The Pennsylvania State University

The Graduate School

PREDICTING GENE EXPRESSION FOR ROBUST GENETIC SYSTEM DESIGN

A Dissertation in

Chemical Engineering

by

Daniel P. Cetnar

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2021

The dissertation of Daniel P. Cetnar was reviewed and approved by the following:

Howard Salis
Associate Professor of Chemical Engineering
& Agricultural and Biological Engineering
Dissertation Advisor
Chair of Committee

Andrew Zydney
Bayard D. Kunkle Chair and
Professor of Chemical Engineering

Philip Savage
Walter L. Robb Family Endowed Chair
Department Head, Chemical Engineering

Philip Bevilacqua
Distinguished Professor of Chemistry and
of Biochemistry and Molecular Biology
Department Head, Chemistry

Seong Kim
Distinguished Professor
Associate Department Head, Chemical Engineering

ABSTRACT

This body of work aims to further the chief goal of synthetic biology, which fundamentally consists of the ability to design, build, and test new biological functions. To achieve this goal, we require predictive control over the fundamentals of genetic systems. In particular, we need clear and predictive models to simulate and design entire genetic systems. A key aspect of this design ability is to have control over each component of the biological system that impacts gene expression. Here I describe my work to better understand and model the fundamental processes that determine gene expression. In chapter 2, I use rational design to build small libraries to test and quantify the core sequence and structure determinants of mRNA decay in *E. coli*. In chapter 3, I expand on the idea of mRNA decay by using Next Generation Sequencing (NGS) to test a large library of constructs and build a comprehensive model of mRNA decay using machine learning. In chapter 4, I use a similar NGS experimental approach to build a library of promoters for use in a variety of organisms. In chapter 5, I discuss the main findings and describe avenues for future research. Overall, this work describes methods of determining and modeling the fundamental processes of gene expression to aid in the building of new genetic systems and explaining the performance of existing systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This completion of this work could not have been accomplished without the incredible support I received over my time in the graduate program. First, and foremost, I would like to thank my advisor Dr. Howard Salis for the opportunity, support, and mentorship for my graduate studies. I would also like to thank my committee members Dr. Andrew Zydney, Dr. Philip Savage, and Dr. Philip Bevilacqua for their guidance and help. I also greatly appreciate my lab mates Alex Reis, Ayaan Hossain, Grace Vezeau, Amin Espah Borujeni, Travis La Fleur, and Sean Halper for their help and contributions to many of the projects I worked on in my time in the Salis Lab. I would also like to thank the undergraduate students, Kavya Vaseekaran, Danielle Hand, Lexie Smathers, Zekun Li, Humood Alanzi, Morgan Roggenbaum, Ammar Norain, and Kaylee Ermine, that I had the privilege of working with on numerous projects. I would also like to thank the Chemical Engineering department head Dr. Philip Savage and the associate department head Dr. Seong Kim for their help throughout the graduate program and working with them in building the Graduate Student Association. I would also like to thank the incredible help of Cathy Krause and Peggy Newel for taking care of all the logistics of ordering, scheduling travel, and much more. My wife Taylor and my family have additionally been a huge support throughout my time in the graduate program. I would also like to thank Craig Praul and Ashley Price at the Genomics Core Facility for their training and help with sequencing, qPCR, and NGS. I would also like to thank Carol Baker for sharing her experimental expertise. Furthermore, I would like to thank Mark Signs and Ashik Sathish in the CLS Behring Fermentation Facility for their training and help.

This project was supported by funds from the Air Force Office of Scientific Research (grant no. FA9550-14-1-0089), the Defense Advanced Research Projects Agency (grant no.

FA8750-17-C-0254) and the Department of Energy (grant no. DE-SC0019090). The findings and conclusions of this dissertation do not necessarily reflect the views of these funding agencies.

