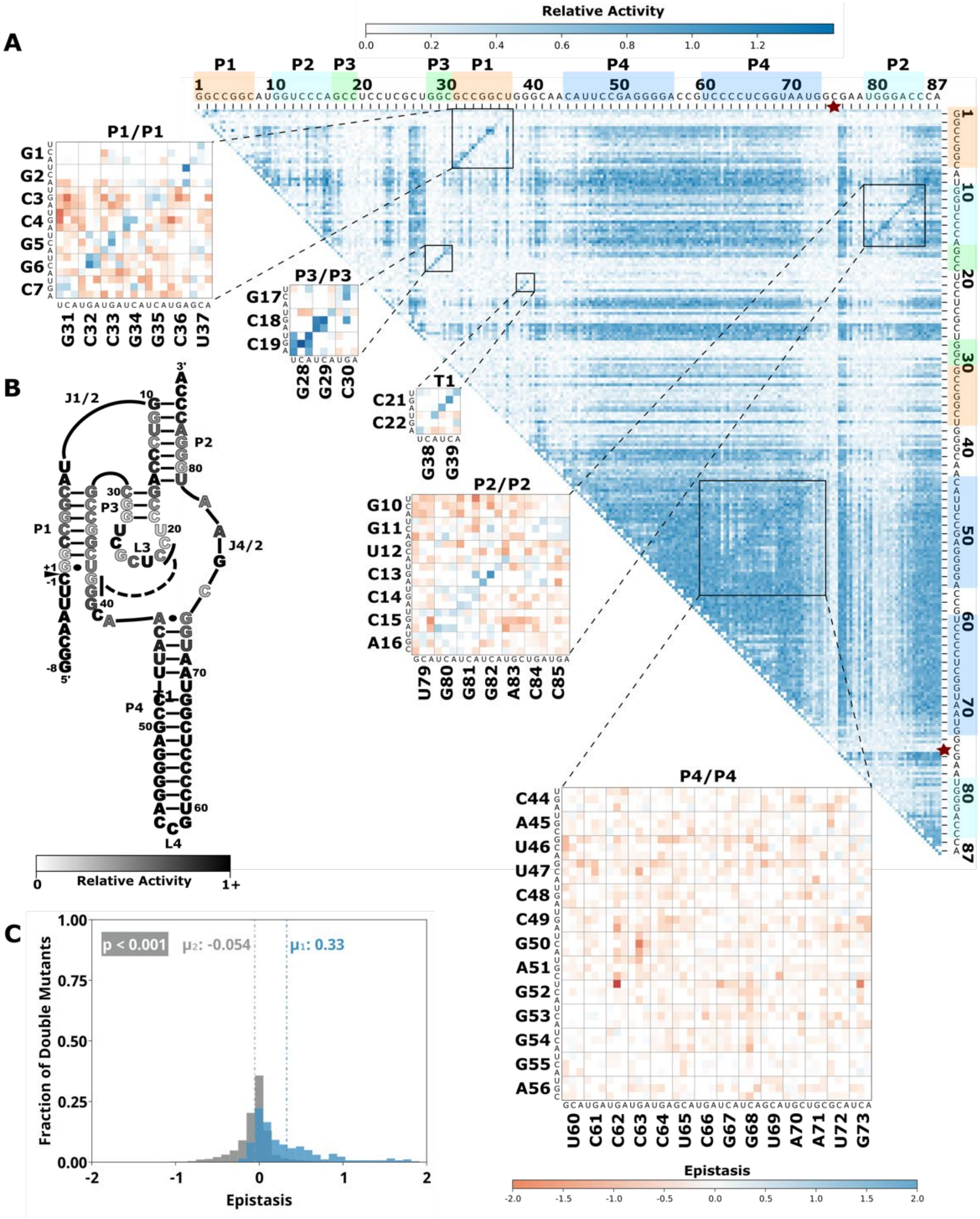
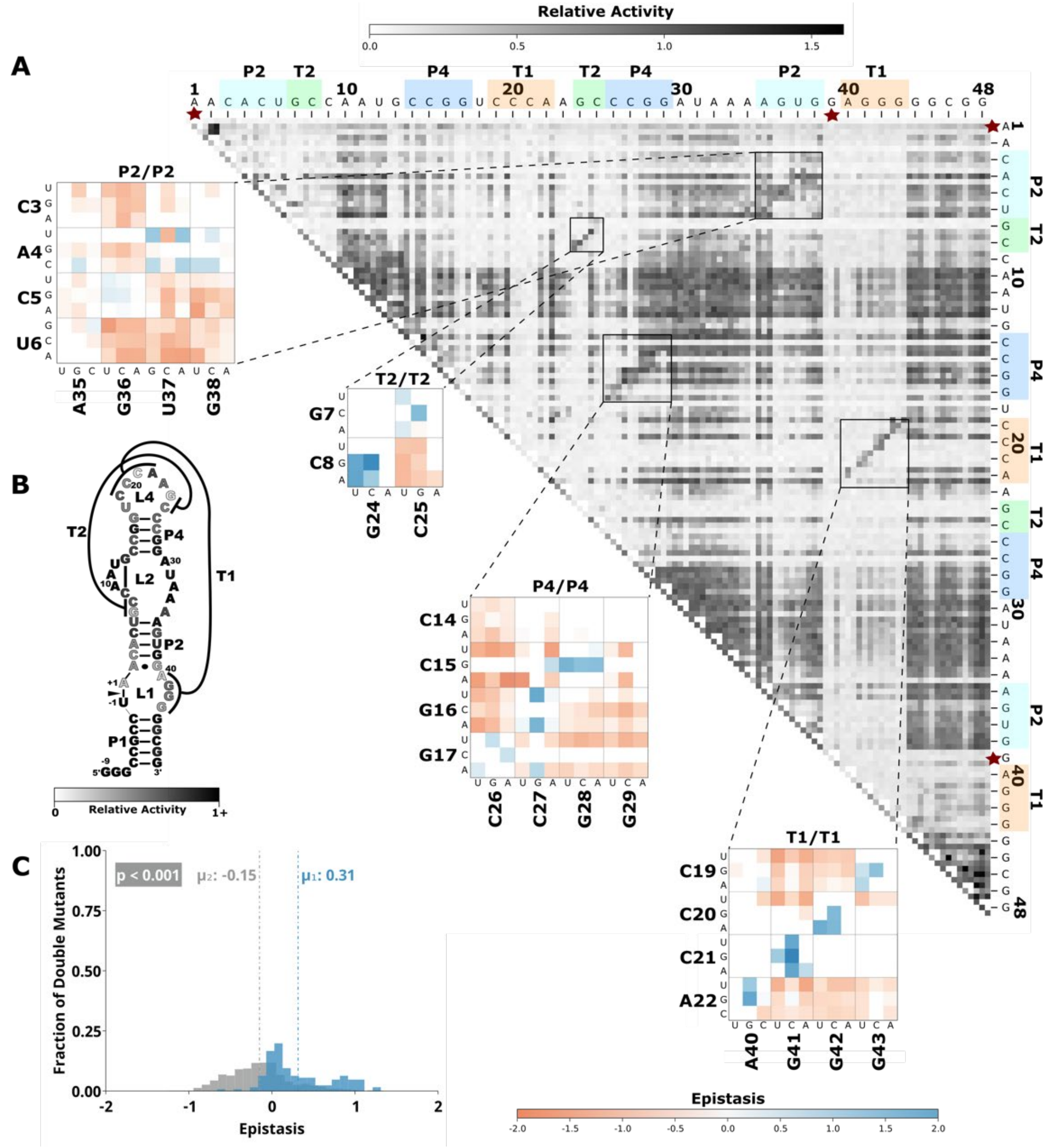


**Figure 2.2. Comprehensive pairwise epistasis landscape for a HDV self-cleaving ribozyme.**

A) Relative activity heatmap depicting all possible pairwise effects of mutations on the cleavage activity of an HDV ribozyme. Base-paired regions P1, P2, P3, P4, and T1 are highlighted and color coordinated along the axes, and surrounded by black squares within the heatmap. Pairwise epistasis interactions observed for each paired regions are each shown as expanded insets for easy identification of the specific epistatic effects measured for each pair of mutations. Instances of positive epistasis are shaded blue, and negative epistasis is shaded red, with higher color intensity indicating a greater magnitude of epistasis. Catalytic residues are indicated by stars along the axes. B) Secondary structure of the HDV ribozyme used in this study. Each nucleotide is shaded to indicate the average relative cleavage activity of all single mutations at that position. C) Histogram showing the distributions of epistasis in the paired regions of HDV. The distribution for double mutants within a paired region that are not involved in a base-pair is shown in grey, and the distribution for nucleotides involved in a base-pair is shown in blue.



## Figure 2.3. Comprehensive pairwise epistasis landscape for a twister self-cleaving ribozyme.

A) Relative activity heatmap depicting all possible pairwise effects of mutations on the cleavage activity of a twister ribozyme. Base-paired regions P2, P4, T1, and T2 are highlighted and color coordinated along the axes, and surrounded by black squares within the heatmap. Pairwise epistasis interactions observed for each paired region are each shown as expanded insets for easy identification of the specific epistatic effects measured for each pair of mutations. Instances of positive epistasis are shaded blue, and negative epistasis is shaded red, with higher color intensity indicating a greater magnitude of epistasis. Catalytic residues are indicated by stars along the axes. B) Secondary structure of the twister ribozyme used in this study. Each nucleotide is shaded to indicate the average relative cleavage activity of all single mutations at that position. C) Histogram showing the distributions of epistasis in the paired regions of twister. The distribution for double mutants within a paired region that are not involved in a base-pair is shown in grey, and the distribution for nucleotides involved in a base-pair is shown in blue.

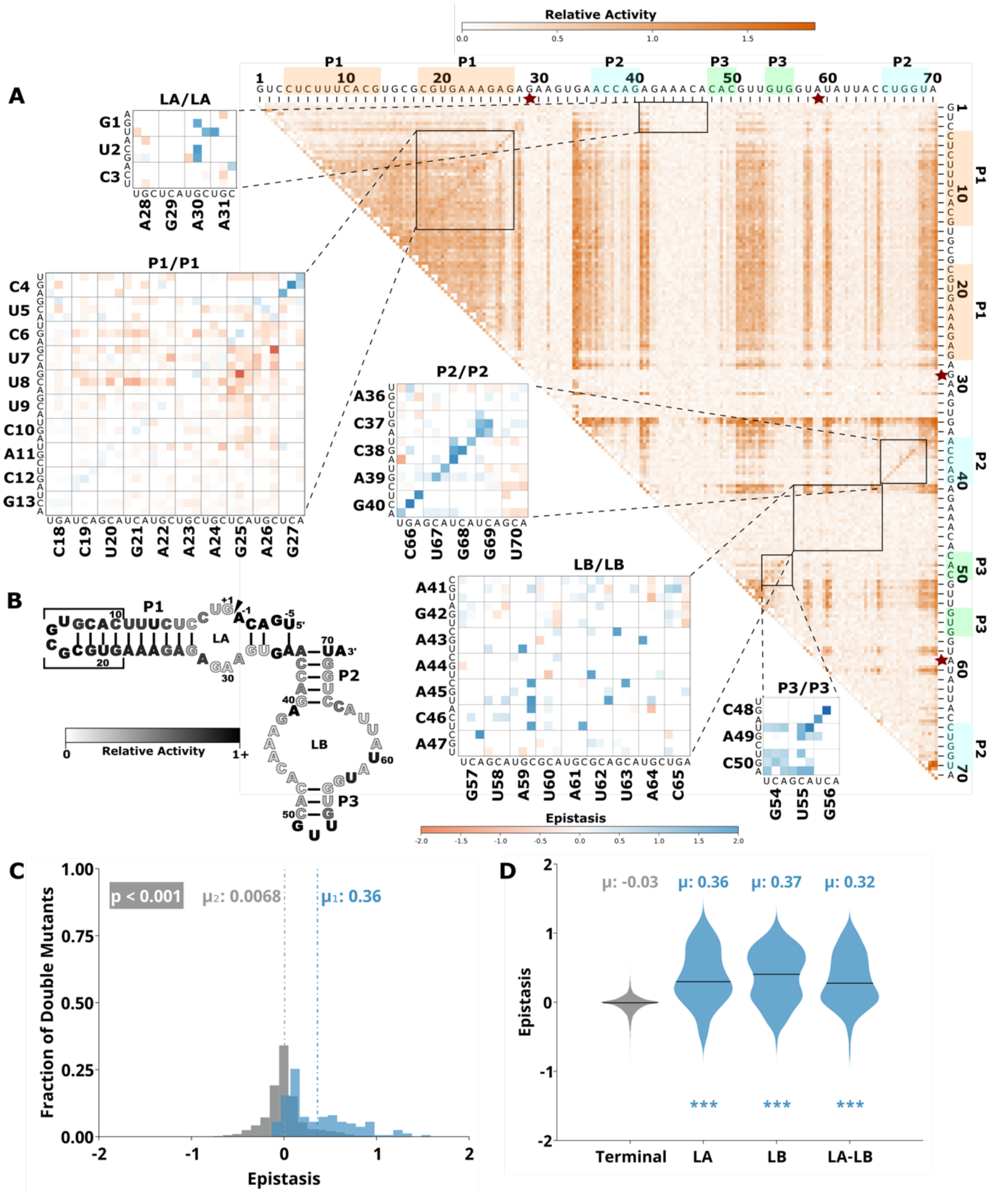


**Figure 2.4. Comprehensive pairwise epistasis landscape for a hairpin self-cleaving ribozyme.**

A) Relative activity heatmap depicting all possible pairwise effects of mutations on the cleavage activity of a hairpin ribozyme. Base-paired regions P1, P2, and P3 are highlighted and color coordinated along the axes, and surrounded by black squares within the heatmap. Pairwise epistasis interactions observed for each paired region are each shown as expanded insets for easy identification of the specific epistatic effects measured for each pair of mutations. Instances of positive epistasis are shaded blue, and negative epistasis is shaded red, with higher color intensity indicating a greater magnitude of epistasis. Catalytic residues are indicated by stars along the axes. B) Secondary structure of the hairpin ribozyme used in this study. Each nucleotide is shaded to indicate the average relative cleavage activity of all single mutations at that position. C) Histogram showing the distributions of epistasis in the paired regions of hairpin. The distribution for double mutants within a paired region that are not involved in a base-pair is shown in grey, and the distribution for nucleotides involved in a base-pair is shown in blue. D) Violin plots showing the distributions of epistasis in all terminal stem loops across all five ribozymes, and epistasis observed within loop A, loop B, and between loop A and loop B in the hairpin ribozyme.

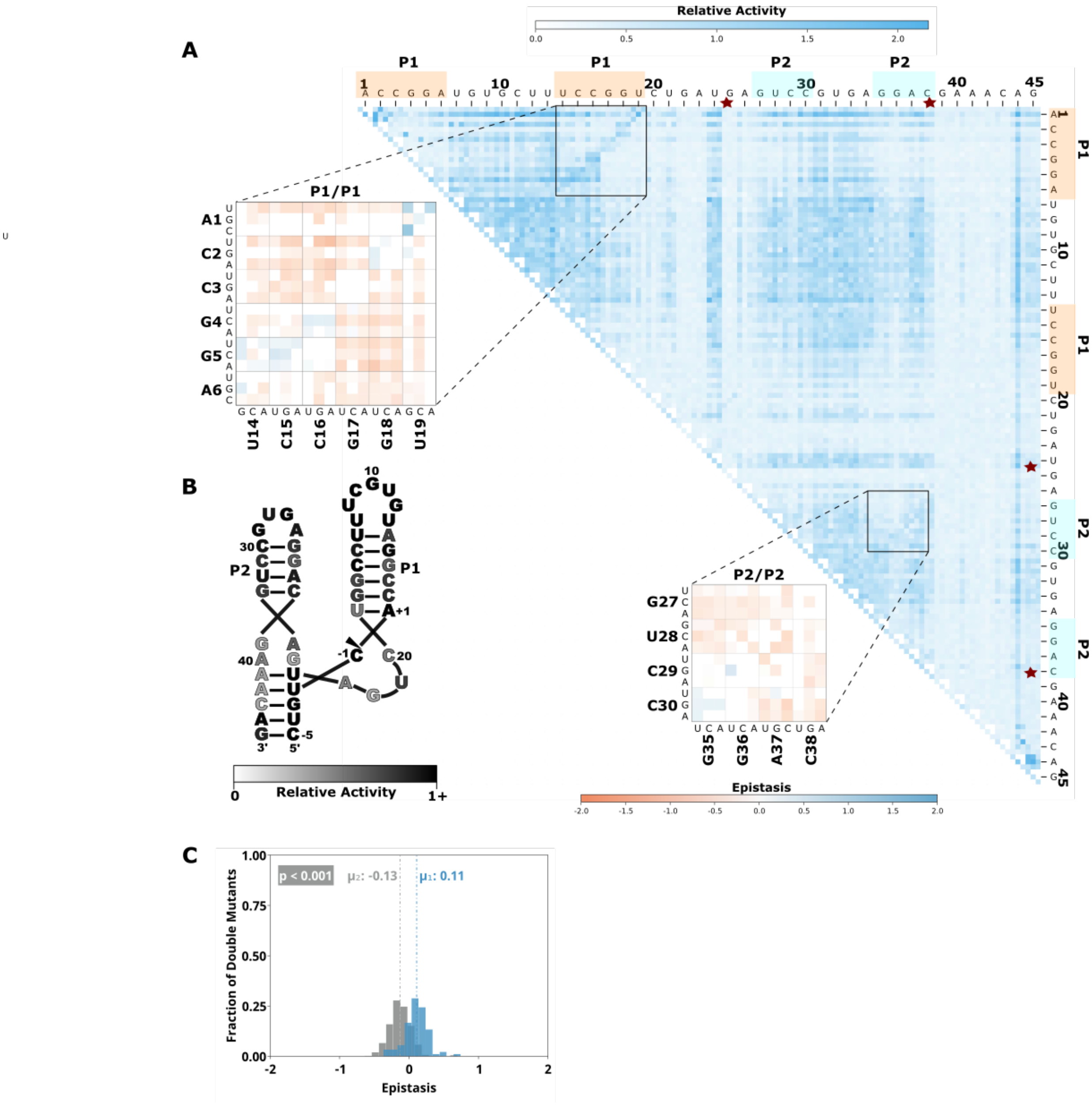


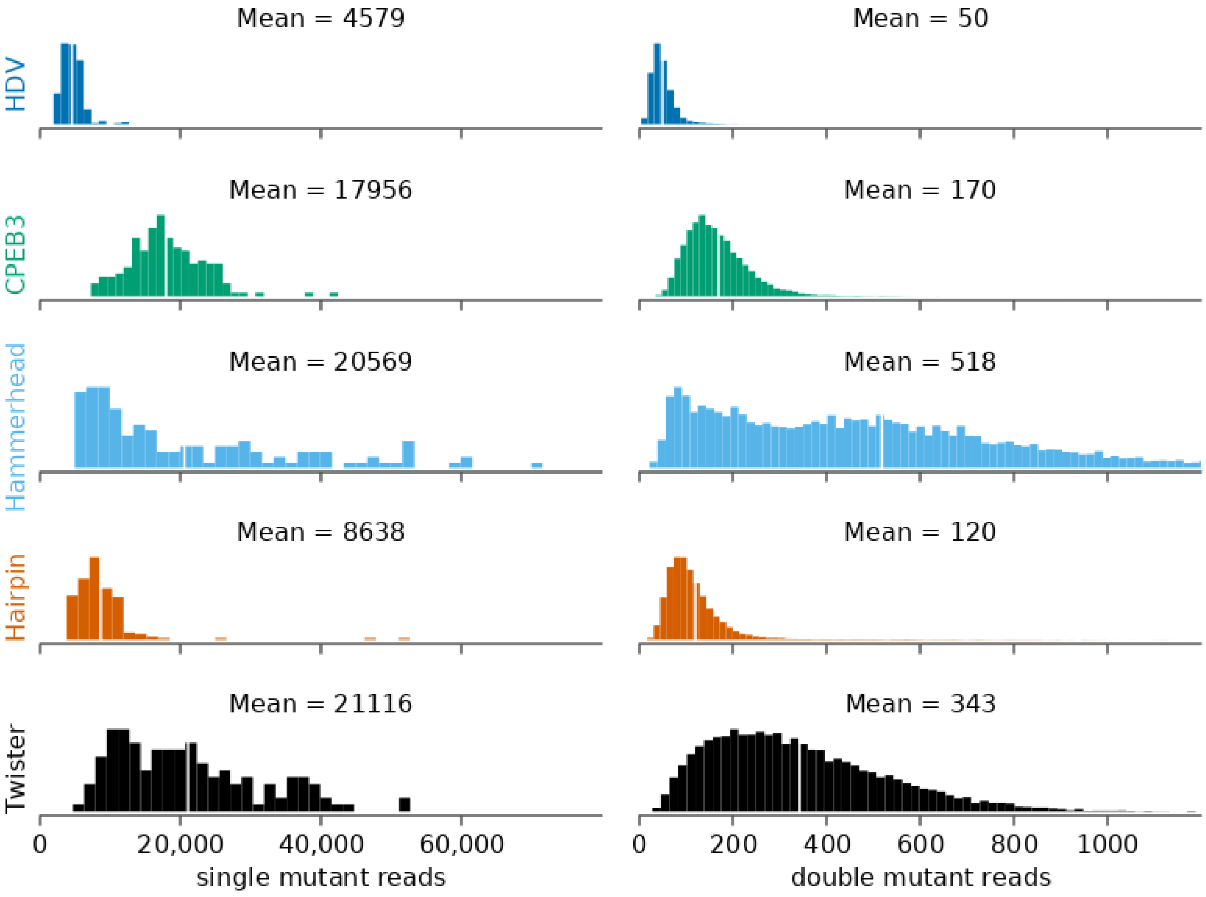
**Figure 2.5. Comprehensive pairwise epistasis landscape for a hammerhead self-cleaving ribozyme.**

A) Relative activity heatmap depicting all possible pairwise effects of mutations on the cleavage activity of a hammerhead ribozyme. Base-paired regions P1, and P2 are highlighted and color coordinated along the axes, and surrounded by black squares within the heatmap. Pairwise epistasis interactions observed for each paired region are each shown as expanded insets for easy identification of the specific epistatic effects measured for each pair of mutations. Instances of positive epistasis are shaded blue, and negative epistasis is shaded red, with higher color intensity indicating a greater magnitude of epistasis. Catalytic residues are indicated by stars along the axes. B) Secondary structure of the hammerhead ribozyme used in this study. Each nucleotide is shaded to indicate the average relative cleavage activity of all single mutations at that position. C) Histogram showing the distributions of epistasis in the paired regions of hammerhead. The distribution for double mutants within a paired region that are not involved in a base-pair is shown in grey, and the distribution for nucleotides involved in a base-pair is shown in blue.

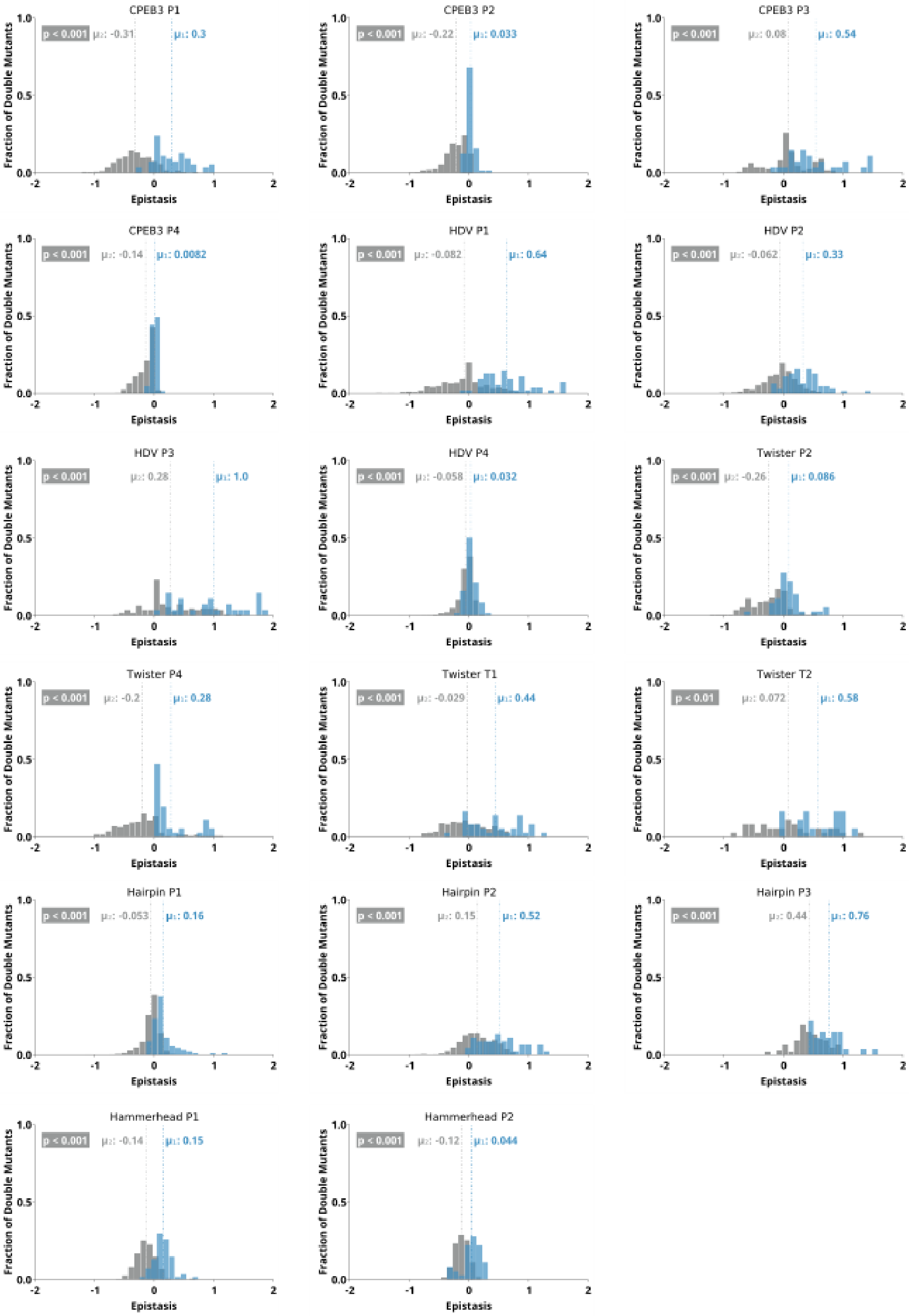
**Table 2.1 Summary of the lengths of each self-cleaving ribozyme used in this study, and the number of single and double mutants whose cleavage activity was analyzed.**

| **Ribozyme Name** | **Hairpin** | **Hammerhead** | **CPEB3** | **HDV** | **Twister** |
|---|---|---|---|---|---|
| **Ribozyme Length** | 71 | 45 | 69 | 87 | 48 |
| **Possible single mutants** | 213 | 135 | 207 | 261 | 144 |
| **Possible double mutants** | 22,365 | 8,910 | 21,114 | 33,669 | 10,152 |
| **Total mapped reads** | 5,067,216 | 8,054,498 | 9,238,603 | 3,316,380 | 7,762,863 |
| **Overall fraction cleaved** | 0.23 | 0.19 | 0.68 | 0.60 | 0.31 |

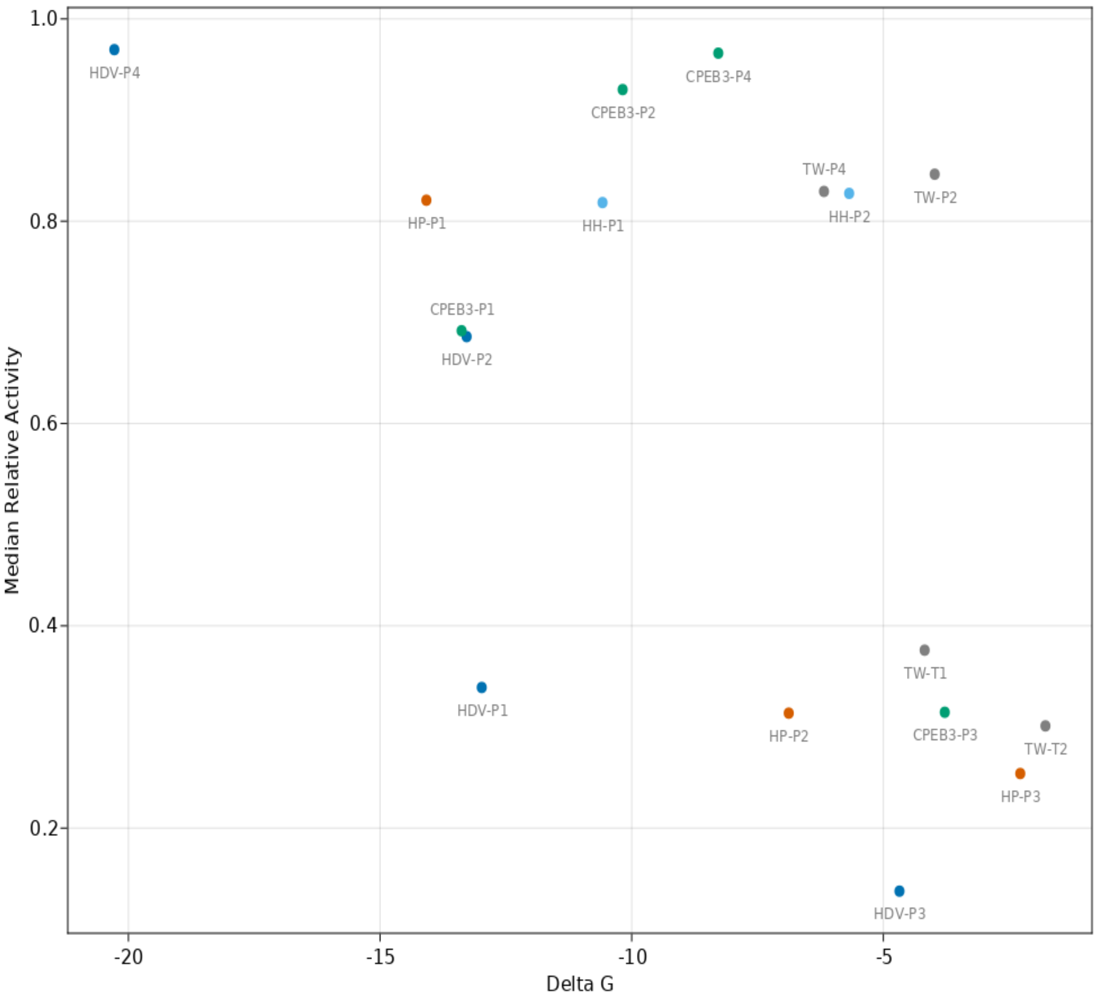
## Supplementary Materials



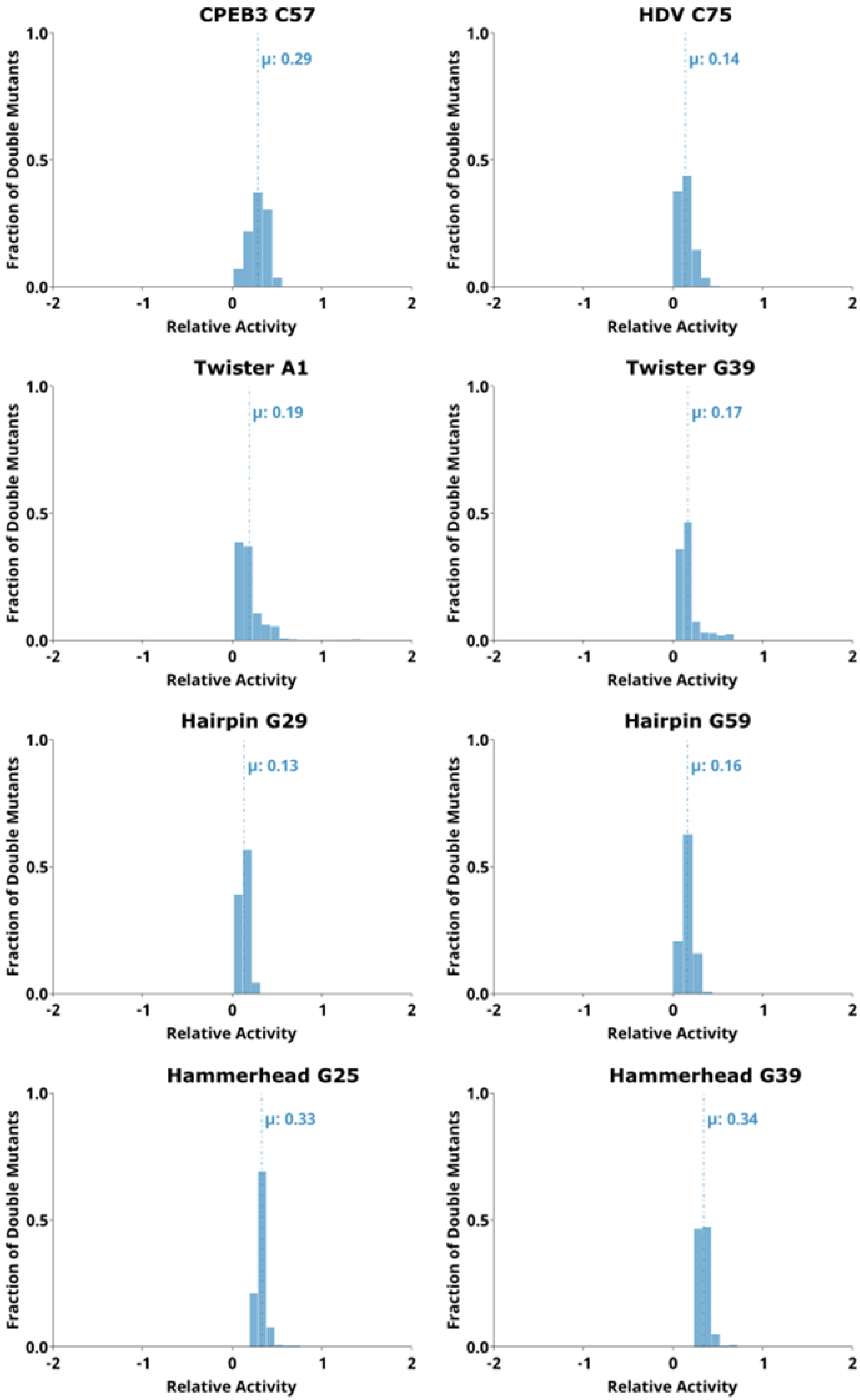
**Supplementary Figure 2.1. Histogram of the distributions of read counts (read depth) for the single and double mutants matching to each ribozyme analyzed in this study (HDV, CPEB3, hammerhead, hairpin, twister).**



**Supplementary Figure 2.2. Distributions for epistasis values seen on and off anti-diagonal in the epistasis heatmaps. The distributions of epistasis values along the anti-diagonal corresponding to double mutations between nucleotides involved in a Watson-Crick base-pair are shown in blue, and the epistasis values seen off diagonal are shown in gray.**



**Supplementary Figure 2.3. Relationship between the Gibbs free energy (ΔG) of each base paired region belonging to the hairpin, hammerhead, CPEB3, HDV, and twister ribozymes, and the median relative activity of all single mutants within each base paired region (Pearson Correlation = -0.53).**


CPEB3 C57

HDV C75

Twister A1

Twister G39

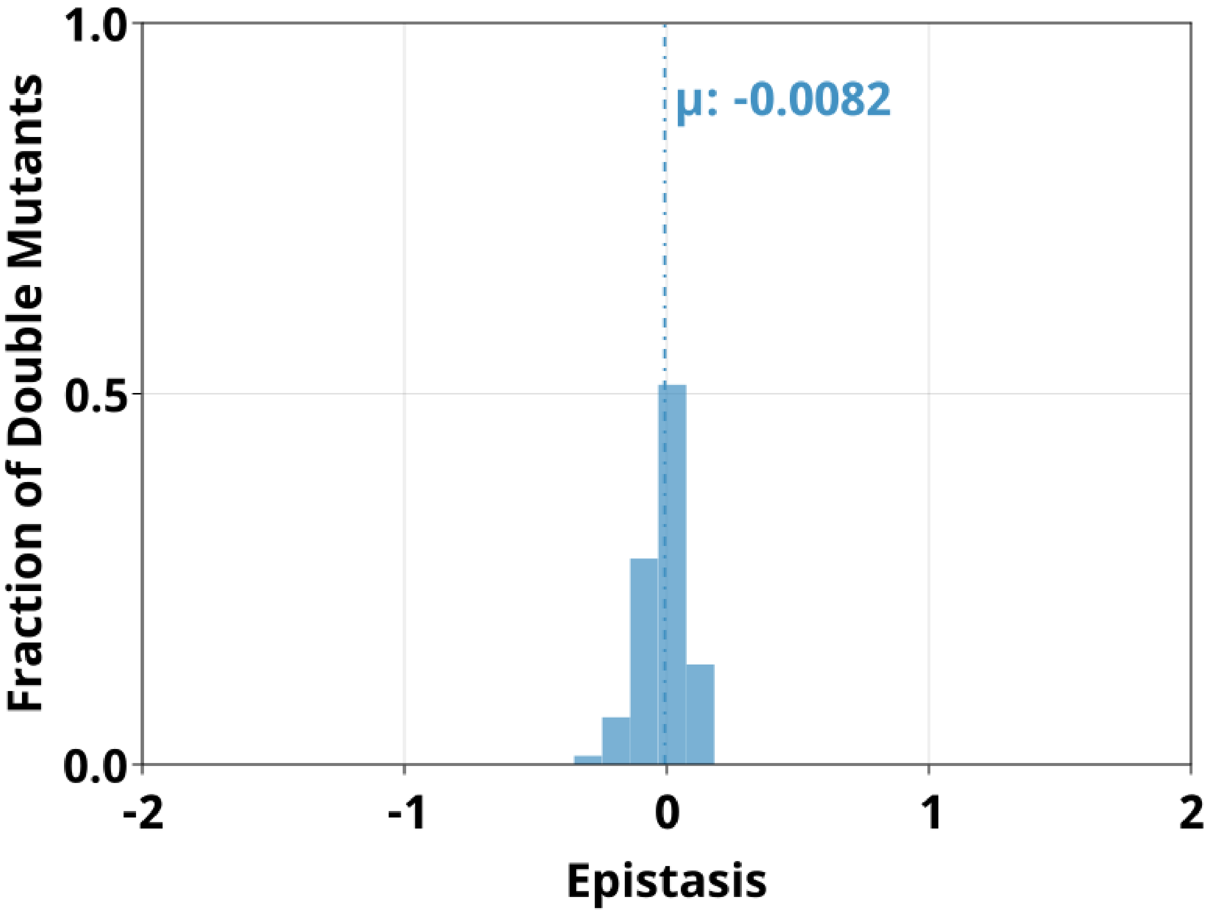
Hairpin G29

Hairpin G59

Hammerhead G25

Hammerhead G39

**Supplementary Figure 2.4. Distributions of relative self-cleavage activity observed for sequences containing mutations to the catalytic nucleotides in the CPEB3, HDV, twister, hairpin, and hammerhead ribozymes.**



**Supplementary Figure 2.5. Distribution of pairwise epistasis observed between the loops of P1 and P2 in the hammerhead ribozyme.**

**Supplementary Table 2.1. Oligonucleotides used in this study.**

| Name | Sequence | Notes |
|---|---|---|
| **HDV template** | GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**TGGGTCCCATTCGCCATTACCGAGGGGACGGTCCCCTCGGAATGTTGCCCAGCCGGCGCCAGCGAGGAGGCTGGGACCATGCCGGCC**ATCAGGCCTATAGTGAGTCGTATTAGCCG | DNA template for in-vitro transcriptions. Bolded nucleotides indicate positions synthesized using doped phosphoramidites (3% mutation rate) |
| **CPEB3 template** | GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**ACAGCAGAATTCGCAGATTCACCAGAATCTGACAGGGGCTGCGACGTGAACGCTTCTGCTGTGGCCCCC**GAATGGTCCTTTTCCTATAGTGAGTCGTATTAGCCG | DNA template for in-vitro transcriptions. Bolded nucleotides indicate positions synthesized using doped phosphoramidites (3% mutation rate) |
| **Twister template** | GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**CCGCCCCCTCCACTTTTATCCGGGCTTGGACCGGCATTGGCAGTGTT**AGGCGGCCCTTTTCCTATAGTGAGTCGTATTAGCCG | DNA template for in-vitro transcriptions. Bolded nucleotides indicate positions synthesized using doped phosphoramidites (3% mutation rate) |
| **Hairpin template** | GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**TACCAGGTAATATACCACAACGTGTGTTTCTCTGGTTCACTTCTCTCTTTCACGCGCACGTGAAAGAGGAC**TGTCATTTTCCTATAGTGAGTCGTATTAGCCG | DNA template for in-vitro transcriptions. Bolded nucleotides indicate positions synthesized using doped phosphoramidites (3% mutation rate) |
| **HH template** | GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**CTGTTTCGTCCTCACGGACTCATCAGACCGGAAAGCACATCCGGT**GACAGTTTTCCTATAGTGAGTCGTATTAGCCG | DNA template for in-vitro transcriptions. Bolded nucleotides indicate positions synthesized using doped phosphoramidites (3% mutation rate) |
| **T7 top strand** | CGGCTAATACGACTCACTATAG | PCR primer |

| | | |
|---|---|---|
| **RT primer** | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CATGCATGC rGrGrG | PCR/RT primer |
| **TSO1** | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCATGCATGCATGCATGC rGrGrG | Phased template switching oligo 1 |
| **TSO2** | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGCATGCATGCATGC rGrGrG | Phased template switching oligo2 |
| **TSO3** | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATGCATGCATGC rGrGrG | Phased template switching oligo 3 |
| **TSO4** | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CATGCATGC rGrGrG | Phased template switching oligo 4 |

# CHAPTER THREE: PREDICTING HIGHER-ORDER MUTATIONAL EFFECTS IN AN RNA ENZYME BY MACHINE LEARNING OF HIGH-THROUGHPUT EXPERIMENTAL DATA

Jim Beck[1†], Jessica M. Roberts[2†], Joey Kitzhaber[3], Ashlyn Trapp[4], Edoardo Serra[1], Francesca Spezzano[1], Eric J. Hayden[2,3,*]

[1]Computing PhD Program, Boise State University, Boise, ID, USA

[2]Biomolecular Sciences Graduate Programs, Boise State University, Boise, ID, USA

[3]Department of Computer Science, Boise State University, Boise, ID, USA

[4]Department of Biological Sciences, Boise State University, Boise, ID, USA

[†]Authors contributed equally to this work and share first authorship

*** Correspondence:** Eric J. Hayden erichayden@boisestate.edu

Keywords: ribozyme, fitness landscape, RNA, epistasis, machine learning, long short-term memory, random forest.

## Abstract

Ribozymes are RNA molecules that catalyze biochemical reactions. Self-cleaving ribozymes are a common naturally occurring class of ribozymes that catalyze site-specific cleavage of their own phosphodiester backbone. In addition to their natural functions, self-cleaving ribozymes have been used to engineer control of gene expression because they can be designed to alter RNA processing and stability. However, the

rational design of ribozyme activity remains challenging, and many ribozyme-based systems are engineered or improved by random mutagenesis and selection (in vitro evolution). Improving a ribozyme-based system often requires several mutations to achieve the desired function, but extensive pairwise and higher-order epistasis prevent a simple prediction of the effect of multiple mutations that is needed for rational design. Recently, high-throughput sequencing-based approaches have produced data sets on the effects of numerous mutations in different ribozymes (RNA fitness landscapes). Here we used such high-throughput experimental data from variants of the CPEB3 self-cleaving ribozyme to train a predictive model through machine learning approaches. We trained models using either a random forest or long short-term memory (LSTM) recurrent neural network approach. We found that models trained on a comprehensive set of pairwise mutant data could predict active sequences at higher mutational distances, but the correlation between predicted and experimentally observed self-cleavage activity decreased with increasing mutational distance. Adding sequences with increasingly higher numbers of mutations to the training data improved the correlation at increasing mutational distances. Systematically reducing the size of the training data set suggests that a wide distribution of ribozyme activity may be the key to accurate predictions. Because the model predictions are based only on sequence and activity data, the results demonstrate that this machine learning approach allows readily obtainable experimental data to be used for RNA design efforts even for RNA molecules with unknown structures. The accurate prediction of RNA functions will enable a more comprehensive understanding of RNA fitness landscapes for studying evolution and for guiding RNA-based engineering efforts.

## Introduction

RNA enzymes, or ribozymes, are structured RNA molecules that catalyze biochemical reactions. One well-studied class of ribozymes are the small self-cleaving ribozymes that catalyze site specific cleavage of phosphate bonds in their own RNA backbone (Ferré-D'Amaré and Scott, 2010). These self-cleaving ribozymes are found in all domains of life, and their biological roles are still being investigated (Jimenez et al., 2015). In addition to their natural functions, these ribozymes have been used as the basis for engineering biological systems. For example, several small ribozymes (hammerhead, twister, pistol and HDV) have been used as genetically encoded gene regulatory elements by combining them with RNA aptamer and embedding them into untranslated regions of genes (Groher and Suess, 2014; Dykstra et al., 2022). This approach continues to gain attention because of the central importance of controlling gene expression and the simple design and build cycles of these small RNA elements. Nevertheless, ribozymes often need optimization for sequence dependent and cell specific effects. This can be achieved by modifying the sequence of the ribozymes, but this often requires multiple mutational changes and the vast sequence space requires extensive trial and error. Given this large sequence space, even the most high-throughput approaches can only find the optimal solutions present in the sequences that can be explored experimentally, which is a fraction of the total possible sequences. The engineering of ribozyme-based systems could benefit from accurate prediction of the effects of multiple mutations in order to narrow the search space towards optimal collections of sequences.

One way to think of the ribozyme optimization problem is in terms of fitness landscapes. Molecular fitness landscapes of protein and RNA molecules are studied by measuring the effects of numerous mutations on the function of a given reference molecule (Athavale et al., 2014; Blanco et al., 2019). Recently, the fitness landscapes of RNA molecules have been studied experimentally by synthesizing large numbers of sequences and using high-throughput sequencing to evaluate the relative activity of the RNA in vitro, or the growth effect of the RNA in a cellular system, both of which are termed “RNA fitness” (Kobori and Yokobayashi, 2016; Li et al., 2016; Pressman et al., 2019). The goal of in vitro evolution is often to find the highest peak in the landscape, or one of many high peaks, by introducing random mutations and selecting for improved activity. However, the RNA fitness landscapes that have been experimentally studied so far have revealed rugged topographies with peaks of high relative activity and adjacent valleys of low activity. Landscape ruggedness is an impediment to finding desired sequences through in vitro evolution approaches (Ferretti et al., 2018). Epistasis, defined as the non-additive effects of mutations, is the cause of ruggedness in fitness landscapes, and epistasis has been used to quantify the ruggedness of fitness landscapes (Szendro et al., 2013). More frequent and more extreme epistasis indicates that a landscape is more rugged. Importantly, more epistasis also means that the effect of combining multiple mutations is challenging to predict even if the effects of each individual mutation are known. In addition, experimental fitness landscapes can only study a limited number of sequences, except for very small RNA molecules (Pressman et al., 2019). It is often not possible to know if the process of in vitro evolution discovered a sequence that is globally optimal, or just a local optimum. For these reasons, it has become a goal to accurately predict the activity of sequences in order to streamline RNA evolution experiments and to study fitness landscapes in a more comprehensive manner (Groher et al., 2019; Schmidt and Smolke, 2021).

Here, we use high-throughput experimental data of mutational variants of a self-cleaving ribozyme to train a model for predicting the effect of higher-order combinations of three or more mutations. The ribozyme used in this study is the CPEB3 ribozyme (Figure 1A). This ribozyme is highly conserved in the genomes of mammals, where it is found in an intron of the CPEB3 gene (Salehi-Ashtiani et al., 2006). For training purposes, we generated a new data set that includes all possible individual and pairs of mutations to the reference CPEB3 ribozyme sequence (Figure 1B). These mutations were made by randomization of the CPEB3 ribozyme sequence with a 3% per nucleotide mutation rate during chemical synthesis of the DNA template. We reasoned that given the extensive amount of pairwise epistasis in RNA (Bendixsen et al., 2017), this data set might be sufficient for predicting higher-order mutants. In addition, we used a second, previously published data set that included 27,647 sequences comprised of random permutations of mutations found in mammals that include up to 13 mutational differences from the same reference ribozyme (Bendixsen et al., 2021). This second data set not only contains higher-order mutational combinations, but also a broad range of self-cleaving activity (Figure 1D). In both data sets, the relative activity of each sequence was determined by the deep sequencing of co-transcriptional self-cleavage data, as previously described. Briefly, the mutated DNA template was transcribed in vitro with T7 RNA polymerase. The transcripts were prepared for Illumina sequencing by reverse transcription and PCR. Relative activity was determined as the *fraction cleaved*, defined by the fraction of sequencing reads that mapped to a specific sequence variant in the shorter, cleaved form relative to the total number of reads for that sequence variant.

We set the goal of being able to predict the activity of the higher-order mutants in the phylogenetically derived fitness landscape (Figure 1D). In addition, we wanted to guide future experiments aimed at producing additional data for training models of ribozyme-based systems. The number of possible sequences increases exponentially with the number of variable nucleotide positions. In addition, the probability of finding active ribozymes at higher mutational distances becomes increasingly unlikely. Experiments aimed at training predictive models will need to choose realistic numbers of sequences that can have the highest impact on model performance. We therefore evaluated the effect of adding to the training data sequences with increasing mutational distances from the wild-type sequence as well as the effect of reducing the number of sequences in the training data. The results of these experiments were expected to be useful in guiding the choice of which sequence variants, and how many, to analyze experimentally in order to produce effective training data sets.

## Results

We first evaluated our new training data set that contained all single and double-mutants of the CPEB3 ribozyme. We found that the data did in fact contain full coverage of the possible 207 single mutants and the 21,114 double mutants. While the number of reads that mapped to each of these sequences varied, we found that, on average, 170 reads mapped to each double mutant, and ~18,000 reads mapped to each single mutant (Supplemental Figure 1). This read depth was sufficient for the determination of the fraction cleaved for all single and double mutants (Figure 1B). Mapping the fraction cleaved to base paired structural elements showed expected patterns of activity caused by compensatory base pairs. Mutations that break a base pair typically showed low activity, but a second mutation that restored the base pair showed high activity. To further evaluate this data, we calculated the non-additive pairwise epistasis in this data set (Figure 1C). Together, this analysis indicated that this data set contained a wide range of ribozyme activity and the effects of all pairwise intramolecular epistatic interactions.

In order to determine the training potential of the comprehensive double-mutant data, we first trained models using only the fraction cleaved data for sequences with two or fewer mutations including the wild-type reference sequence. We then tested the models' performance in predicting the fraction cleaved for sequences with increasing numbers of mutations. We trained two models with two approaches (see Materials and Methods). The first approach used a Random Forest regressor. In the second approach, we added a Long Short-Term Memory (LSTM) recurrent neural network to extract hidden features from the data. We then fed the hidden features with associated fraction cleaved to a Random Forest regressor. We will refer to this approach as "LSTM". We found that models trained on 2 or fewer mutations with Random Forest outperformed LSTM at predicting the activity of sequences with five or fewer mutations (Figure 2 A-C), but LSTM performed better when predicting the activity of sequences with six or more mutations relative to the wild-type (Figure 2 D-I). However, both approaches showed a decrease in the correlation between predicted and observed when challenged to predict the activity of sequences with higher numbers of mutations, and both resulted in relatively low correlation (Pearson $r < 0.7$) for sequences with seven or more mutations when trained only on this double mutant data (Figure 2 and Supplementary Table 2). We concluded that models trained on simple random mutagenesis containing all double mutants can be useful for predicting lower mutational distances, but we anticipated that additional data might improve the ability to predict the effect of higher numbers of mutations.

To determine the effect of adding higher-order mutants to the training data, we divided the phylogenetic derived sequence data by mutational distance and re-trained models with increasing orders of mutations in the training set. As expected, adding higher-order mutants improved the predicted to observed correlation at higher mutational distances (Figure 3 and Supplemental Figures 2-14). Interestingly, we found that the Random Forest approach outperformed the LSTM approach when sequences with more mutations were included in the training data. This is especially apparent for predicting the activity of sequences with 8-10 mutations. The Random Forrest approach resulted in models with high correlation between predicted and observed for all mutational distances when trained with data from sequences with four or more mutations (Figure 3 A-C). For both approaches, the largest improvements in the correlations occurred when sequences with 3 mutations (relative to wild-type) were added to the data. Subsequently appending additional sequences with greater numbers of mutations had diminishing improvements on the correlation. We note that all the testing data was set aside prior to training and identical testing data was used for all models. The results demonstrate that adding higher order mutants to the training data improves the Pearson correlation of sequences at higher distances in this data set. It is important to note that the phylogenetically derived data has different numbers of sequences for each class of mutations (Table 1), and sequences with higher numbers of mutations in our data show mostly low activity (Supplementary Figure

15). This helps interpret the effect of sequentially adding higher-order mutant sequences to the training data. It is also important to note that the phylogenetic derived sequences only contain mutations at thirteen different positions. The higher order sequences in this data are therefore combinations of the lower order sequences. For example, a sequence with six mutations can be constructed by combining two sequences with three mutations, both of which would be in the “3 mutations” training data. Our model is therefore predicting the effects of combining sets of mutations, and adding precise sets of lower order mutations that re-occur in higher order mutations clearly improves the correlations between prediction and experimental observation in our data.

In order to inform future experiments for collecting training data, we next set out to determine the effect of decreasing the amount of data in the training sets. Starting from the 80% of data used as prior training data, we randomly sampled sequences from this data to create new training data sets with 60%, 40%, 20%, 10% and 1% of the total data. These subsampled data sets were used to train models using the random forest regressor. The same testing data was set aside for all models and used to compare the Pearson correlation coefficient of each model trained with decreasing amounts of data. As an illustrative example, we focused on a model trained with sequences with 5 or fewer mutations relative to wild-type used to predict the activity of sequences with 7 mutations (Figure 4 and Supplementary Table 1). We chose this example because it achieved very high correlation (Pearson *r* = 0.99) when trained with 80% (25,733 unique sequences) of the data and therefore provided an opportunity to observe how rapidly the correlation decreased with less data. We found that the models trained on 5 or fewer mutations predicted with high correlation when as little as 40% (12,866) of the data was used for training (Pearson $r$ = 0.97). With only 20% (6,433) and 10% (3,217) of the data, the model still showed good prediction accuracy with a Pearson correlation $r \cong 0.9$. Surprisingly, we still observed reasonably high correlation when including only 1% (322) of the training data, and this was reproducible over five different models trained with different random samples of the data (Pearson $r$ = 0.81, *stdev* = 0.046, $n = 5$). Similar results were observed with other training and testing scenarios. To illustrate general trends, we have plotted the Pearson correlation for the same model trained on 5 or fewer mutations when predicting the activity of sequences with 6, 7, 8 or 9 mutations, and for a model trained on 9 or fewer mutations used to predict sequences with 5, 6, 7, or 8 mutations (Figure 4). This analysis suggests that the total amount of training data is not critical for predicting the activity of sequences in our data set. When combined with the diminishing returns of adding more higher order mutations (Figure 3), this analysis emphasizes the importance of collecting appropriate experimental data sets for training that include ribozymes with more mutations that still maintain relatively high activity. However, given the low probability of finding higher-order sequences with higher activity, an iterative approach with several cycles of predicting and testing might be necessary to acquire such data.

While the primary goal was to predict the relative activity of RNA sequences, we wondered if the models might also be useful for predicting structurally important nucleotides. To address this question, we analyzed the “feature importance” in several of our Random Forest models. Feature importance is a method to assign importance to specific input data. Because our data only uses sequence as input, the features in our data are specific nucleotides (A, G, C or U) at specific positions. We found that for the

Random Forest models, the most important feature all clustered around the active site of the ribozyme (Supplemental Figures 16 and 17). Further, the CPEB3 ribozyme uses metal ion catalysis and several of the most important features were nucleotides that have been observed coordinated to the active site magnesium ion in the CPEB3 ribozyme, or the analogous nucleotides in the structurally similar HDV ribozyme (Kapral et al., 2014; Skilandat et al., 2016). For example, for all the models trained with some higher order mutants, the most important feature was G1, which positions the cleaved phosphate bond in contact with the catalytic magnesium ion. The second most important feature was G25, which forms a wobble base pair with U20 (Lévesque et al., 2012), another important feature (top 4-6), and this nucleotide pair coordinates the active site magnesium ion through outer sphere contacts. The catalytic nucleotide C57 binds the same catalytic magnesium as the G25:U20 wobble pair, and had a high feature importance similar to U20. Most of the other important features are involved in base pairs that stack or interact with the metal ion coordinating bases. Interestingly, we found that nine of the ten most important features were identical for models trained with only single and double mutants or with increasing amounts of higher-order mutants. However, the G1 and G25 features became increasingly more important as sequences with higher mutational distance were added to the training data. This indicates that the higher-order mutants in the training data helped emphasize structurally critical nucleotides. We conclude that the machine learning models presented identified nucleotides involved in forming the active sites of the CPEB3 ribozyme. Because we did not use structural data to train our models, the results suggest that similar data could identify active sites in RNA molecules with unknown structures.

## Discussion

We have shown that a model trained on ribozyme activity data can accurately predict the self-cleavage activity of sequences with numerous mutations. This approach can be used to guide experiments based on a relatively small set of initial data. Importantly, the approach did not use structural information such as X-ray crystallography or cryo-EM, and used only sequence and activity data, which can be obtained with common molecular biology approaches (in vitro transcription, RT-PCR, and sequencing). In addition, the training data starts with small amounts of synthetic DNA. The comprehensive double mutant data and the phylogenetic derived data each started from a single DNA oligo synthesis that used doped phosphoramidites at the variable positions. Each data set was collected on a single lane of an Illumina sequencer. The approach presented in this paper is therefore accessible, rapid and inexpensive as compared to approaches that use structural data to train their models.

Sequence conservation of naturally occurring RNA molecules has been another useful data type for training models to predict RNA structure from sequence (De Leonardis et al., 2015; Weinreb et al., 2016). This approach is based on the observation that nucleotide positions that form a base pair often show co-evolutionary patterns of sequence conservation. In some cases, this co-evolutionary data has been combined with thermodynamic predictions or structural data from chemical probing, such as SHAPE experiments (Calonaci et al., 2020). Numerous ribozymes, aptamers and aptazymes have been discovered through in vitro evolution experiments and conservation data is not available unless sequencing experiments were applied during the selection process. Our approach could be used to expand functional information of non-natural RNA molecules which could then be used to guide structure prediction of these molecules in a way similar to how naturally occurring sequence conservation has been used. In addition, sequence conservation does not necessarily predict relative activity. For example, while the CPEB3 ribozyme is highly conserved in nature, not all of the sequence are equally proficient at catalyzing self-cleavage (Chadalavada et al., 2010; Bendixsen et al., 2021). Our approach using machine learning from experimentally derived data may prove useful for guiding experiments with non-natural RNA molecules discovered through in vitro selection or SELEX-like approaches. However, adopting this machine learning approach will require that each experimenter acquire specific data for their system necessary to train and test sequences with the functions they are investigating.

With future work, it may be possible to produce more general models of ribozyme activity. For example, a model trained on data sets from several different self-cleaving ribozymes with different nucleotide lengths might learn to predict the activity of sequences of arbitrary length and sequence composition. In fact, recent advances in RNA structure prediction have used the crystal structures of several different self-cleaving ribozymes as training data to develop predictive modes that achieve near-atomic level resolution of arbitrary sequences (Townshend et al., 2021). Alternatively, models trained on ribozymes with different activities beyond self-cleavage might be able to classify sequences as ribozymes of various functions. There has been some success with generating general models for predicting protein functions. The latent features identified by deep generative models of protein sequences are being used to better understand the complex, higher-order amino acid interactions necessary to achieve a functional protein structure (Riesselman et al., 2018; Detlefsen et al., 2022). We hypothesize that latent features could aid in the identification of generalized parameters that govern the epistatic interactions of higher-order mutants of RNA sequences as well. We hope that the accuracy and accessibility of the approach presented here will inspire others to carry out similar experiments and initiate the data sharing that will be needed to develop more general models, similar to what is being accomplished for protein functional predictions (Biswas et al., 2021).

One challenge to our predictive models appears to be the low frequency of active sequences at higher mutational distances. In our phylogenetically derived data the vast majority of sequences have very low activity (Figure 1D), and the probability of finding sequence with high fraction cleaved decreases with the number of mutations relative to wild-type. As a consequence, models trained on lower-order mutant variants tend to overestimate the activity of sequences at higher mutational distances. It has been previously observed that experimental RNA fitness landscapes are dominated by *negative epistasis,* which means that mutations in combination tend to have lower fitness than would be expected from the additive effects of individual mutations (Bendixsen et al., 2017). The overestimation of fraction cleaved at higher mutational distances suggests that our models have a difficult time learning to predict negative epistasis. It has been previously observed that mutations with “neutral” or “beneficial” effects on protein function often have destabilizing effects on protein structure (Soskine and Tawfik, 2010). We postulate that the same effect is causing negative epistasis in the RNA data. This suggests that additional information, such as measurements or estimates of thermodynamic stability of helices, might be necessary for increasing accuracy at even higher distances beyond those offered by this data set (Groher et al., 2019; Yamagami et al., 2019). For example, we have recently demonstrated that our sequencing based approach to measuring ribozyme activity can be extended to include magnesium titrations in order to evaluate RNA folding/stability (Peri et al., 2022). In the future, combining structural and functional information might be the best approach to accurately design RNA molecules with desired functional properties.

## Materials and Methods

### Ribozyme activity data

Ribozyme activity was determined as previously described (Bendixsen et al., 2021). Briefly, DNA templates were synthesized with the promoter for T7 RNA polymerase to enable in vitro transcription. Templates were synthesized with mixtures of phosphoramidites at variable positions. For the comprehensive double-mutant data set, templates were synthesized with 97% wild-type nucleotides and 1% each of the other three nucleotides. For the phylogenetic derived data set, the template was synthesized with an equal mixture of the naturally occurring nucleotides that were found at 13 positions that varied across 99 mammalian genomes. During in vitro transcription, RNA molecules self-cleaved at different rates. The reaction was stopped at 30 minutes, and the RNA was concentrated and reverse transcribed with a 5’-RACE protocol that appends a new primer site to the cDNA of both cleaved and uncleaved RNA (SMARTScribe, Takara). The cDNA was PCR amplified with primers that add the adaptors for Illumina sequencing. This procedure was done in triplicate with unique dual-indexes for each replicate. DNA was combined equimolar and sent for sequencing (GC3F, University of Oregon). Sequencing was performed on a single lane of a HiSeq 4000 using paired-end 150 reads.

## Ribozyme activity from sequence data

FastQ sequencing data were analyzed using custom Julia and Python scripts. Briefly, the scripts identified the reverse transcription primer binding site at the 3'-end to determine nucleotide positions and then determined if the sequence was cleaved or uncleaved by the absence or presence of the 5'-upstream sequence. For the single and double mutants, all possible sequences were generated and stored in a list, and reads that matched the list elements were counted and cleaved or uncleaved was determined by the presence or absence of the 5'-upstream sequence. For the phylogenetically derived data, nucleotide identities were determined at the expected 13 variable positions by counting the string character position from the fixed regions. Sequencing reads were discarded if they contained unexpected mutations in the primer binding site, the uncleaved portion, or the ribozyme sequence. For each unique genotype in the library the number of cleaved and uncleaved sequences were counted and ribozyme activity (fraction cleaved) was calculated as *fraction cleaved* = $\mathrm{counts}_{\mathrm{cleaved}}/(\mathrm{counts}_{\mathrm{cleaved}} + \mathrm{counts}_{\mathrm{uncleaved}})$.

## Machine Learning

Random Forest regression uses an ensemble of decision trees to improve prediction accuracy. Each tree in the ensemble is created by partitioning the sequences within a sample into groups possessing little variation. Each sample is drawn with replacement and the resulting trees are aggregated into forests that best predict the cleavage rates of the sequences. The Random Forest regression was performed using the python package scikit-learn. Each sequence was transformed into a 69 by 4 one-hot encoding representation of the sequence. Each of the four possible nucleotides within the sequence was represented by a vector of length 4 possessing a uniquely located "1" within the vector to signify the nucleotide's identity. Each sequence in the training set was fit using scikit-learn's RandomForestRegressor ensemble module. Feature importance was computed via a forest of randomized trees using the *features_importances* function in the module under default settings. Briefly, the relative importance of a feature was determined by the depth of the feature when it was used as a decision node in a tree. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples they contributed to was used as an estimate of the relative importance of the features.

LSTM is a recurrent neural network commonly used for the predictive modeling of written text data, which has sequential dependencies. Here we used an LSTM to compute a set of hidden features given a set of nucleotide sequences. These hidden features are learned by the LSTM in a supervised way for the purpose of relating the nucleotide sequence to the corresponding ribozyme activity (fraction cleaved). The LSTM network has an architecture where each cell *C* outputs the next state $h_t$ ($1 \leq t \leq n$) by taking in input from the previous state $h_{t-1}$ and the embedding $x_t$ of the current nucleotide in the sequence. The output $h_n$ of the last cell of the LSTM is then used as input to a Random Forest regressor to predict the sequence functional activity rate. The LSTM model was built using PyTorch's open-source machine learning framework. Sequences were trained using an LSTM layer with 32 hidden dimensions and a dropout rate of 0.2. Each sequence was embedded in a 69 by 4 tensor (where 4 is the size of the nucleotide embedding) and then batched in groups of 64 sequences for input to the model. The gradient descent was performed using PyTorch's built-in Adam optimizer and MSELoss criterion. Twenty-five training epochs were performed on each training set.

Training and Test Data

The data set containing the fraction cleaved data from the 27,647 phylogenetically derived sequences was binned based on the number of mutations relative to the wild-type ribozyme. For each bin, a portion of the data (20%) was chosen at random and set aside as test data. This resulted in test data sets that were also separated by the number of mutations relative to the wild-type sequence. Training data sets were created from the 80% of data in each mutational bin that was not set aside for testing. Training data sets were created by combining bins at a given number of mutations to all the bins with lower numbers of mutations. Training data included 100% of the single and double mutant data. For reduced training sets were created by randomly sampling different numbers of sequences from the original full training data sets.

## Data Availability

Sequencing reads in FastQ format are available at ENA (PRJEB51631). Sequence and activity data and computer code is available at GitLab (https://gitlab.com/bsu/biocompute-public/ml-ribo-predict.git).

## Author Contributions

JB – involved in conceptualizing the project, managed data, performed computational work for formal analysis and visualization, reviewed and edited the manuscript; JR – involved in conceptualizing the project, performed experiments, managed the project, supervised and facilitated computational work, prepared figures, reviewed and edited the manuscript, JK – involved in conceptualizing the project, performed computational work for formal analysis and validation; AT– was involved in conceptualizing the project, helped with experimental validation, supervised and facilitated computational work;  ES – Involved in conceptualizing the project, supervised computational work, reviewed and edited the manuscript; FS – Involved in conceptualizing the project, supervised computational work, reviewed and edited the manuscript; EH – involved in conceptualizing the project, supervised experimental work, supervised computational work, wrote the original draft, and reviewed and edited the manuscript.

## Funding

The authors acknowledge funding from the National Science Foundation (EH, grant number OIA-1738865, OIA-1826801, REU site grant #1950599), National Aeronautics and Space Administration (EH, grant number 80NSSC17K0738), and the Human Frontier Science Program (EH, Ref.-No: RGY0077/2019).

## Acknowledgments

We thank Devin Bendixsen for producing the fitness landscape image used in the manuscript and for valuable discussions.

## References

Athavale, S. S., Spicer, B., and Chen, I. A. (2014). Experimental fitness landscapes to understand the molecular evolution of RNA-based life. *Current Opinion in Chemical Biology* 22, 35–39. doi: 10.1016/j.cbpa.2014.09.008.

Bendixsen, D. P., Østman, B., and Hayden, E. J. (2017). Negative Epistasis in Experimental RNA Fitness Landscapes. *J. Mol. Evol.* doi: 10.1007/s00239-017-9817-5.

Bendixsen, D. P., Pollock, T. B., Peri, G., and Hayden, E. J. (2021). Experimental resurrection of ancestral mammalian CPEB3 ribozymes reveals deep functional conservation. *Mol Biol Evol*. doi: 10.1093/molbev/msab074.

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. (2021). Low-N protein engineering with data-efficient deep learning. *Nat Methods* 18, 389–396. doi: 10.1038/s41592-021-01100-y.

Blanco, C., Janzen, E., Pressman, A., Saha, R., and Chen, I. A. (2019). Molecular Fitness Landscapes from High-Coverage Sequence Profiling. *Annu Rev Biophys*. doi: 10.1146/annurev-biophys-052118-115333.

Calonaci, N., Jones, A., Cuturello, F., Sattler, M., and Bussi, G. (2020). Machine learning a model for RNA structure prediction. *NAR Genomics and Bioinformatics* 2, lqaa090. doi: 10.1093/nargab/lqaa090.

Chadalavada, D. M., Gratton, E. A., and Bevilacqua, P. C. (2010). The Human HDV-like CPEB3 Ribozyme Is Intrinsically Fast-Reacting. *Biochemistry* 49, 5321–5330. doi: 10.1021/bi100434c.

De Leonardis, E., Lutz, B., Ratz, S., Cocco, S., Monasson, R., Schug, A., et al. (2015). Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res* 43, 10444–10455. doi: 10.1093/nar/gkv932.

Detlefsen, N. S., Hauberg, S., and Boomsma, W. (2022). Learning meaningful representations of protein sequences. *Nat Commun* 13, 1914. doi: 10.1038/s41467-022-29443-w.

Dykstra, P. B., Kaplan, M., and Smolke, C. D. (2022). Engineering synthetic RNA devices for cell control. *Nature Reviews Genetics*, 1–14.

Ferré-D'Amaré, A. R., and Scott, W. G. (2010). Small self-cleaving ribozymes. *Cold Spring Harbor perspectives in biology* 2, a003574.

Ferretti, L., Weinreich, D., Tajima, F., and Achaz, G. (2018). Evolutionary constraints in fitness landscapes. *Heredity* 121, 466–481. doi: 10.1038/s41437-018-0110-1.

Groher, A.-C., Jager, S., Schneider, C., Groher, F., Hamacher, K., and Suess, B. (2019). Tuning the Performance of Synthetic Riboswitches using Machine Learning. *ACS Synth. Biol.* 8, 34–44. doi: 10.1021/acssynbio.8b00207.

Groher, F., and Suess, B. (2014). Synthetic riboswitches—a tool comes of age. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1839, 964–973.

Jimenez, R. M., Polanco, J. A., and Lupták, A. (2015). Chemistry and Biology of Self-Cleaving Ribozymes. *Trends in Biochemical Sciences* 40, 648–661. doi: 10.1016/j.tibs.2015.09.001.

Kapral, G. J., Jain, S., Noeske, J., Doudna, J. A., Richardson, D. C., and Richardson, J. S. (2014). New tools provide a second look at HDV ribozyme structure, dynamics and cleavage. *Nucleic Acids Research* 42, 12833–12846. doi: 10.1093/nar/gku992.

Kobori, S., and Yokobayashi, Y. (2016). High-Throughput Mutational Analysis of a Twister Ribozyme. *Angew. Chem. Int. Ed. Engl.* 55, 10354–10357. doi: 10.1002/anie.201605470.

Lévesque, D., Reymond, C., and Perreault, J.-P. (2012). Characterization of the Trans Watson-Crick GU Base Pair Located in the Catalytic Core of the Antigenomic HDV Ribozyme. *PLoS ONE* 7, e40309. doi: 10.1371/journal.pone.0040309.

Li, C., Qian, W., Maclean, C. J., and Zhang, J. (2016). The fitness landscape of a tRNA gene. *Science*, aae0568. doi: 10.1126/science.aae0568.

Peri, G., Gibard, C., Shults, N. H., Crossin, K., and Hayden, E. J. (2022). Dynamic RNA fitness landscapes of a group I ribozyme during changes to the experimental environment. *Molecular Biology and Evolution*, msab373. doi: 10.1093/molbev/msab373.

Pressman, A. D., Liu, Z., Janzen, E., Blanco, C., Müller, U. F., Joyce, G. F., et al. (2019). Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA. *J. Am. Chem. Soc.* 141, 6213–6223. doi: 10.1021/jacs.8b13298.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* 15, 816–822. doi: 10.1038/s41592-018-0138-4.

Salehi-Ashtiani, K., Lupták, A., Litovchick, A., and Szostak, J. W. (2006). A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene. *Science* 313, 1788–1792. doi: 10.1126/science.1129308.

Schmidt, C. M., and Smolke, C. D. (2021). A convolutional neural network for the prediction and forward design of ribozyme-based gene-control elements. *eLife* 10, e59697. doi: 10.7554/eLife.59697.

Skilandat, M., Rowinska-Zyrek, M., and Sigel, R. K. O. (2016). Secondary structure confirmation and localization of Mg2+ ions in the mammalian CPEB3 ribozyme. *RNA* 22, 750–763. doi: 10.1261/rna.053843.115.

Soskine, M., and Tawfik, D. S. (2010). Mutational effects and the evolution of new protein functions. *Nat. Rev. Genet* 11, 572–582. doi: 10.1038/nrg2808.

Szendro, I. G., Schenk, M. F., Franke, J., Krug, J., and Visser, J. A. G. M. de (2013). Quantitative analyses of empirical fitness landscapes. *J. Stat. Mech.* 2013, P01005. doi: 10.1088/1742-5468/2013/01/P01005.

Townshend, R. J. L., Eismann, S., Watkins, A. M., Rangan, R., Karelina, M., Das, R., et al. (2021). Geometric deep learning of RNA structure. *Science* 373, 1047–1051. doi: 10.1126/science.abe5650.

Weinreb, C., Riesselman, A., Ingraham, J. B., Gross, T., Sander, C., and Marks, D. S. (2016). 3D RNA and functional interactions from evolutionary couplings. *Cell* 165, 963–975. doi: 10.1016/j.cell.2016.03.030.

Yamagami, R., Kayedkhordeh, M., Mathews, D. H., and Bevilacqua, P. C. (2019). Design of highly active double-pseudoknotted ribozymes: a combined computational and experimental study. *Nucleic Acids Research* 47, 29–42. doi: 10.1093/nar/gky1118.

# Figures

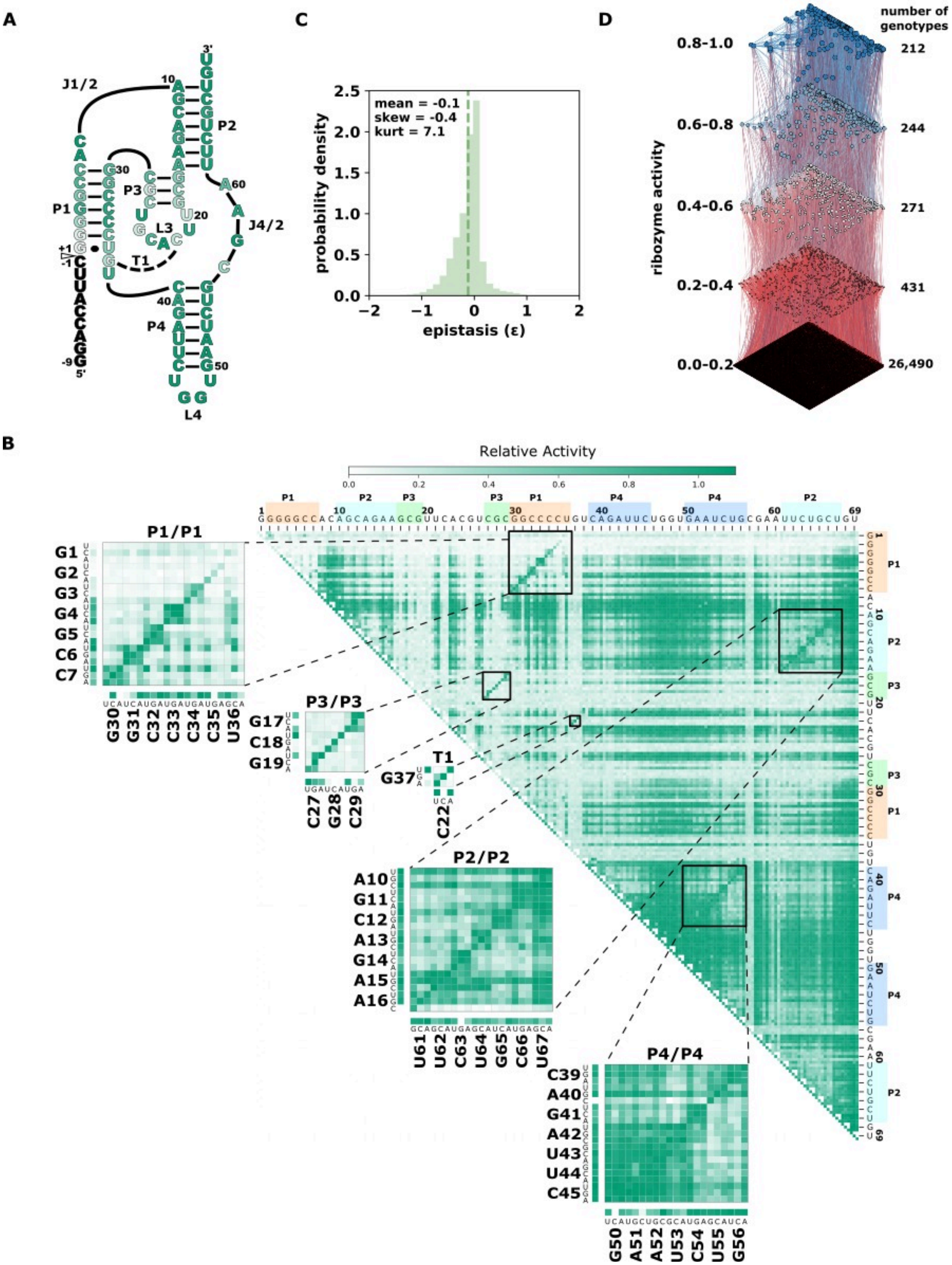


**Figure 3. 1. The CPEB3 ribozyme and data prediction challenge.**

(A) Secondary structure diagram of the CPEB3 ribozyme. The white arrow indicates the site of self-cleavage. Nucleotide color indicates the average relative activity of the three possible point mutations at each position. (B) Heatmap representation of comprehensive single and double-mutant data. Each pixel in the heatmap shows the ribozyme activity for a specific double mutant indicated by the nucleotide positions on the top and right of the heatmap. Insets show base paired regions and specific mutations. Ribozyme activity is determined as the fraction of total reads that map to each sequence that are in the cleaved form (fraction cleaved) relative to the wildtype fraction cleaved. (C) Distribution of pairwise epistasis from double mutant data. Epistasis was calculated as $\varepsilon = \log_{10} (W_{AB} * W_{wt} / W_A * W_B)$, where $W_{wt}$ is the fraction cleaved of the wild-type ribozyme, $W_A$ and $W_B$ are fraction cleaved of sequences with individual mutations and $W_{AB}$ is the fraction cleaved of the sequence with both individual mutations. (D) Higher mutational distance variants of the CPEB3 ribozyme represented as a fitness landscape. *Ribozyme activity* (fraction cleaved) is shown for 27,647 sequence variants derived from permutations of naturally occurring mutations. Each node represents a different sequence and the size and color of the node is scaled to the ribozyme activity. Edges connect nodes that differ by a single mutation. Sequences are binned into quintiles of ribozyme activity and the *number of genotypes* reports the number of sequences in each quintile.

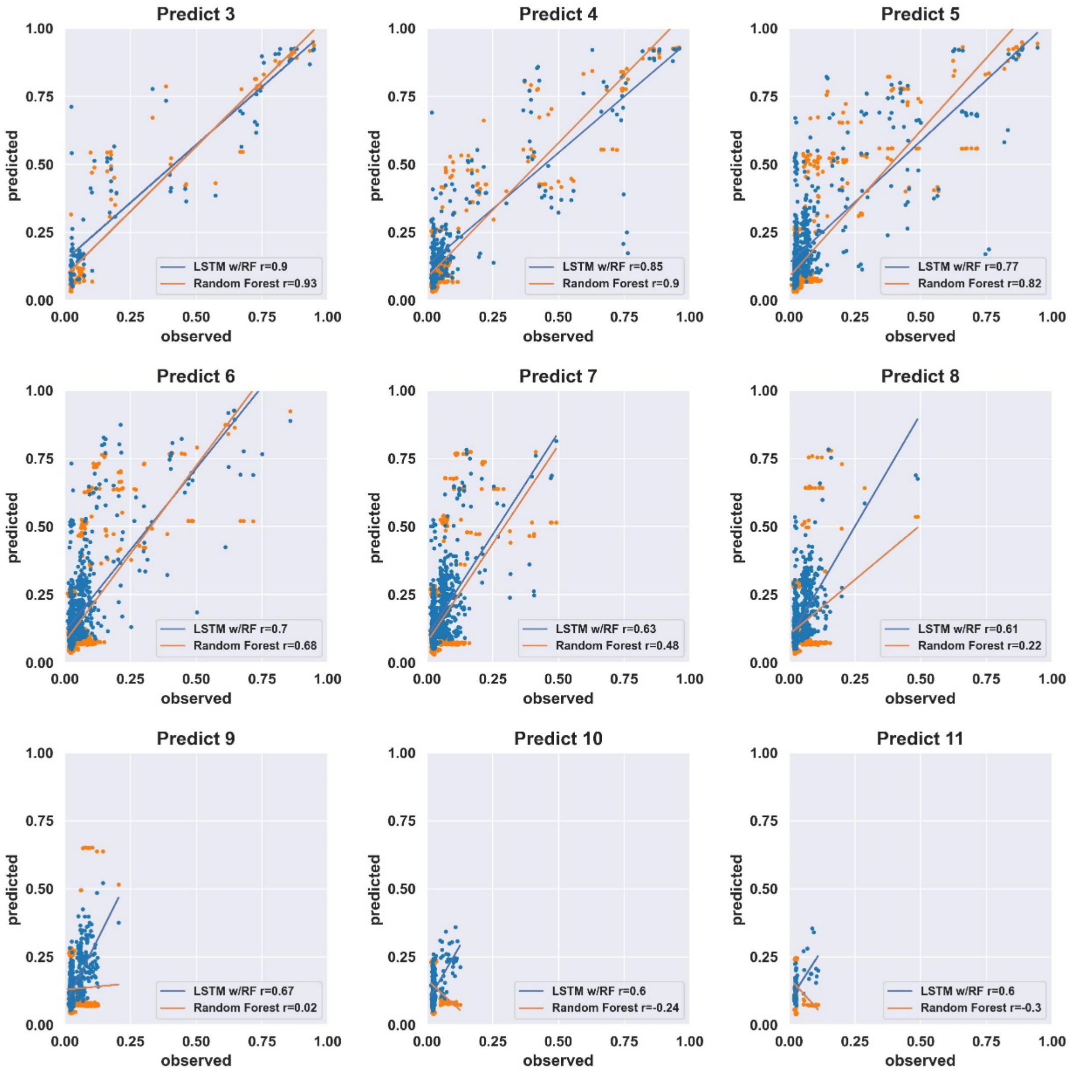


**Figure 3.2. Prediction accuracy of models trained on comprehensive individual and pairs of mutations.**

Scatter plots of *Predicted* (fraction cleaved from the models) and *Observed* (fraction cleaved from experiments). The models were trained on the experimentally determined fraction cleaved for the wild-type and all possible sequences with one mutation (207 sequences) or two mutations (21,114 sequences). Pearson correlation coefficients *r* reported for the model trained by LSTM-RF (blue) and the Random Forest (orange) approach. The sequences are separated by the number of mutations relative to the wild-type, as indicated by the title of each graph.

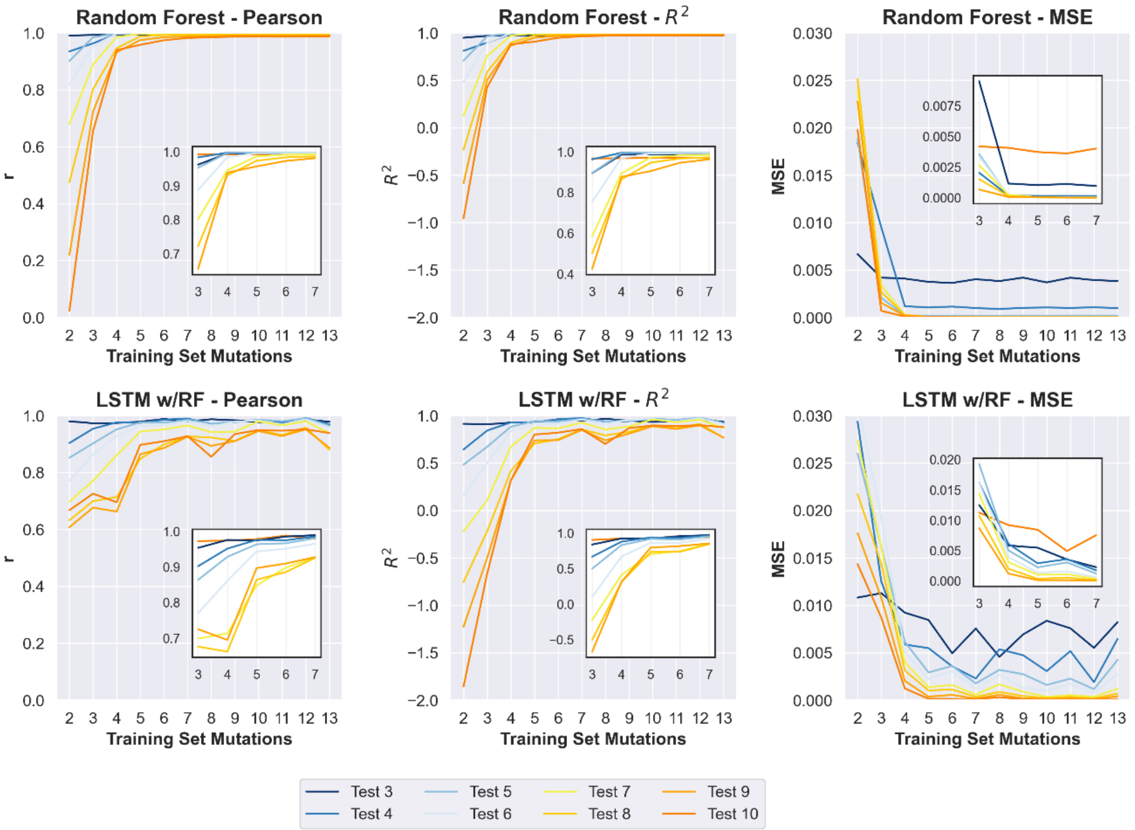


**Figure 3.3. Improvement in prediction accuracy from increased mutational distances in the training data.**

Changes in Pearson *r*, $R^2$, and mean squared error (MSE) of prediction-observed correlation (y-axis) with increasing numbers of max mutations within the training data (x-axis). Training sets included all sequences up to and including the y-axis value. For each plot, colors indicate the numbers of mutations in sequences in the test data (see key). Insets show changes to the same prediction accuracy measurement with the 3-7 mutation training data, to allow more visual resolution.

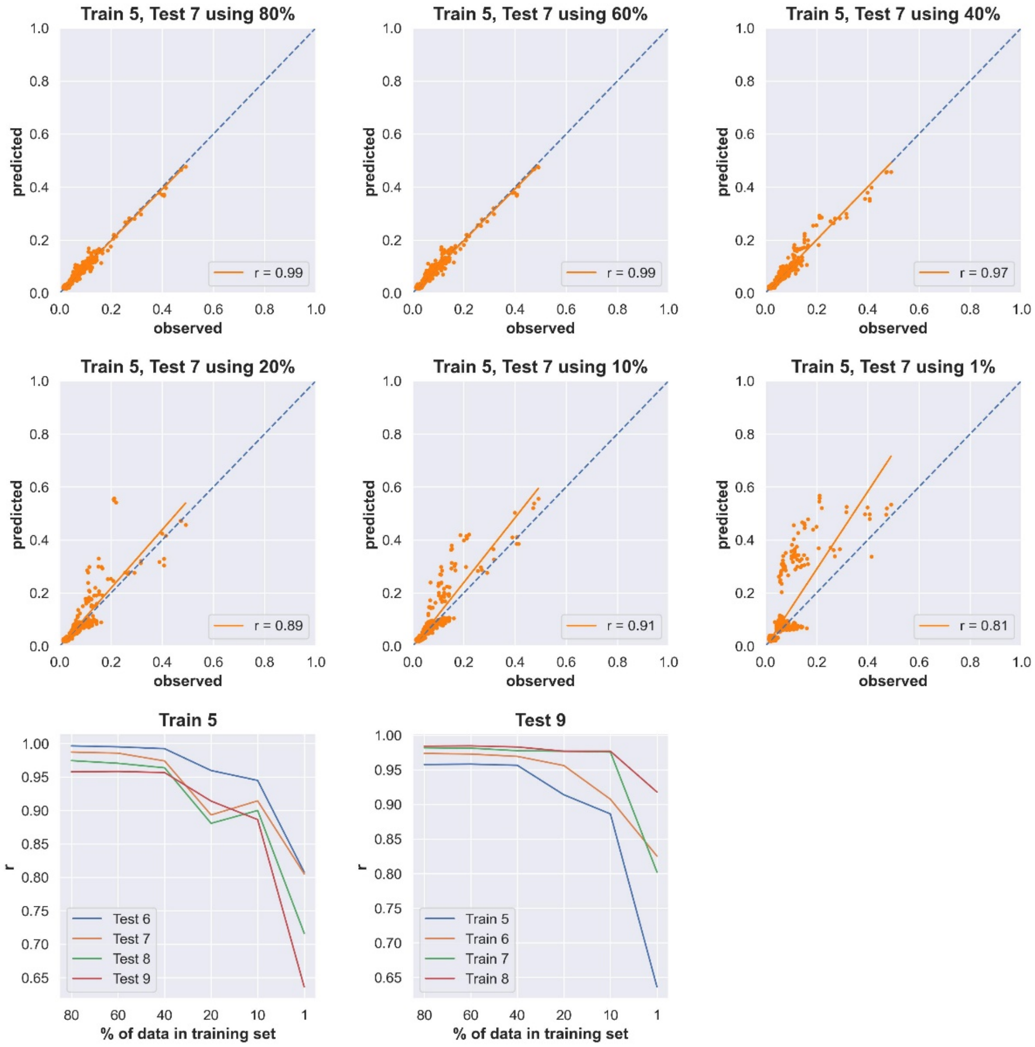
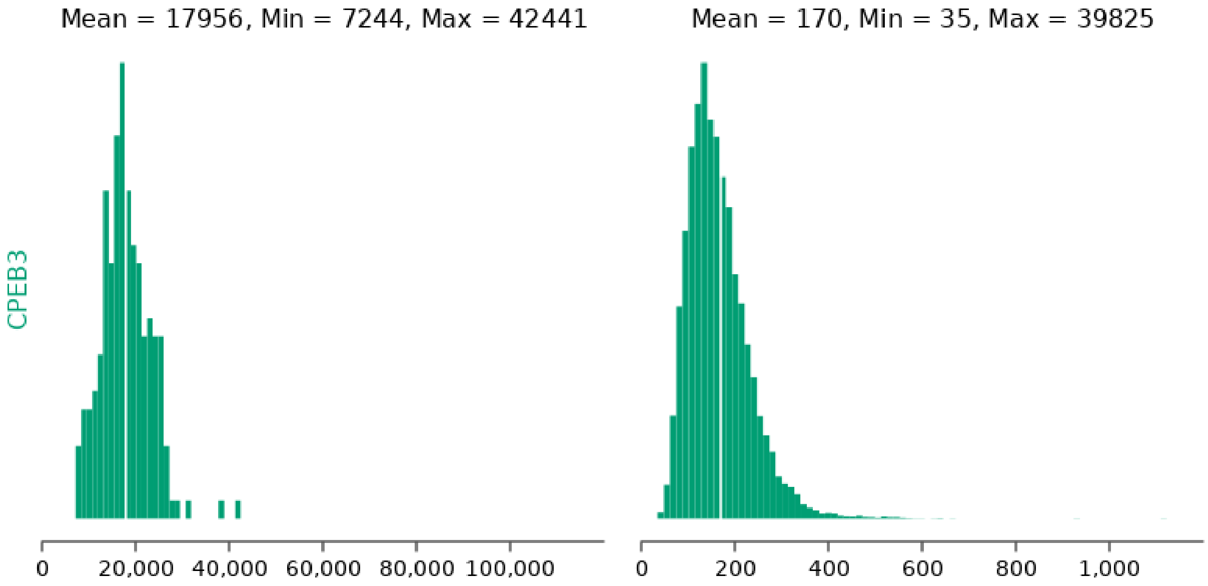


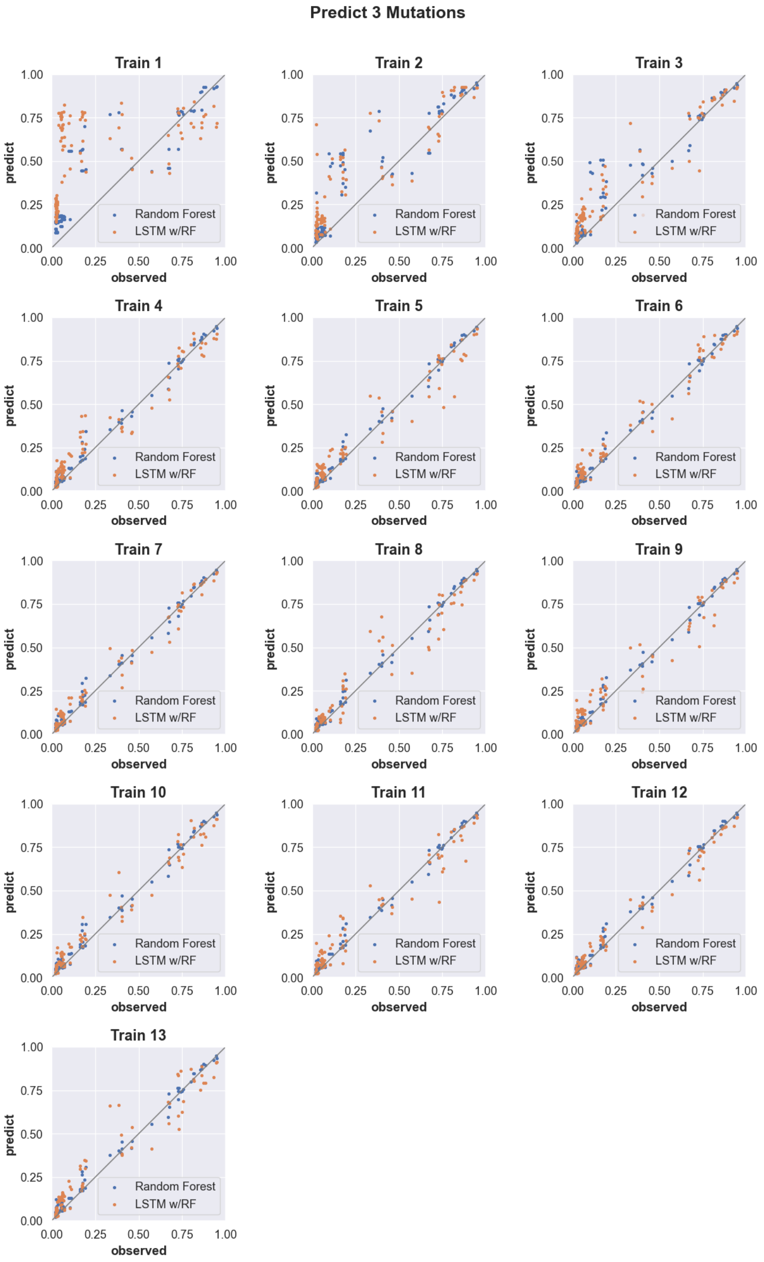
**Figure 3.4. Effects of reducing the number of sequences in the training data.** Scatter plots of *Predicted* (fraction cleaved from the models) and *Observed* (fraction cleaved from experiments). Shown are the results for models trained with decreasing amounts of sequences with 5 or fewer mutations using the random forest approach and predicting the fraction cleaved of sequences with 7 mutations. The percent of the total sequences used in the training data is indicated in the title of each plot. Pearson correlation coefficients are indicated as insets. The results of training with 5 or fewer mutations (Train 5) on different test data sets (6-9) and 9 or fewer mutations (Test 9) on different train data sets (6-8) are also shown.

Table 3.1. **Counts of sequences in training and testing data sets.**

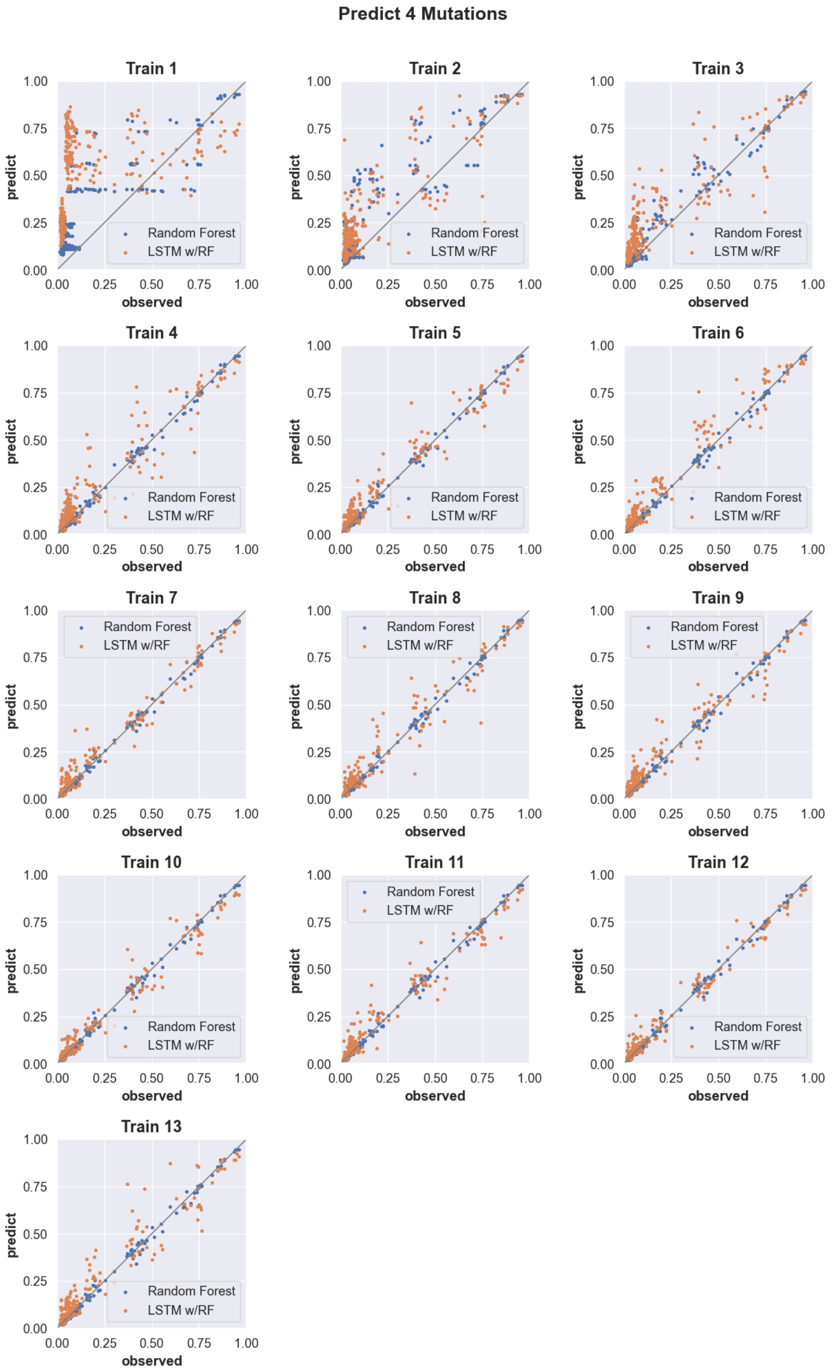
| No. of mutations | Training | Testing |
|---|---|---|
| 1 | 207 | --- |
| 2 | 21114 | --- |
| 3 | 414 | 104 |
| 4 | 1240 | 310 |
| 5 | 2650 | 662 |
| 6 | 4162 | 1040 |
| 7 | 4867 | 1217 |
| 8 | 4241 | 1060 |
| 9 | 2720 | 680 |
| 10 | 1249 | 312 |
| 11 | 389 | 97 |
| 12 | 74 | 18 |
| 13 | 6 | 2 |



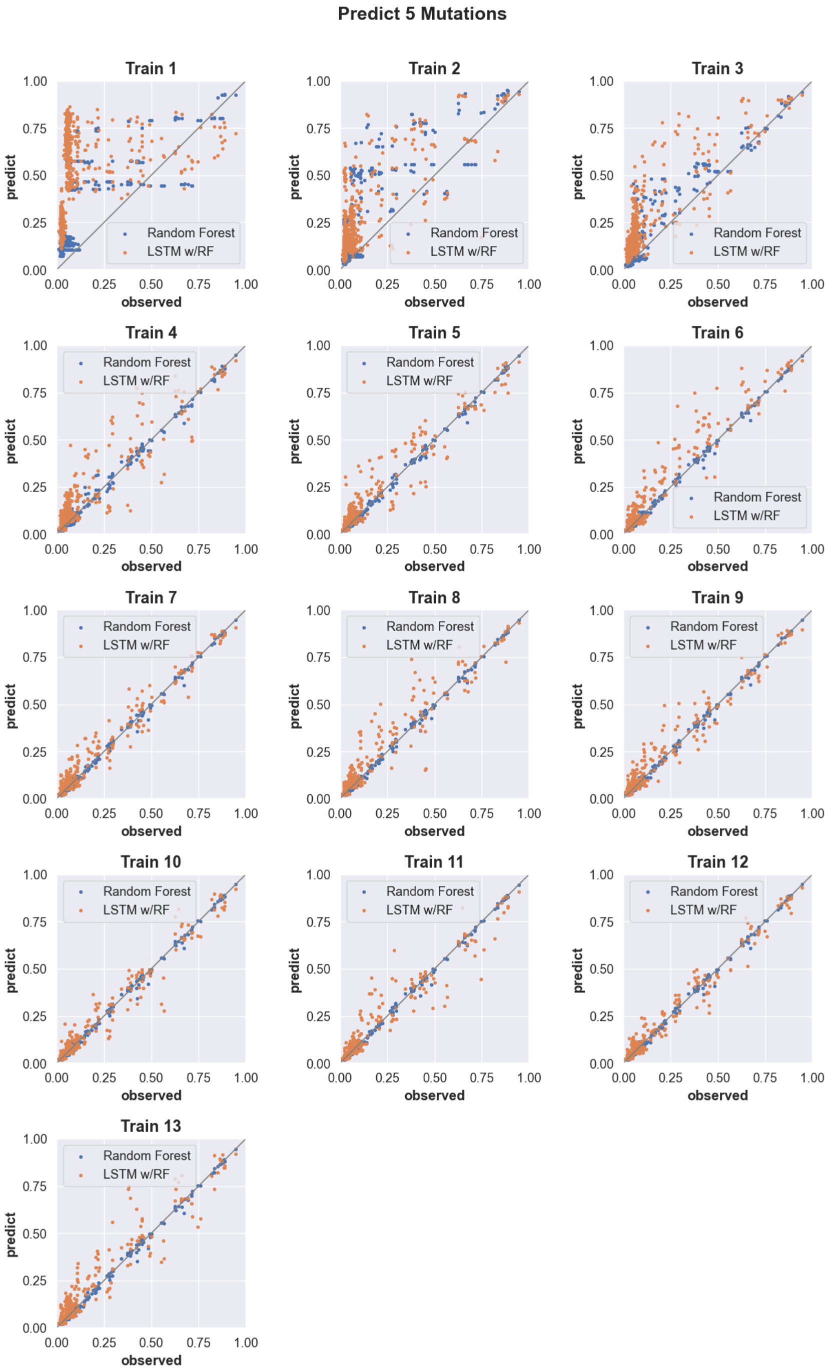
**Supplementary Figure 3.1. Histogram of CPEB3 variant counts for single (left) and double (right) mutants. Mean, minimum and maximum values for each distribution are indicated.**



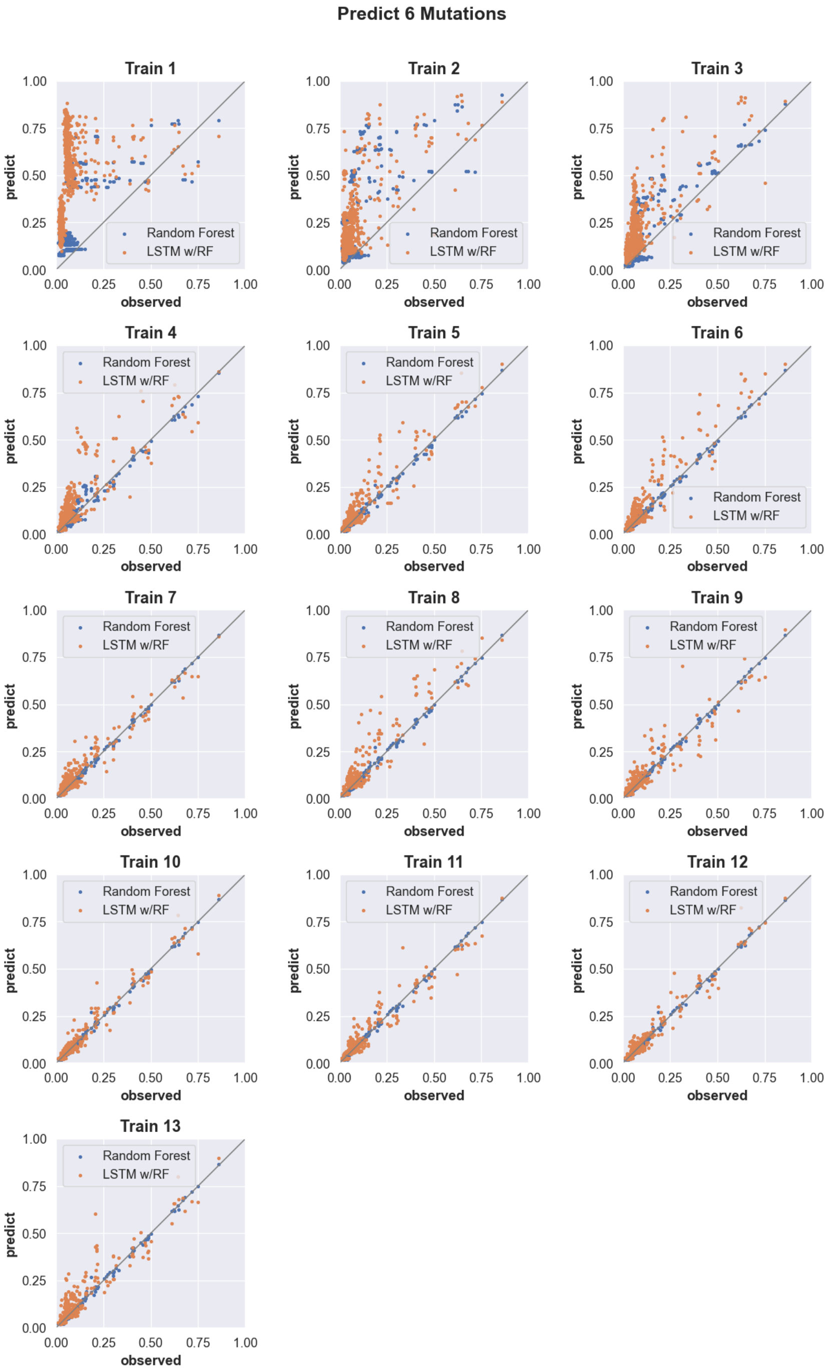
**Supplementary Figure 3.2. Three-mutation sequence activity predictions.** Scatter plots comparing fraction cleaved values measured from experiments (observed) to those predicted by models (predict) trained by either random forest (blue) or the LSTM approach (orange). Each scatter plot shows the predictions from a different training data set. The training data contained sequences with up to the number of mutations in the title (Train N). For example, 'Train 5' indicates that the model was trained using data for sequences containing 1,2,3,4, and 5 mutations. The line indicates unity, not a fit to the data.



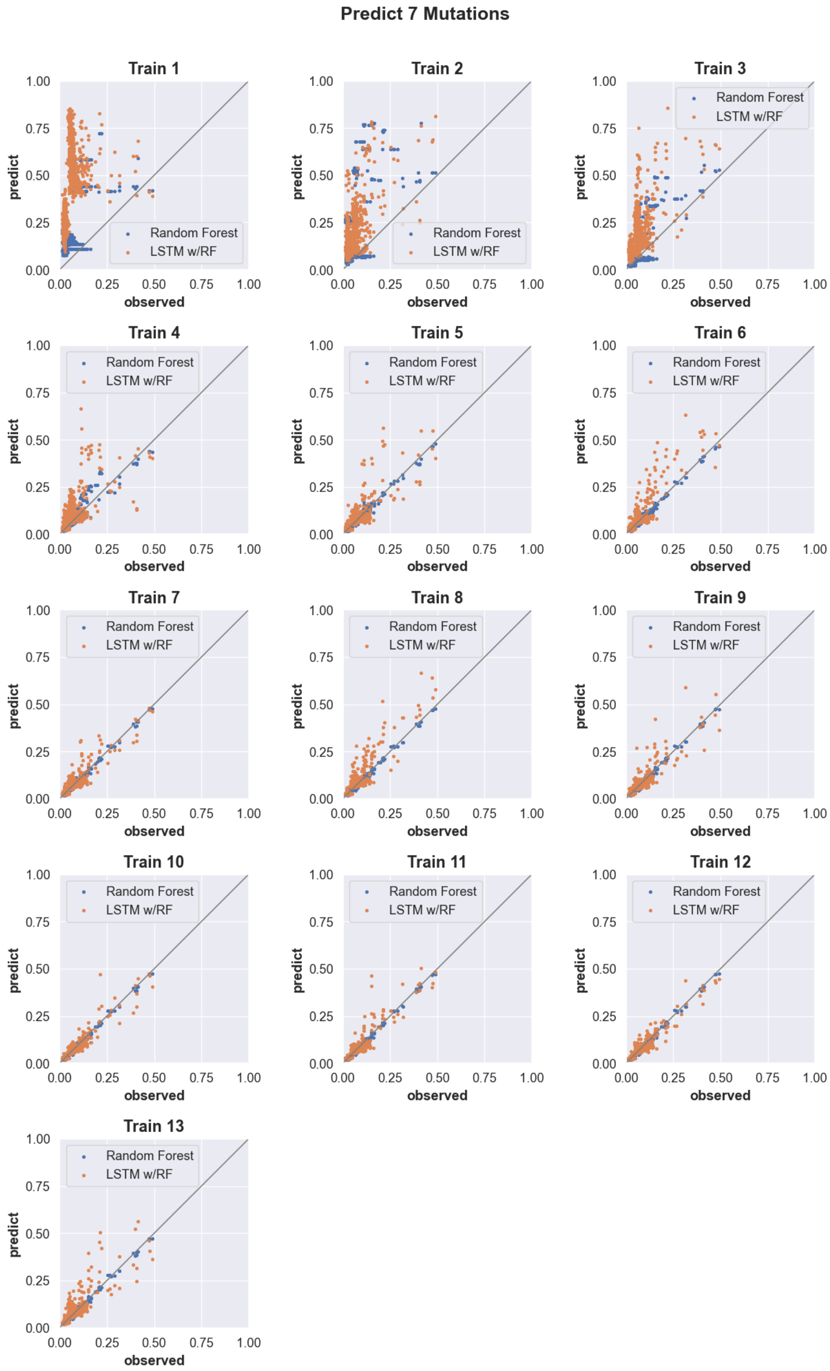
**Supplementary Figure 3.3. Predicting the activity of sequences with four mutations (see Supp. Fig. 3.2 for details).**



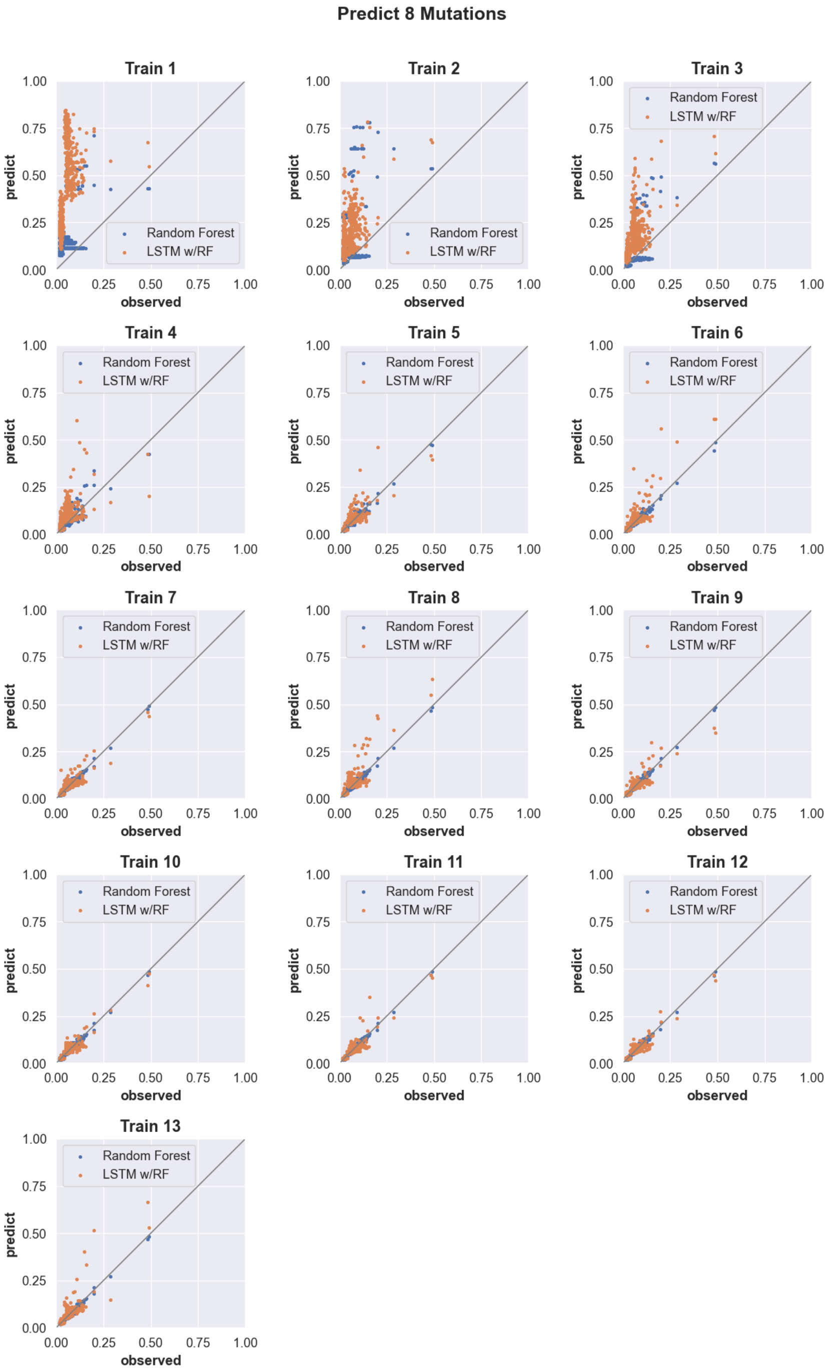
**Supplementary Figure 3.4. Predicting the activity of sequences with five mutations. (see Supp. Fig. 3.2 for details).**



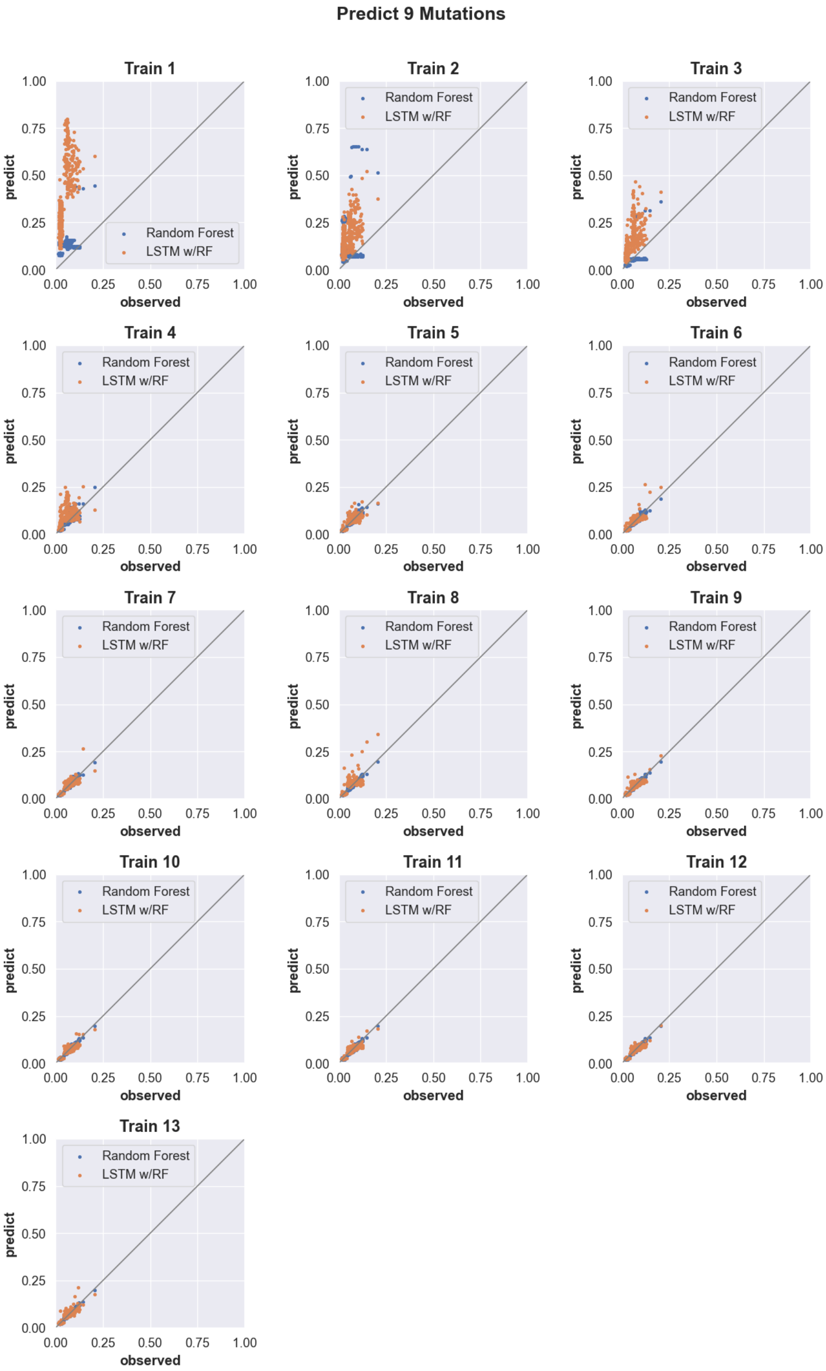
**Supplementary Figure 3.5. Predicting the activity of sequences with six mutations (see Supp. Fig. 3.2 for details).**



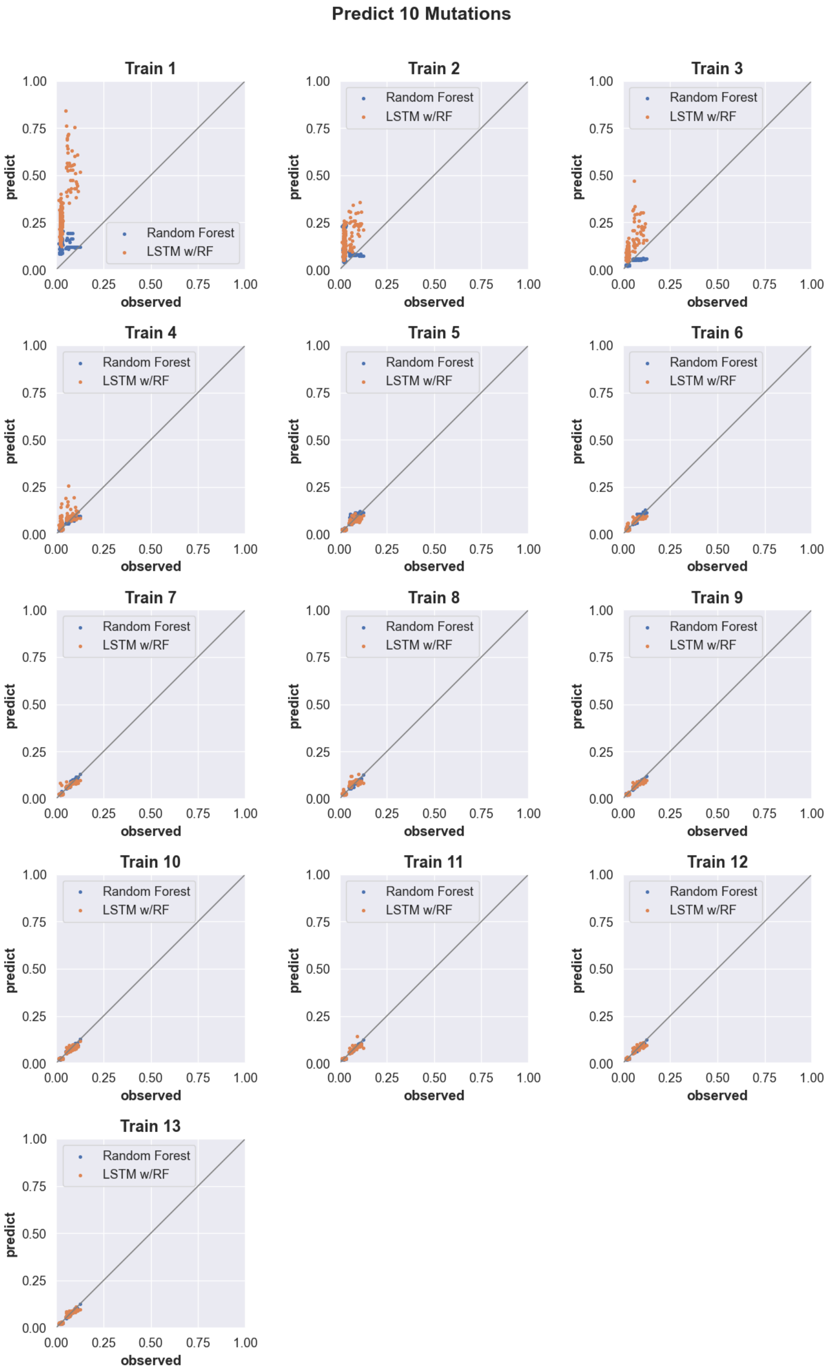
**Supplementary Figure 3.6. Predicting the activity of sequences with seven mutations (see Supp. Fig. 3.2 for details).**



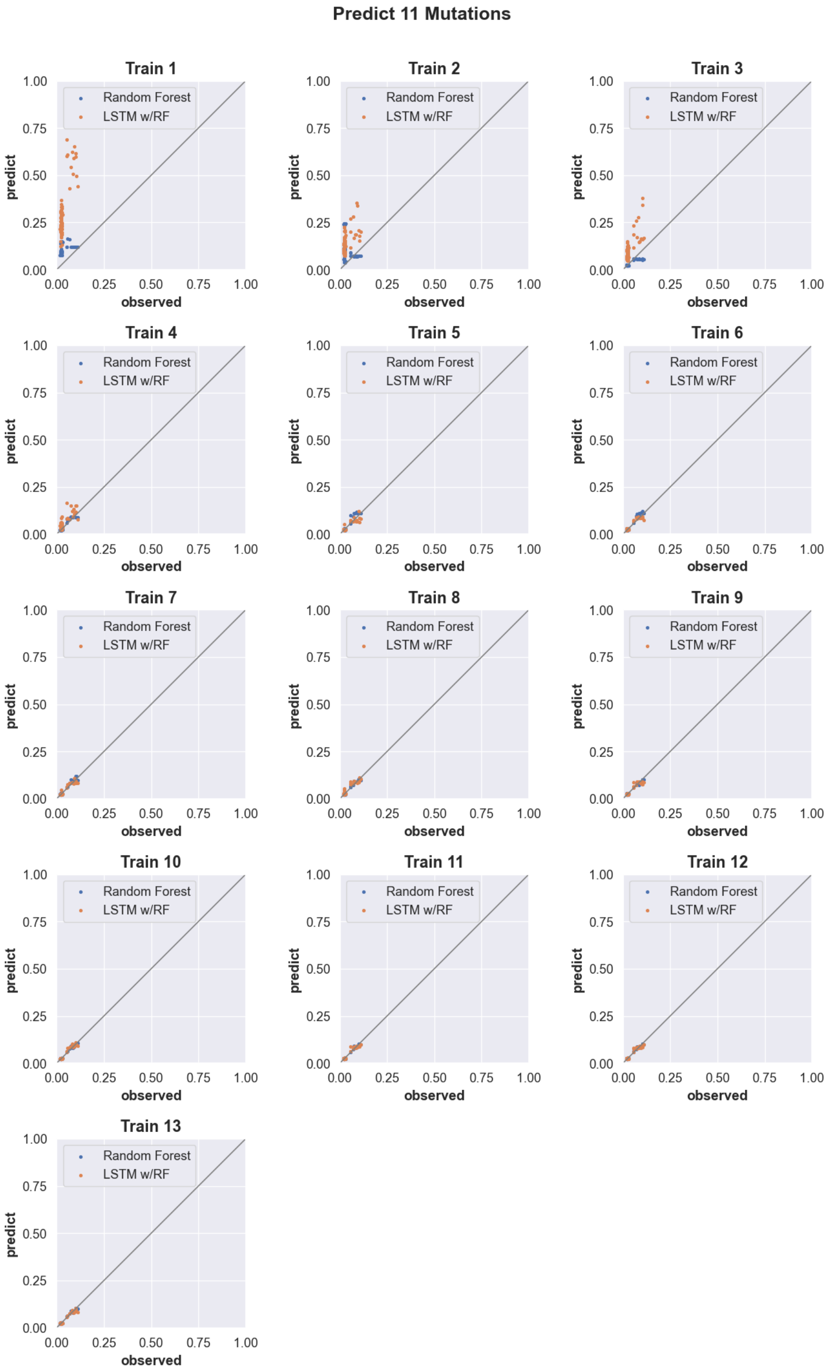
**Supplementary Figure 3.7. Predicting the activity of sequences with eight mutations (see Supp. Fig. 3.2 for details).**



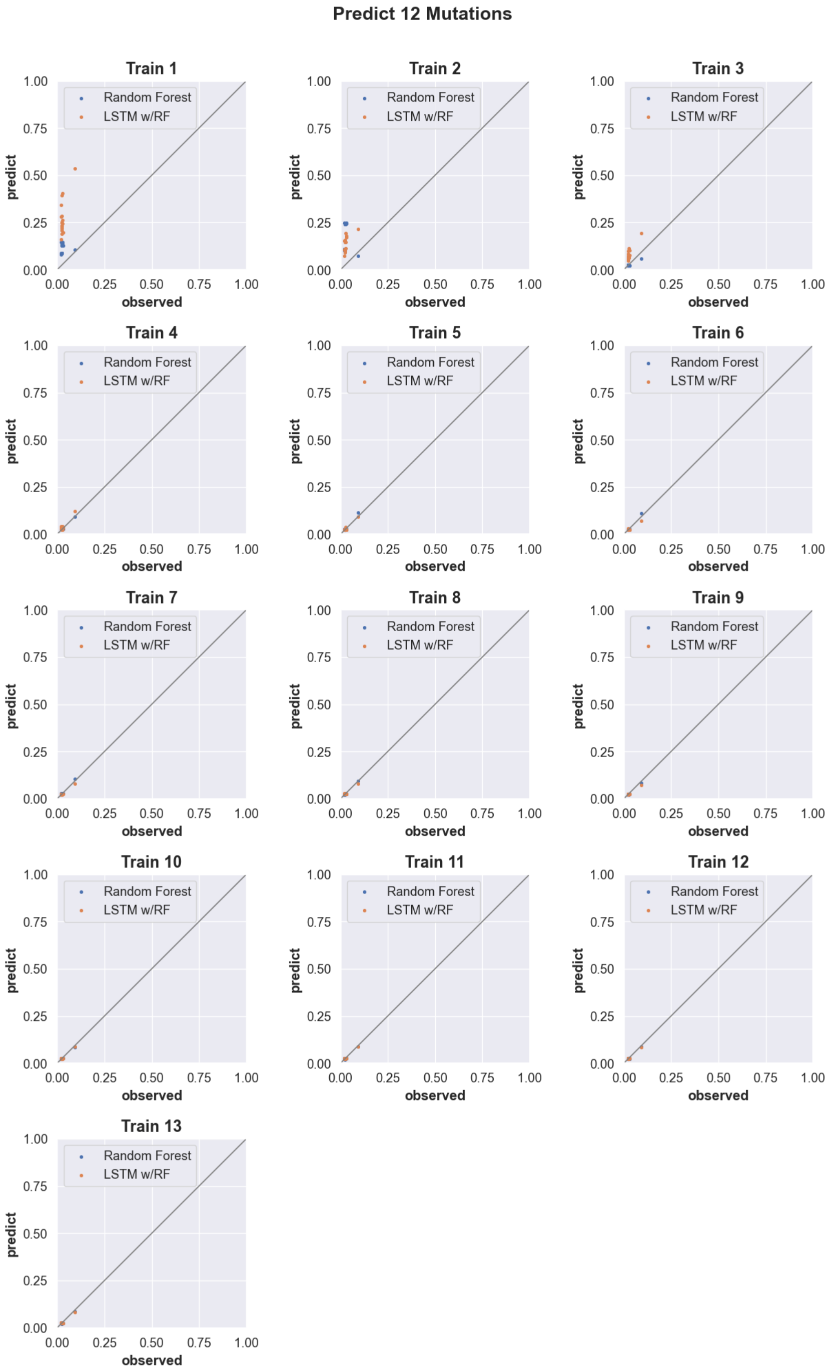
**Supplementary Figure 3.8. Predicting the activity of sequences with nine mutations (see Supp. Fig. 3.2 for details).**



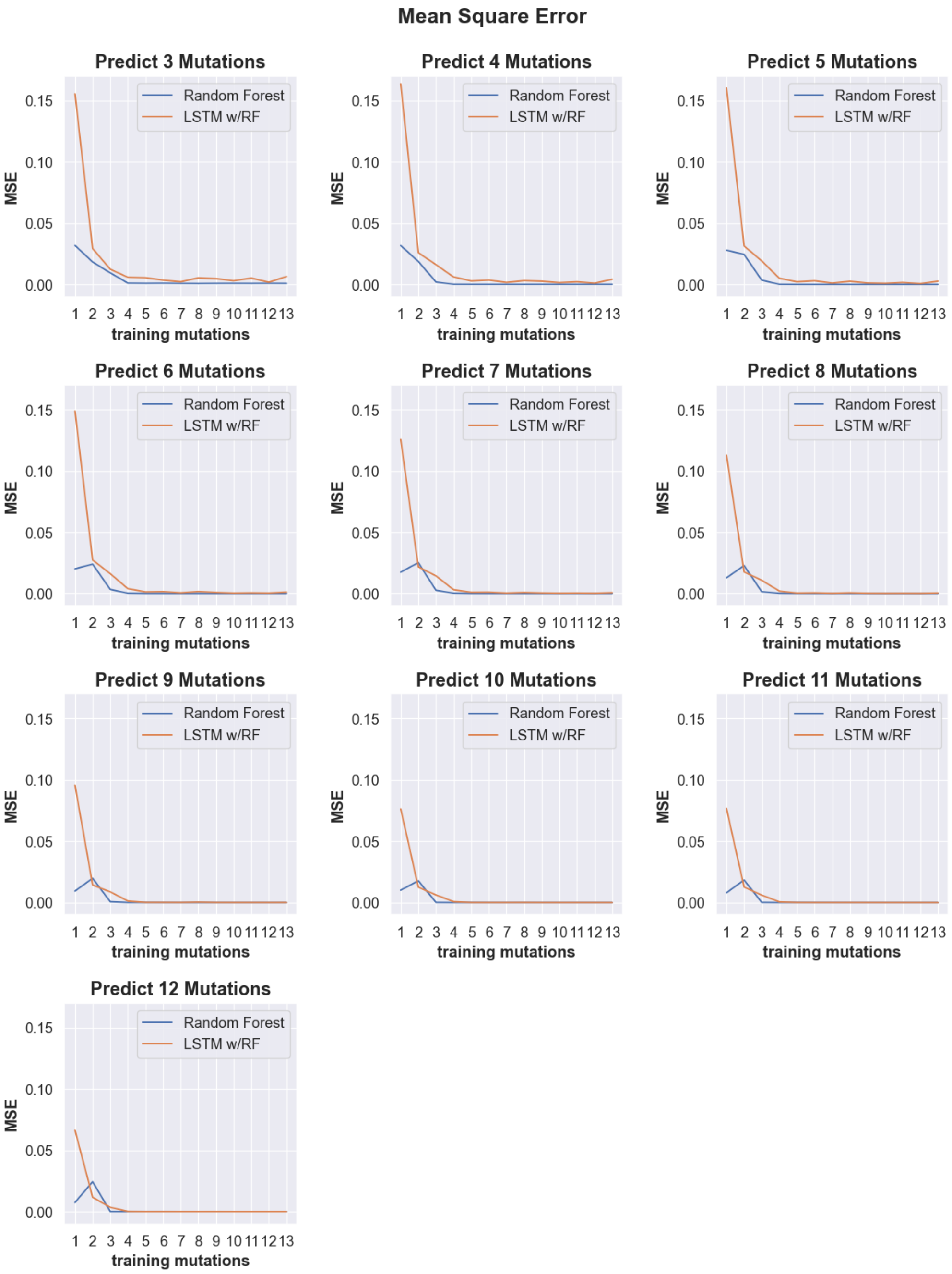
**Supplementary Figure 3.9. Predicting the activity of sequences with 10 mutations (see Supp. Fig. 3.2 for details).**



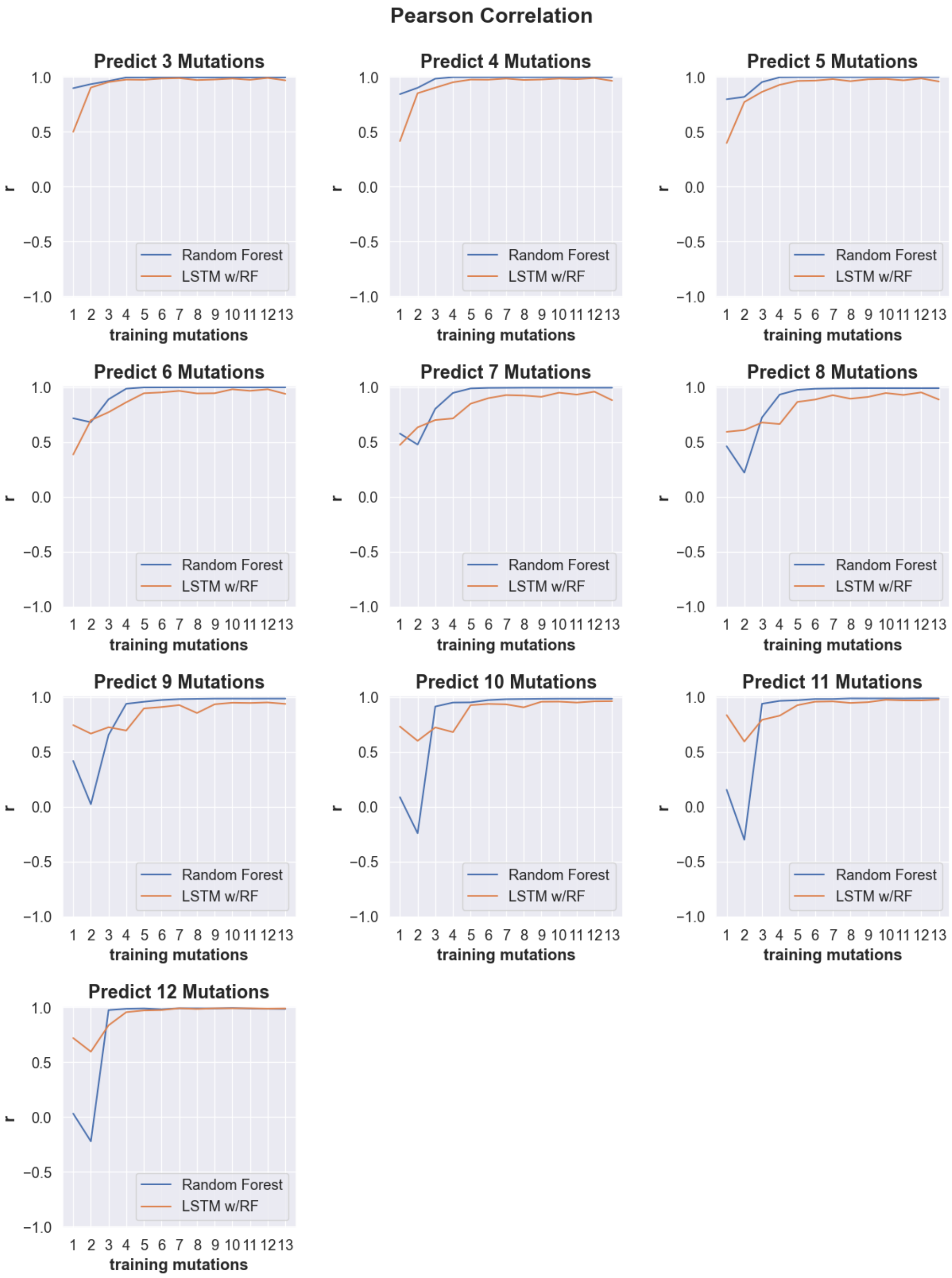
**Supplementary Figure 3.10. Predicting the activity of sequences with 11 mutations (see Supp. Fig. 3.2 for details).**



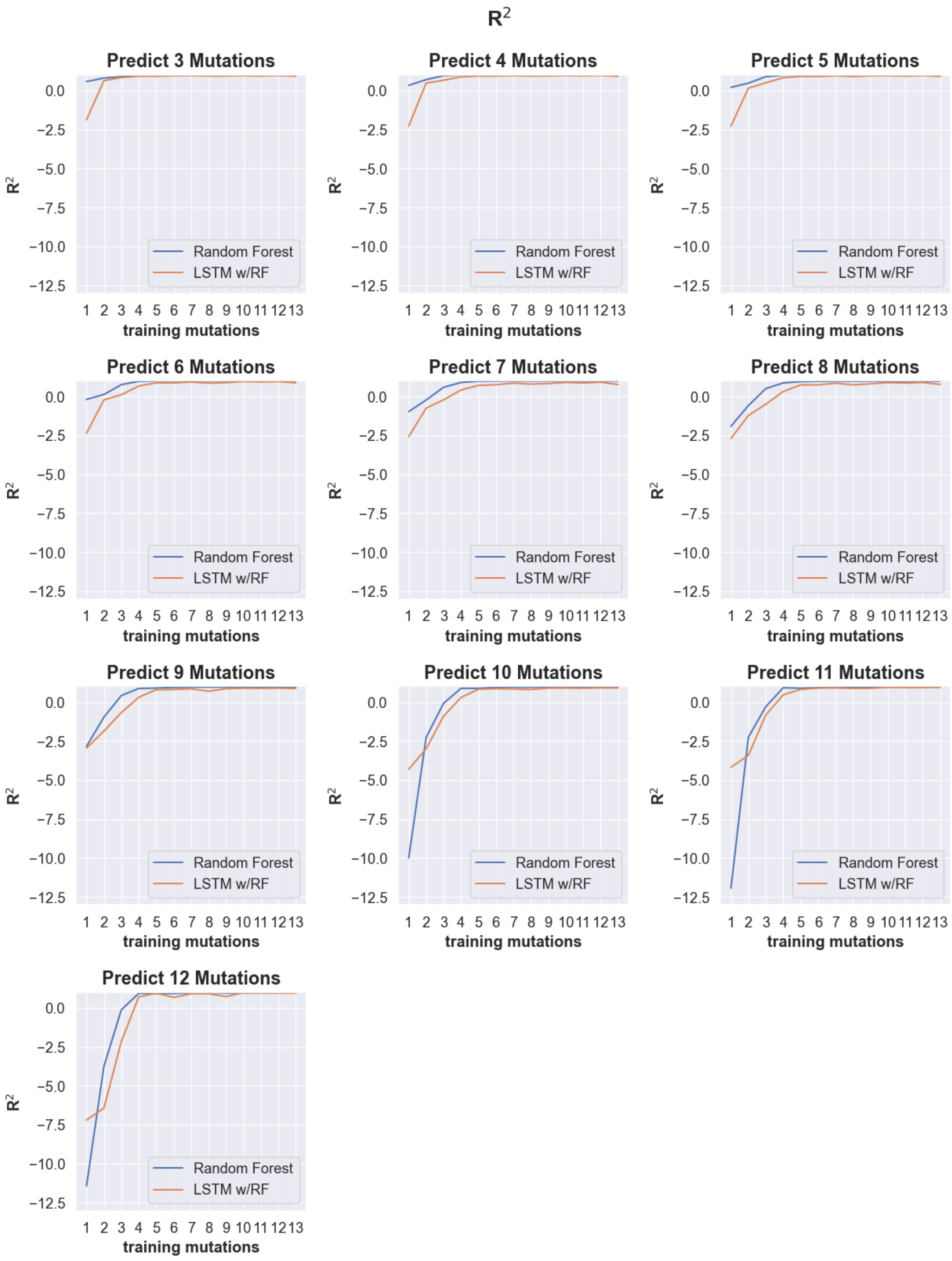
**Supplementary Figure 3.11. Predicting the activity of sequences with 12 mutations (see Supp. Fig. 3.2 for details).**



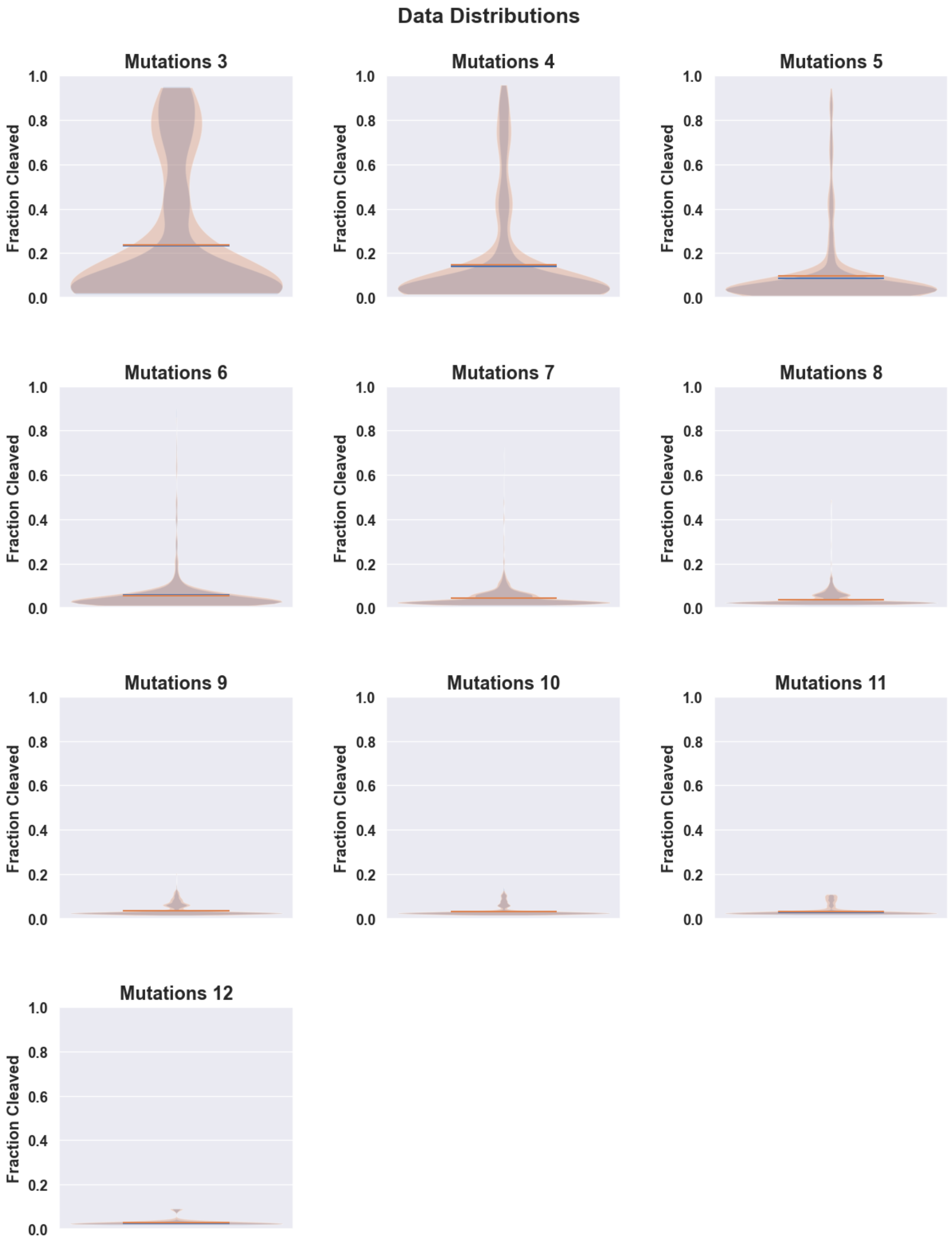
**Supplementary Figure 3.12. Line plots showing the mean square error (MSE) of predicted cleavage activity values obtained from random forest (blue) and LSTM with random forest (orange) machine learning models trained on data with incrementally increasing numbers of mutations shown along the x-axis. Each plot shows the MSE for predictions obtained for sequences containing the number of mutations indicated by the plot title.**



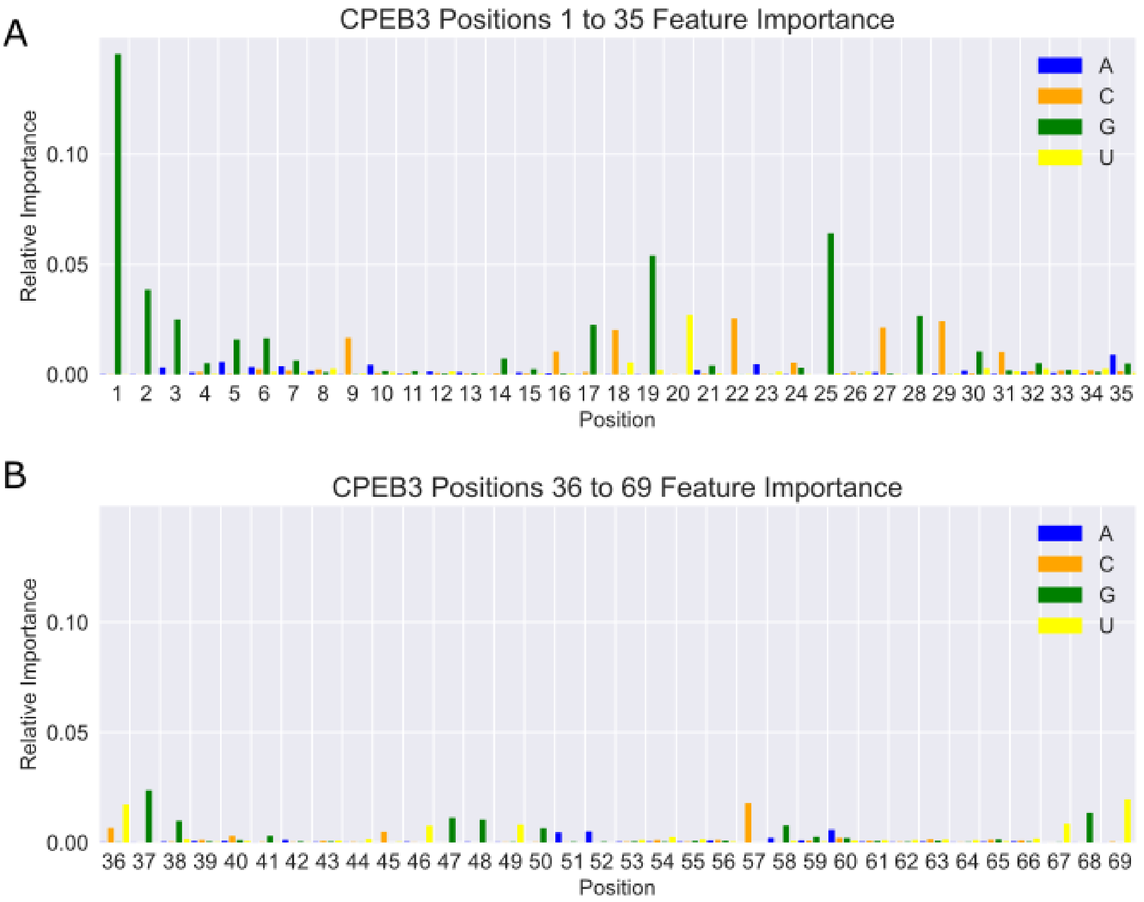
**Supplementary Figure 3.13.** Line plots showing the Pearson correlation values of predicted cleavage activity obtained from random forest (blue) and LSTM with random forest (orange) machine learning models trained on data with incrementally increasing numbers of mutations shown along the x-axis. Each plot shows the Pearson correlation for predictions obtained for sequences containing the number of mutations indicated by the plot title.



**Supplementary Figure 3.14** Line plots showing the R$^2$ values of predicted cleavage activity obtained from random forest (blue) and LSTM with random forest (orange) machine learning models trained on data with incrementally increasing numbers of mutations shown along the x-axis. Each plot shows the R$^2$ for predictions obtained for sequences containing the number of mutations indicated by the plot title.



**Supplementary Figure 3.15. Violin plots showing the distribution of cleavage rates observed in the test data (orange) and the total data set for a given mutation (blue). The distributions are shown separately for each data set containing increasing numbers of mutations, from 3 to 12.**



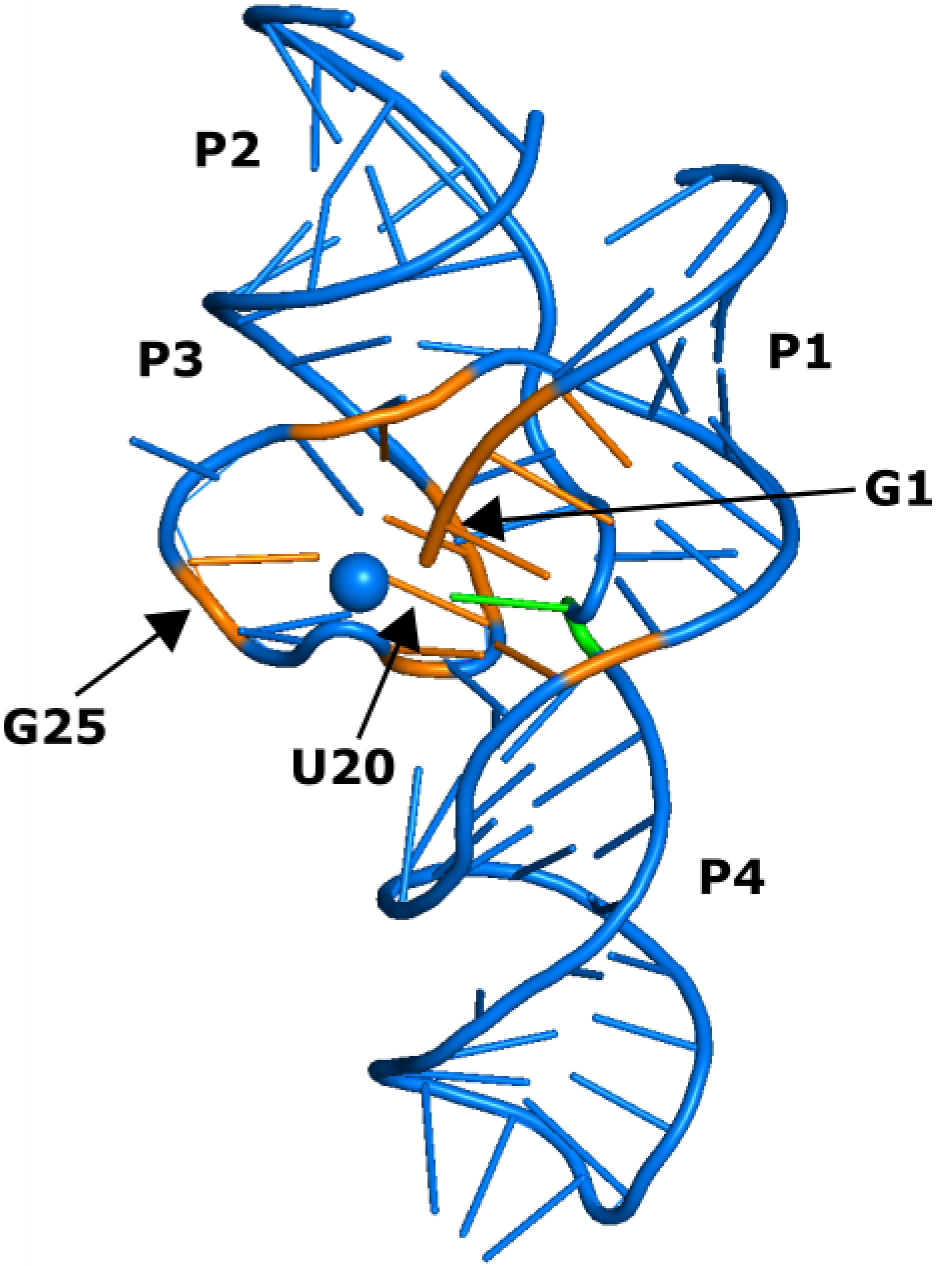
C

Top Ten Feature Importance

| Rank | Train 2 | Train 3 | Train 4 | Train 5 | Train 6 | Train 7 | Train 8 | Train 9 | Train 10 | Train 11 | Train 12 | Train 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | G2:0.044 | G1:0.049 | G1:0.084 | G1:0.145 | G1:0.207 | G1:0.251 | G1:0.278 | G1:0.294 | G1:0.301 | G1:0.304 | G1:0.305 | G1:0.305 |
| 2 | G1:0.041 | G2:0.046 | G25:0.046 | G25:0.064 | G25:0.082 | G25:0.093 | G25:0.096 | G25:0.1 | G19:0.096 | G25:0.101 | G25:0.098 | G25:0.096 |
| 3 | G25:0.034 | G25:0.037 | G2:0.043 | G19:0.054 | G19:0.068 | G19:0.08 | G19:0.089 | G19:0.089 | G25:0.095 | G19:0.09 | G19:0.092 | G19:0.095 |
| 4 | T20:0.034 | G28:0.034 | G19:0.042 | G2:0.039 | G2:0.032 | G2:0.028 | G2:0.027 | G2:0.026 | G2:0.026 | G2:0.026 | G2:0.026 | G2:0.026 |
| 5 | G28:0.033 | T20:0.034 | T20:0.032 | T20:0.027 | C29:0.024 | C29:0.025 | C29:0.024 | C29:0.024 | C29:0.023 | C29:0.021 | C29:0.022 | C29:0.02 |
| 6 | C22:0.033 | G19:0.033 | G28:0.032 | G28:0.027 | T20:0.024 | T20:0.021 | T20:0.019 | T20:0.019 | T20:0.018 | T20:0.018 | T20:0.018 | T20:0.018 |
| 7 | G19:0.032 | C22:0.033 | C22:0.03 | C22:0.026 | G28:0.024 | G28:0.021 | G28:0.019 | G28:0.018 | G28:0.018 | G28:0.018 | G28:0.018 | G28:0.018 |
| 8 | G3:0.032 | G3:0.032 | G3:0.029 | G3:0.025 | C22:0.023 | C22:0.02 | C22:0.018 | C22:0.018 | C22:0.017 | C22:0.017 | C22:0.017 | C22:0.017 |
| 9 | G37:0.03 | G37:0.029 | G37:0.027 | C29:0.024 | G3:0.022 | G3:0.019 | G3:0.018 | G3:0.017 | G3:0.017 | G3:0.017 | G3:0.017 | G3:0.017 |
| 10 | T69:0.029 | T69:0.026 | G17:0.024 | G37:0.024 | G37:0.02 | G37:0.018 | T36:0.017 | T36:0.016 | T36:0.016 | T36:0.016 | T36:0.016 | T36:0.015 |
| Total | 0.342 | 0.353 | 0.387 | 0.454 | 0.527 | 0.576 | 0.606 | 0.621 | 0.628 | 0.627 | 0.629 | 0.627 |

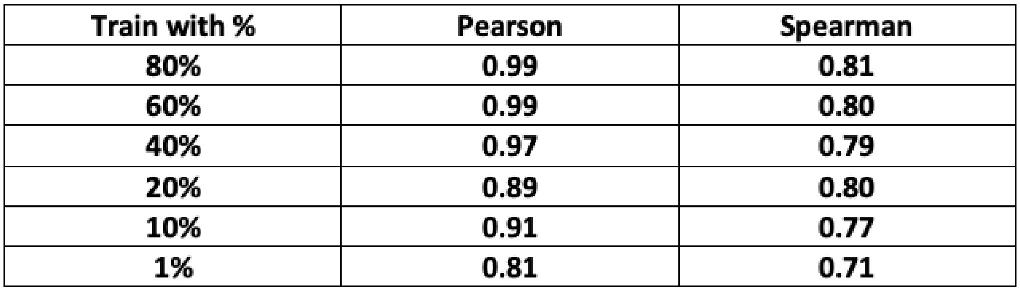
**Supplementary Figure 3.16. Summary of important features extracted from random forest models.**

A-B) Bar graphs of feature importance when training with up to five mutations. Each feature represents a specific nucleotide at a specific location, as indicated by the X-axis label (position), and color (nucleotide identity). Positions 1-35 are shown in (A), and positions 36-69 are shown in (B). The height of the bar indicates the relative importance. C) Table ranking the top ten important features extracted from random forest models trained with increasing numbers of mutations. Nucleotides discussed in the main text are highlighted.



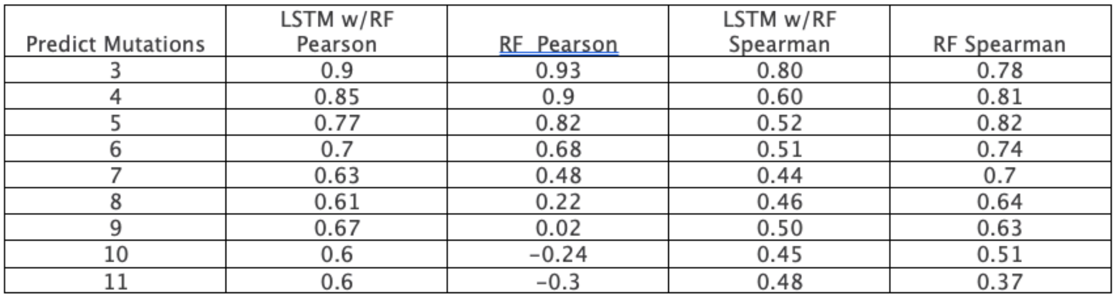
**Supplementary Figure 3.17. Crystal structure of an HDV ribozyme (PDB 3NKB) showing the CPEB3 analogous positions representing the top ten important features identified in our random forest models.**
The feature importance depicted was extracted from the random forest model trained on CPEB3 data including up to 5 mutations. The nucleotides identified as the top ten important features are shaded in orange, the catalytic nucleotide is shaded green (C57/75), and the catalytic $Mg^{2+}$ ion is depicted as a blue sphere.

**Supplementary Table 3.1. Table comparing Pearson and Spearman correlation metrics for reduced training sets containing sequences with up to 5 mutations predicting sequences with 7 mutations. Both Pearson and Spearman correlations show similar, limited reductions in correlation as training set size is reduced.**



| Train with % | Pearson | Spearman |
|---|---|---|
| 80% | 0.99 | 0.81 |
| 60% | 0.99 | 0.80 |
| 40% | 0.97 | 0.79 |
| 20% | 0.89 | 0.80 |
| 10% | 0.91 | 0.77 |
| 1% | 0.81 | 0.71 |

**Supplementary Table 3.2. Table comparing Pearson and Spearman correlation metrics for training set containing sequences with up to 2 mutations predicting sequences with 3 to 11 mutations using the LSTM with Random Forest and the Random Forest models. Both Pearson and Spearman correlations show similar reductions in correlation as predictive distance grows.**



| Predict Mutations | LSTM w/RF Pearson | RF Pearson | LSTM w/RF Spearman | RF Spearman |
|---|---|---|---|---|
| 3 | 0.9 | 0.93 | 0.80 | 0.78 |
| 4 | 0.85 | 0.9 | 0.60 | 0.81 |
| 5 | 0.77 | 0.82 | 0.52 | 0.82 |
| 6 | 0.7 | 0.68 | 0.51 | 0.74 |
| 7 | 0.63 | 0.48 | 0.44 | 0.7 |
| 8 | 0.61 | 0.22 | 0.46 | 0.64 |
| 9 | 0.67 | 0.02 | 0.50 | 0.63 |
| 10 | 0.6 | −0.24 | 0.45 | 0.51 |
| 11 | 0.6 | −0.3 | 0.48 | 0.37 |

# CHAPTER FOUR: CONSTRUCTING MAMMALIAN CELL LINES WITH CHROMOSOMALLY INTEGRATED RNA REPORTERS FOR CO-TRANSCRIPTIONAL ANALYSIS OF SYNTHETIC RIBOSWITCHES

Jessica M. Roberts[1], Matthew L. Ferguson[1,2], Eric J. Hayden[1,3]

[1]Biomolecular Sciences Graduate Programs, Boise State University, Boise, ID, USA

[2]Department of Physics, Boise State University, Boise, ID, USA

[3]Department of Biological Sciences, Boise State University, Boise, ID, USA

## Abstract

Advances in gene editing technologies such as CRISPR-Cas9 now enable the precise modification of genetic material. However, the advancement of synthetic gene regulatory technologies is still needed to realize safe and effective implementation of genetic engineering practices. RNA riboswitches have emerged as promising candidates for such realization. Riboswitches are naturally occurring structured RNA elements found in the untranslated regions of some mRNA transcripts, where they exert metabolite responsive regulation of gene expression. Riboswitches are modular in structure, containing an aptamer domain that specifically binds to a metabolite, and a dynamically structured expression platform domain whose conformation is influenced by the bound or unbound state of the aptamer. Synthetic riboswitches, or aptazymes, have been developed with self-cleaving ribozymes as the expression platform module, and aptamers as the metabolite binding module. Previously developed aptazymes often exhibit undesired

basal gene expression when in the ‘off’ state, and a narrow dynamic range between gene activation or repression. Additionally, efforts to develop synthetic riboswitches in mammalian cells has compounded issues with their performance, and further research and development is needed before widespread implementation in real world applications. However, progress has been relatively slow, in part because the performance of riboswitches is generally inferred via bulk protein read outs, losing important kinetic and mechanistic insight into the self-cleavage activity of the aptazyme at the RNA level. Because mammalian mRNA transcription and processing are physically and kinetically coupled, the co-transcriptional kinetics of synthetic aptazyme riboswitches is undoubtedly of central importance to their performance. Here, we developed a live-cell platform to observe aptazyme cleavage activity at the site of transcription. Several aptazymes and ribozyme controls were imbedded into the 3’-UTR of an established RNA reporter gene fluorescently labelled with phage coat proteins fused to GFP or mCherry. A clonal cell line expressing both GFP and mCherry fusion constructs was established using lentiviral transfection and colony isolation. Each reporter gene was integrated into the same chromosomal location using FRT/flp recombination, confirmed by sequencing, establishing several cells lines that can be directly compared. This platform will enable observation of synthetic riboswitch activity in mammalian cells, providing mechanistic details that will guide future efforts aimed at their improvement.

## Introduction

Riboswitches are naturally occurring protein independent gene regulatory elements found in untranslated regions of some mRNA transcripts. Many distinct classes of these RNA based gene regulatory elements are widely distributed in bacteria, and some classes have also been identified in the genomes of eukaryotic species (McCown et al., 2017). Riboswitches are generally composed of two domains, a metabolite sensing domain referred to as an aptamer, and an expression platform module. Aptamers form complex 3D structures capable of selectively binding and detecting the presence of intracellular metabolites such as coenzymes, nucleotide derivatives, signaling molecules, amino acids, and ions. The binding of the aptamer's ligand induces a structural conformation change that permeates through the expression platform portion of the riboswitch, which alters gene expression (Serganov & Patel, 2012).

Riboswitches modulate gene expression through several mechanisms. The most common mechanisms influence either transcription termination or translation initiation. In the case of transcription termination, binding of the ligand induces the switching between terminator and anti-terminator stems, which effectively terminates or promotes transcription. For translation initiation, ligand induced conformational changes lead to the sequestration or exposure of a ribosomal binding site. Other mechanisms are less common, such as altering splicing or RNA stability (Etzel & Mörl, 2017; Pavlova et al., 2019; Roth & Breaker, 2009; Serganov & Nudler, 2013; Serganov & Patel, 2012).

The several classes of naturally occurring riboswitches in the genomes of organisms have gained attention from synthetic biologists and bioengineers looking to develop tools to synthetically regulate gene expression. Advances in gene editing technologies now enable precise modification and engineering of biological systems with a variety of applications such as personalized medicine and gene therapy, metabolic engineering for the production of biofuels and medicines, and crop improvement (Barrangou & Doudna, 2016; Gupta & Shukla, 2017; Pickar-Oliver & Gersbach, 2019;

Shanmugam et al., 2020). However, in order to safely and effectively engineer desired metabolic pathways or modify genes in an individual organism, robust and orthogonal gene regulatory mechanisms are still needed.

Riboswitches are promising candidates for synthetic gene regulatory platforms for several reasons. First, they are fairly modular in nature, and variation in expression platforms and aptamer domains presents opportunities to develop riboswitches that target a desired ligand, and affect gene regulation at desired process (e.g. transcription initiation or termination vs translation initiation). Importantly, in vitro selection experiments (SELEX) allows for the generation of new aptamer domains that can target virtually any desired ligand, enabling the development of orthogonal and external control over gene regulation (Dixon et al., 2010; Etzel & Mörl, 2017). Second, riboswitches present a protein-free avenue to directly regulate gene expression of the mRNA transcripts in which they reside, and do not depend on the expression of intermediate protein-based transcription factors. Because of this, riboswitches require a much smaller genetic footprint than protein transcription factors, with reduced potential for immunogenicity, and a decreased energetic burden on the cell (Yokobayashi, 2019). Finally, high-throughput in vivo selection strategies allows for the selection and identification of synthetic candidate riboswitches which respond to a desired ligand *and* alter gene expression at the desired step (Desai & Gallivan, 2004; Townshend et al., 2015; Wieland et al., 2012).

Several such efforts to develop synthetic riboswitches have utilized catalytically active RNA molecules that catalyze the cleavage of their own phosphodiester backbone, called self-cleaving ribozymes. Riboswitches developed using self-cleaving ribozymes as an expression platform are referred to as aptazymes, as they are typically generated via the addition of an aptamer domain to a self-cleaving ribozyme (Etzel & Mörl, 2017; Yokobayashi, 2019). Aptazymes are thought to modulate gene expression by reducing the stability and subsequent processing and translation of mRNA transcripts upon self-cleavage. Binding of the ligand by the aptamer domain can stabilize a catalytically active conformation, and the presence of the ligand would increase self-cleavage and reduce gene expression (“off switch”). Or the ligand could stabilize an inactive conformation, and the presence of the ligand would decrease self-cleavage and increase gene expression (“on switch”) (Zhong et al., 2016).

While efforts by several research groups have successfully generated synthetic aptazyme based riboswitches in a variety of model organisms, improvement is still needed to realize widespread adoption in real world applications (Lee & Oh, 2015; Ogawa & Maeda, 2008; Townshend et al., 2015). One issue is that many previously developed aptazyme riboswitches exhibit undesired levels of background expression and apparent ‘leakiness’. Additionally, low activation ratios of only 2-4 fold are common, and a more robust change in gene expression is essential for their effective implementation (Etzel & Mörl, 2017). One of the major challenges in addressing these limitations lies in the fact that current methods for developing synthetic riboswitches is based on protein level detection. This approach only provides an indirect read-out of the synthetic riboswitch performance, and lacks important kinetic information at the RNA level where they function.

Mammalian transcription and mRNA processing is complex, and involves the co-transcriptional recruitment of multiple enzyme complexes that facilitate splicing, capping, polyadenylation, and nuclear export of a nascent RNA transcript (Bentley, 2014; Komili & Silver, 2008). This is largely achieved via interactions with the C-terminal domain (CTD) of RNA polymerase II, where changes in phosphorylation patterns are associated with differential recruitment of appropriate mRNA processing enzymes (Heidemann et al., 2013; Phatnani & Greenleaf, 2006). Additionally, these highly coordinated and regulated processes are greatly influenced by the kinetics of transcription, and variation in alternative splicing and other processing steps such as transcript release have been shown to be influenced by alterations to such kinetics (Chauvier et al., 2017; Dujardin et al., 2013). Therefore, nuanced differences in the kinetics and timing of self-cleavage undoubtedly have a significant effect on the performance of the riboswitch. Therefore, in order to improve the regulatory potential of synthetic riboswitches, it is necessary to understand the kinetic differences that exist between riboswitches with high and low dynamic ranges, and between those that exhibit high levels of leakiness from those with acceptable basal levels of expression.

In order to facilitate the analysis of synthetic riboswitch performance at the RNA level in mammalian cells, we developed several cell lines with chromosomally integrated fluorescent RNA reporters. To do this, we modified a previously developed human β-globin reporter that enables the real time fluorescent visualization of nascent RNA (Coulon et al., 2014; Martin et al., 2013) to contain synthetic riboswitches in the 3' untranslated region (UTR). When the reporter is transcribed into RNA, two distinct types of multimeric phage hairpins form in the first intron and final exon. The first intron contains RNA hairpins from the phage PP7 (Chao et al., 2008), and the final exon contains hairpins from the MS2 phage (Fusco et al., 2003). Once transcribed, the hairpins are subsequently bound by fluorescently labeled  PP7-mCherry and MS2-GFP phage coat proteins, enabling the fluorescent visualization of nascent RNA transcripts. Fluorescent cross correlation analysis (FCCS) of the fluorescent signals emerging from the two color labeling system allows for the simultaneous single molecule measurements of RNA transcription elongation, co-transcriptional splicing, and transcript release kinetics (Coulon et al., 2014).  By modifying this reporter system to contain synthetic aptazyme based riboswitches in the 3'UTR, alterations to the kinetics of these processes can elucidate kinetic insight into self-cleavage activity of the riboswitches.

We chose to integrate variants of hammerhead ribozyme based synthetic riboswitches that were previously selected for in yeast, and were shown to exhibit variation in both dynamic range and basal gene expression levels (Townshend et al., 2015). Two unique variants of theophylline responsive hammerhead riboswitches were chosen for this work and are summarized in table 4.1. These constructs exhibit an increase in gene expression in response to theophylline. The first construct (11X Theo) was previously shown to exhibit an 11-fold increase in gene expression in the presence of theophylline. The second construct (9X Theo), was previously shown to exhibit a 9-fold increase in gene expression in the presence of theophylline, but had a lower basal level of expression in the absence of theophylline, indicating less 'leakiness' in the off state. In addition, several ribozyme constructs that do not exhibit changes in gene expression in response to theophylline were utilized as controls. These are referred to 'graded' ribozymes in this work because they spanned the dynamic range of the theophylline responsive riboswitches described above when placed in the 3'UTR of an mRNA transcript in yeast. These are also summarized in table 4.1.

**Table 4.1. Summary of riboswitch constructs used**

| Riboswitch Construct | Smolke Naming | Expression in yeast |
|---|---|---|
| **11X Theo** | Theo(A)-AAAGA | 11X activation |
| **9X Theo** | Theo(A)-AAAAA | 9X activation |
| **Gated High (GH)** | Grz_TGTT_GTGA | High |
| **Gated Medium (GM)** | Grz_TACA_AGCT | Medium |
| **Gated Low (GL)** | Grz_TGTT_ATAA | Low |
| **Gated Control (GC)** | sTRSVctl | High |

The adapted β-globin reporters each containing a unique riboswitch or graded ribozyme variant were subsequently stably integrated into a single genomic location of 293T-Rex$^{TM}$ cell lines, allowing for the comparison of their activity independent of genomic location. This is important because epigenetic regulation and chromatin organization greatly influences the process of transcription (Felsenfeld et al., 1996; Shilatifard, 2006). Therefore, in order to assure that any observed perturbations to transcriptional kinetics in our system was attributed to the riboswitch construct, stable integration into a single genomic location was desired. Hence, we have developed a platform that will allow for the direct comparison of synthetic riboswitch activity at the site of transcription in live mammalian cells.

## Results

Stable integration of fluorescent coat proteins

In order to minimize any variation in the fluorescent signals due to differential lentiviral integration of the fluorescent phage coat proteins, we first set out to isolate clonal populations of the 293T-Rex$^{TM}$ cells after their transduction. The parent 293T-Rex$^{TM}$ cell line with an empty FRT site was first co-transduced with both MS2-GFP and PP7-mCherry phage coat proteins by lentivirus. This resulted in a mixed population of cells where some contained only MS2-GFP, some contained only PP7-mCherry, and a desired subset contained both of the fluorescently labelled phage coat proteins. Initial efforts to isolate clonal populations that contained both fluorescent phage coat proteins were focused on utilizing fluorescence-activated cell sorting (FACS). However, multiple attempts failed to generate viable clonal populations. We determined that this cell line appeared to be sensitive to the cell sorting process, and also likely grew poorly in the absence of other cells. We then attempted to isolate single cells via the gentler process of serial dilution, and similarly found that growth was inhibited when only single cells were seeded onto a 96 well plate. After multiple attempts, we reasoned that we may be more successful at expanding clonal populations from isolated colonies rather than from single cells. To do this, we implemented colony picking guided by fluorescence microscopy. Cells were thinly seeded onto a 10 cm plate and allowed to grow into isolated colonies. Colonies expressing both phage coat proteins were carefully picked up from the plate using a pipette, and transferred to individual wells of a 96 well plate and allowed to expand. This approach was successful, and we were able to isolate a colony of empty 293T-Rex$^{TM}$ cells expressing desired levels of both phage coat proteins to be used for integration of our β-globin reporter constructs (Figure 4.1 A). In addition to being utilized to stably integrate the β-globin construct containing the select riboswitch constructs in this study, this cell line containing both phage coat protein and an empty FRT site is poised to easily allow for the integration of any future design iterations.

### Gibson assembly of aptazyme reporter plasmids

We next set out to modify a developed human β-globin reporter system to contain the synthetic riboswitch constructs. Starting with the Flp-In$^{TM}$ system pcDNA$^{TM}$5/FRT expression vector plasmid containing the β-globin reporter with the PP7 hairpins located in intron 1, Gibson assembly was used to integrate the riboswitch constructs into the 3'UTR, just before the poly A signal. Using primers that target the plasmid and flank the integration site, PCR was used to screen for plasmids that successfully integrated a riboswitch construct. Positively screened plasmids were sent for Sanger sequencing to further confirm that the riboswitch constructs were successfully integrated into the desired location in the 3'UTR of the β-globin reporter.

After confirming the integration of our synthetic riboswitch constructs into the pcDNA$^{TM}$5/FRT expression vector containing the β-globin reporter, we then stably integrated the reporter constructs into the 293T-Rex$^{TM}$ cells containing the phage coat proteins via site directed FLP-FRT recombination. Co-transfection of the cells with the pOG44 Flp recombinase expression vector and the pcDNA$^{TM}$5/FRT expression vector containing our reporter results in FRT site based integration of our reporter into a single genomic site. The pcDNA$^{TM}$5/FRT expression vector contains a hygromycin resistance gene that lacks a promoter and ATG initiation codon, and is therefore not expressed directly from the plasmid. Successful genomic integration into the FRT site in the 293T-

Rex$^{TM}$ cells brings an SV40 promotor and ATG initiation into proximity of the hygromycin resistance gene, enabling its expression. This allows for the selection of cells that successfully integrated our reporter system into their FRT site via treatment with Hygromycin B. The surviving cells were expanded, and successful integration of our reporter constructs was further validated by genomic DNA isolation and subsequent PCR screening using primers that amplify our β-globin reporter. The expanded and verified cell lines contain a stable integration of the β-globin reporter harboring unique variants of theophylline responsive hammerhead based synthetic riboswitches all at the same genomic location.

### Fluorescence imaging based confirmation of reporter integrations

In order to verify that that resulting cell lines enabled fluorescent detection of transcription sites actively expressing our β-globin constructs, we implemented high-resolution fluorescent microscopy. As an initial confirmation, we obtained images of the gated control cell line using the apotome 3 feature of the Zeiss axio observer microscope in both the GFP and mCherry channels. The images were overlayed and showed that the cells still expressed both MS2-GFP and PP7-mCherry fluorescent phage coat proteins. In addition, many cells showed sites of active transcription that were observed as singular sub-nuclear foci of high intensity fluorescent signal in both channels (Figure 4.1 B-C). From this we determined that these cells have stable integration of both the reporter construct and fluorescent phage coat proteins, and could be used to measure the kinetics of several co-transcriptional mRNA processing events.

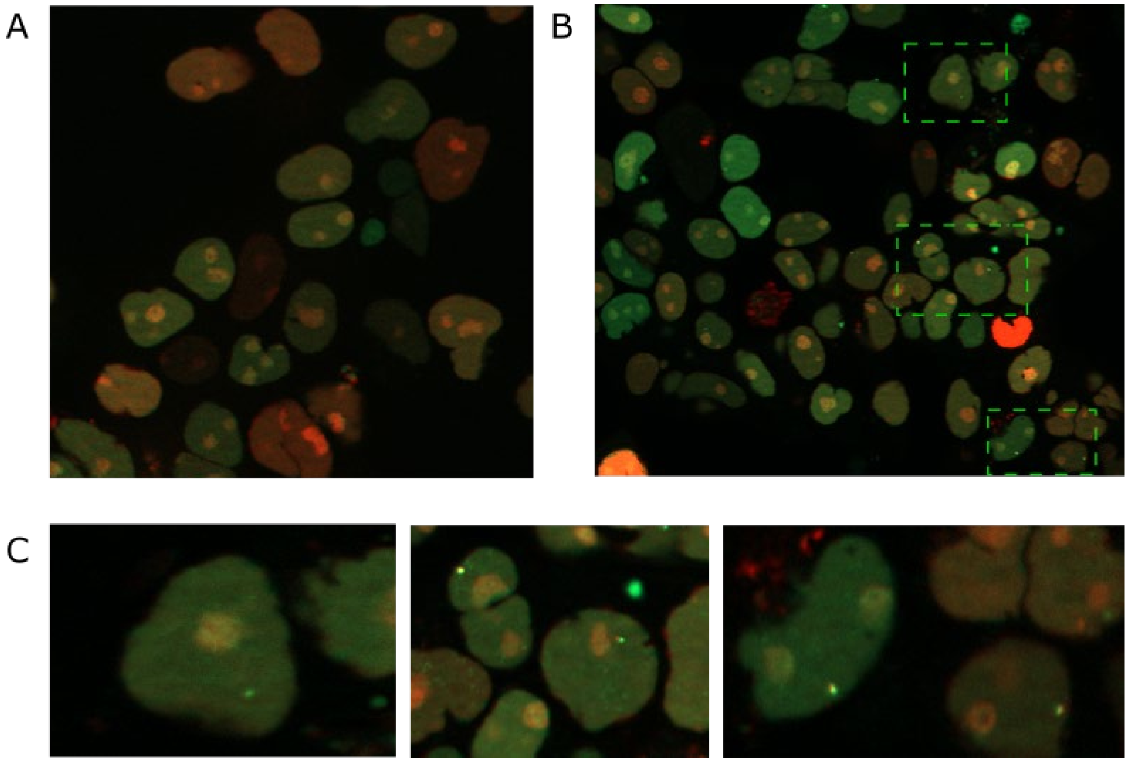


**Figure 4.1. Fluorescent microscopy images of select resulting cell lines**
A) Overlaying fluorescent image of 293T-RexTM cells containing an empty FRT site and stably integrated MS2-GFP and PP7-mCherry fluorescent phage coat proteins. B) Overlaying fluorescent image of the 293T-RexTM cells with stable integration of both phage proteins, and containing stable integration of the gated control β-globin reporter in the genomic FRT site. Select cells with active sites of transcription are boxed off and expanded in (C).

## Discussion

The resulting cell lines provide a ready platform for the real time imaging and analysis of synthetic riboswitch activity in mammalian cells. The theophylline responsive hammerhead based synthetic riboswitch constructs chosen exhibit variation in both their dynamic ranges of control over gene expression, and their basal expression levels. This should allow for the direct comparison of the kinetics of their self-cleavage activity at the RNA level, providing insight into differences that may contribute to both leakiness in basal level expression as well as in achieving greater dynamic ranges in gene expression modulation. The graded ribozyme controls that do not exhibit changes in gene expression in response to theophylline provide an internal metric to assess co-transcriptional self-cleavage activity of our constructs in the context of their new cellular environment (human vs yeast). In addition to the cell lines that contain synthetic riboswitch constructs, the clonal 293T-Rex$^{TM}$ cell line containing both PP7-mCherry and MS2-GFP fluorescent phage coat proteins is poised for the integration of future iterations to allow for the continued exploration and improvement of synthetic RNA regulatory elements.

Taken together, this work provides an avenue to directly measure synthetic aptazyme based riboswitch self-cleavage kinetics directly at the site of transcription. This will provide a window into the single molecule self-cleavage activity of these RNA regulatory elements at the RNA level, effectively removing the black box associated with bulk protein read-outs. We expect that this approach will be a valuable avenue to gain guiding insight towards the improvement and successful implementation of synthetic riboswitches as widespread, effective, and robust RNA gene regulatory elements with broad applications in synthetic biology and engineering.

## Materials and Methods

### Cell line growth and maintenance

293T-Rex$^{TM}$ cells (Invitrogen$^{TM}$) were obtained through ThermoFisher Scientific and maintained according to the manufacturers recommended conditions. Briefly, cells were grown in DMEM (Corning, 4.5g/L glucose, with L-glutamine and sodium pyruvate), supplemented with 10% fetal bovine serum, and 1X penicillin and streptomycin (Manufacturer info). The cells were maintained at 37°C with 5% $CO_2$ and 95% humidity.

Plasmids, DNA oligos, and cloning

The plasmid containing the Tet inducible dual color β-globin reporter was obtained from collaborators, and is available through addgene (Plasmid #61762). The β-globin reporter was modified to contain the PP7 hairpins in the first intron, and then cloned into the pcDNA™FRT/TO vector designed for use with the Flp-In$^{TM}$ T-Rex$^{TM}$ system using restriction enzyme based cloning methods by previous members of our group.

The synthetic hammerhead based riboswitch sequences were previously published (Townshend et al., 2015), and were ordered from IDT as Gblocks® Gene Fragments. Areas of homology targeting the 3'UTR of the β-globin reporter construct were appended to the riboswitch sequences via PCR using primers that bind to the riboswitch sequence and contain additional nucleotides homologous to the β-globin reporter. The resulting riboswitch sequences were cloned into the 3'UTR of the β-globin reporter contained in the pcDNA™FRT/TO vector using the NEBuilder® HiFi DNA Assembly cloning kit (E5520S, New England Biolabs®). The assembly reaction was transformed into NEB® 5-alpha Competent *E. coli, the resulting colonies were screened for riboswitch insertion via PCR, and the PCR products were sequence validated using Sanger sequencing (Elim Biopharmaceuticals Inc.).*

Lentiviral vectors containing the MS2-GFP (addgene plasmid ID # 61764) and PP7-mCherry (addgene plasmid ID # 61763) phage coat proteins under the control of a human ubiquitin promotor were previously developed (Coulon et al., 2014; Larson et al., 2013), and were used in conjunction with the ViraPower$^{TM}$ lentiviral packaging mix (Life Technologies).

Stable Integration of fluorescent reporters

Lentiviral particles containing the MS2-GFP and PP7-mCherry were separately prepared via co-transfection of HEK293T cells with the ViraPower$^{TM}$ lentiviral packaging plasmids and the lentiviral vectors containing the respective fluorescently labeled phage coat proteins. Briefly, PEI/DNA complexes were formed at a 5:1 ratio in serum free DMEM via brief vortex and room temperature incubation, and then added dropwise to a tissue culture plate containing HEK293T cells at ~50% confluency. The media was replaced at 24 hours, and then collected at 72 hours to harvest the viral particles. The media containing the viral particles was filtered through a 100 μM cell strainer (VWR® Cell Strainers, Catalog #10199-658) to remove cellular debris, and then vacuum filtered through a 0.45 μM filter (VWR® Tube Top Vacuum Filters, Catalog #76012-772). The resulting mediums containing the purified viral particles were combined (MS2-PP7, and PP7-mCherry), and then immediately transferred to a 10 cm tissue culture plate containing 293T-Rex$^{TM}$ cells at low confluency for transduction.

Clonal cell isolation

293T-Rex$^{TM}$ cells with an empty FRT site that were transduced with both MS2-GFP and PP7-mCherry phage coat proteins were expanded on a 10 cm plate until ~80% confluent. Cells were trypsinized, pelleted, and resuspended in 3 mL of complete DMEM. Cells were thinly seeded at a 1:300 dilution onto a new 10 cm plate containing 10 mL of complete DMEM media. Cells were incubated at 37°C, 5% $CO_2$, 95% humidity until small and distinct colonies formed. Using an EVOS fluorescent microscope, colonies expressing roughly equal levels of both GFP (EVOS GFP light cube, Ex 482/25 nm, Em 524/24 nm) and mCherry (EVOS Invitrogen Texas Red light cube, Ex 585/29, Em 628/32) were identified. The isolated colonies were carefully picked up off of the plate with a sterile pipette tip, and transferred to a single well of a 96 well plate containing 200 μL of complete DMEM and allowed to expand. Resulting colonies were qualitatively assessed for normal morphology, growth rates, and fluorescence, and a single colony was chosen for expansion and future integration of β-globin reporter containing the riboswitch constructs.

### Single genomic integration of riboswitch reporter

Each riboswitch construct was stably integrated into a single genomic location in the clonal population of 293T-Rex$^{TM}$ cells containing stably integrated PP7-mCherry and MS2-GFP coat proteins via Flp-In site specific recombination. Briefly, cells were co-transfected with the pcDNA™FRT/TO vector containing the β-globin riboswitch reporter, and the pOG44 Flp-In recombinase expression vector using PEI transfection reagent as described above. After a 24-hour incubation, the media on the cells was replaced with selective medium (150 μg/mL Hygromycin B, DMEM, 10% FBS, 1X Pen/Strep). The cells were expanded under selective media, and stable integration was confirmed via genomic DNA purification and PCR screening for insertion of the β-globin reporter(s).

### Fluorescence imaging and transcription spot visualization

Fluorescence images were obtained using a Zeiss Axio Observer microscope and the resolution was enhanced using the Apotome 3 feature. Images were consecutively obtained in both the GFP and mCherry channels using auto-expose settings and were subsequently overlayed.

## References

Barrangou, R., & Doudna, J. A. (2016). Applications of CRISPR technologies in research and beyond. *Nature Biotechnology*, *34*(9), 933–941. https://doi.org/10.1038/nbt.3659

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics*, *15*(3), 163–175. https://doi.org/10.1038/nrg3662

Chao, J. A., Patskovsky, Y., Almo, S. C., & Singer, R. H. (2008). Structural basis for the coevolution of a viral RNA–protein complex. *Nature Structural & Molecular Biology*, *15*(1), 103–105. https://doi.org/10.1038/nsmb1327

Chauvier, A., Picard-Jean, F., Berger-Dancause, J.-C., Bastet, L., Naghdi, M. R., Dubé, A., Turcotte, P., Perreault, J., & Lafontaine, D. A. (2017). Transcriptional pausing at the translation start site operates as a critical checkpoint for riboswitch regulation. *Nature Communications*, *8*(1), 13892. https://doi.org/10.1038/ncomms13892

Coulon, A., Ferguson, M. L., de Turris, V., Palangat, M., Chow, C. C., & Larson, D. R. (2014). Kinetic competition during the transcription cycle results in stochastic RNA processing. *ELife*, *3*, 1–22. https://doi.org/10.7554/eLife.03939

Desai, S. K., & Gallivan, J. P. (2004). Genetic Screens and Selections for Small Molecules Based on a Synthetic Riboswitch That Activates Protein Translation. *Journal of the American Chemical Society*, *126*(41), 13247–13254. https://doi.org/10.1021/ja048634j

Dixon, N., Duncan, J. N., Geerlings, T., Dunstan, M. S., McCarthy, J. E. G., Leys, D., & Micklefield, J. (2010). Reengineering orthogonally selective riboswitches. *Proceedings of the National Academy of Sciences*, *107*(7), 2830–2835. https://doi.org/10.1073/pnas.0911209107

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L. I., Fiszbein, A., Godoy Herz, M. A., Nieto Moreno, N., Muñoz, M. J., Alló, M., Schor, I. E., & Kornblihtt, A. R. (2013). Transcriptional elongation and alternative splicing. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1829*(1), 134–140. https://doi.org/10.1016/j.bbagrm.2012.08.005

Etzel, M., & Mörl, M. (2017). Synthetic Riboswitches: From Plug and Pray toward Plug and Play. *Biochemistry*, *56*(9), 1181–1198. https://doi.org/10.1021/acs.biochem.6b01218

Felsenfeld, G., BoYES, J., Chung, J., CLARKt, D., & Studitsky, V. (1996). Chromatin structure and gene expression. *Proc. Natl. Acad. Sci. USA*, 5.

Fusco, D., Accornero, N., Lavoie, B., Shenoy, S. M., Blanchard, J.-M., Singer, R. H., & Bertrand, E. (2003). Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells. *Current Biology*, *13*(2), 161–167. https://doi.org/10.1016/S0960-9822(02)01436-7

Gupta, S. K., & Shukla, P. (2017). Gene editing for cell engineering: Trends and applications. *Critical Reviews in Biotechnology*, *37*(5), 672–684. https://doi.org/10.1080/07388551.2016.1214557

Heidemann, M., Hintermair, C., Voß, K., & Eick, D. (2013). Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1829*(1), 55–62. https://doi.org/10.1016/j.bbagrm.2012.08.013

Komili, S., & Silver, P. A. (2008). Coupling and coordination in gene expression processes: A systems biology view. *Nature Reviews Genetics*, *9*(1), 38–48. https://doi.org/10.1038/nrg2223

Larson, D. R., Fritzsch, C., Sun, L., Meng, X., Lawrence, D. S., & Singer, R. H. (2013). Direct observation of frequency modulated transcription in single cells using light activation. *ELife*, *2*, e00750. https://doi.org/10.7554/eLife.00750

Lee, S.-W., & Oh, M.-K. (2015). A synthetic suicide riboswitch for the high-throughput screening of metabolite production in Saccharomyces cerevisiae. *Metabolic Engineering*, *28*, 143–150. https://doi.org/10.1016/j.ymben.2015.01.004

Martin, R. M., Rino, J., Carvalho, C., Kirchhausen, T., & Carmo-Fonseca, M. (2013). Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity. *Cell Reports*, *4*(6), 1144–1155. https://doi.org/10.1016/j.celrep.2013.08.013

McCown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E., & Breaker, R. R. (2017). Riboswitch diversity and distribution. *RNA*, *23*(7), 995–1011. https://doi.org/10.1261/rna.061234.117

Ogawa, A., & Maeda, M. (2008). An Artificial Aptazyme-Based Riboswitch and its Cascading System in E. coli. *ChemBioChem*, *9*(2), 206–209. https://doi.org/10.1002/cbic.200700478

Pavlova, N., Kaloudas, D., & Penchovsky, R. (2019). Riboswitch distribution, structure, and function in bacteria. *Gene*, *708*, 38–48. https://doi.org/10.1016/j.gene.2019.05.036

Phatnani, H. P., & Greenleaf, A. L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. *Genes & Development*, *20*(21), 2922–2936. https://doi.org/10.1101/gad.1477006

Pickar-Oliver, A., & Gersbach, C. A. (2019). The next generation of CRISPR–Cas technologies and applications. *Nature Reviews. Molecular Cell Biology*, *20*(8), 490–507. https://doi.org/10.1038/s41580-019-0131-5

Roth, A., & Breaker, R. R. (2009). The Structural and Functional Diversity of Metabolite-Binding Riboswitches. *Annual Review of Biochemistry*, *78*(1), 305–334. https://doi.org/10.1146/annurev.biochem.78.070507.135656

Serganov, A., & Nudler, E. (2013). A decade of riboswitches. *Cell*, *152*(1–2), 17–24. https://doi.org/10.1016/j.cell.2012.12.024

Serganov, A., & Patel, D. J. (2012). Metabolite Recognition Principles and Molecular Mechanisms Underlying Riboswitch Function. *Annual Review of Biophysics*, *41*(1), 343–370. https://doi.org/10.1146/annurev-biophys-101211-113224

Shanmugam, S., Ngo, H.-H., & Wu, Y.-R. (2020). Advanced CRISPR/Cas-based genome editing tools for microbial biofuels production: A review. *Renewable Energy*, *149*, 1107–1119. https://doi.org/10.1016/j.renene.2019.10.107

Shilatifard, A. (2006). Chromatin Modifications by Methylation and Ubiquitination: Implications in the Regulation of Gene Expression. *Annual Review of Biochemistry*, *75*(1), 243–269. https://doi.org/10.1146/annurev.biochem.75.103004.142422

Townshend, B., Kennedy, A. B., Xiang, J. S., & Smolke, C. D. (2015). High-throughput cellular RNA device engineering. *Nature Methods*, *12*(10), 989–994. https://doi.org/10.1038/nmeth.3486

Wieland, M., Ausländer, D., & Fussenegger, M. (2012). Engineering of ribozyme-based riboswitches for mammalian cells. *Methods*, *56*(3), 351–357. https://doi.org/10.1016/j.ymeth.2012.01.005

Yokobayashi, Y. (2019). Aptamer-based and aptazyme-based riboswitches in mammalian cells. *Current Opinion in Chemical Biology*, *52*, 72–78. https://doi.org/10.1016/j.cbpa.2019.05.018

Zhong, G., Wang, H., Bailey, C. C., Gao, G., & Farzan, M. (2016). Rational design of aptazyme riboswitches for efficient control of gene expression in mammalian cells. *ELife*, *5*, e18858. https://doi.org/10.7554/eLife.18858