Chapter 1

Synthetic Biology Approach to Design and Modeling

Living organisms possess an unparalleled ability to synthesize complex molecules and structures, sense their own inner and surrounding environment, and maintain tight control over the execution of vast regulatory networks. Many of these biological processes have clear application in industrial, therapeutic, or agricultural purposes. Biomolecules have a nearly unlimited set of applications in the chemical industry in areas such as fuels[1], flavors and fragrances[2], and pesticides[3]. The growth of biologic drugs in the pharmaceutical industry further demonstrates the broad range of application of bioderived pharmaceutical for the treatment of cancer[4], immune disorders[5], and neurodegenerative disease[6]. Nearly 50% of all cancer drugs[7] and nearly 70% of all anti-infective drugs[8] come from natural products from living organisms. Furthermore, the ability of organisms to sense and respond to environmental stimulus makes them ideal candidates for detecting and remediating environmental toxins and pathogens, which has clear applications for food and transportation security[9]. There currently exists a huge wealth of sequenced genomes full of gene clusters that synthesize and modify an astronomical variety of chemicals. This redefines the challenge going forward from discovering desirable products in nature to instead creating tools to engineer living organisms to produce these products at an economically viable level. Many discovered biochemical products have useful applications, but cannot be implemented due to the high cost and difficulty of engineering biology. Current methods of genetic engineering remain particularly cost and labor intensive, preventing the development of many natural products industrially. Synthetic biology looks to expedite the development of industrial applications of genetic engineering. To do this, it aims to create clear and predictive models of the central dogma of biology, namely how transcription of DNA into

RNA and translation of RNA into protein occurs and how long the various biological molecules persist.

Historical Approach to Genetic Engineering

Historically, the medical and chemical industries looked to nature as a source of inspiration and discovery of novel therapeutics and chemicals without having a clear understanding of the underlying processes. Despite the fact the recombinant DNA technology revolution occurred in the early 1970s[10-11], empirical, black box methods of engineering living systems persisted for years. This led to slow and costly commercialization of complex molecules and pathways. In bacteria, single and multi-cistronic operons form the central architectural unit of all natural and engineered genetic systems. An operon's sequence ultimately determines the expression levels of its RNAs and proteins. Without understanding the fundamental operon design, each product commercialization gains little benefit from previous work. Historically, and in many contemporary industrial settings, crude methods of gene alteration combined with mass screening of variants remains the tool of choice in optimizing production in living organisms. For example, many of the industrial strains in the antibiotic producing bacterial genus, actinomycetes, only developed high producing strains by employing many rounds of random mutagenesis, recombination, or transposon insertion followed by screening. Common approaches include duplicating the genes of interest in the native organism, overexpressing upregulating regulatory genes, and reducing the expression of downregulating regulatory genes[12]. This, however, often fails because transcription factors and other regulatory elements of the gene cluster may remain unknown or gross changes to the regulatory genes could have unintended deleterious impacts on other genes, rendering the cell inviable.

The growth of next generation sequencing and sequence homology analysis led to the discovery of many genes and gene clusters predicted to have useful properties. Unfortunately, many of these gene clusters come from difficult to culture organisms or very low productivity strains, which prevents confirmation and exploitation of their properties. A new approach reconstructs large genetic systems outside the native host operons by appending together several pre-existing promoters, ribosome binding sites, coding sequences, and terminators. However, while avoiding the complexities of using a non-model organism, new problems surface. Due to differences in codon use, the coding regions of the genes of interest frequently require optimization for the new host organism. This, along with the use of stock parts, often results in engineered operons full of overlapping, context-dependent, and undesired genetic elements that confound rational design and inevitably break the operon's function.

Genetic Design Space

An additional fundamental challenge of designing genetic systems remains the astronomically large design space. A stretch of just 20 nucleotides has over a trillion possible design combinations. This makes it infeasible to sample and test every possible design. This necessitates models and groupings of designs to make sense of the huge number of possible combinations. To illustrate the multiple actors controlling gene expression, the following set of differential equations governing transcription, translation, and steady-state protein level in the cell illustrate these ideas respectively.

$$\frac{d[mRNA]}{dt} = \alpha N_{DNA} - (\delta_{mRNA} + \mu)[mRNA]$$

$$\frac{d[Protein]}{dt} = \beta[mRNA] - (\delta_{Protein} + \mu)[Protein]$$

$$[Protein]_{ss} = \frac{\alpha\beta N_{DNA}}{(\delta_{mRNA} + \mu)(\delta_{Protein} + \mu)}$$

Figure **1-1**: Equations governing gene expression.

In these equations, α represents the transcription initiation rate, β represents the translation initiation rate, δ_{mRNA} represents the mRNA degradation rate, $\delta_{Protein}$ represents the protein degradation rate, and μ represents the cell growth rate. From the steady state protein calculation, we have four major parameters, α, β, $\delta_{mRNA,}$ and $\delta_{Protein,}$ to engineer overall protein production. As discussed earlier, significant work and progress has greatly improved biophysical models of the production of mRNA and protein [13-14]. In particular, considerable work has gone into understanding the influence of codon usage in the speed of translational elongation and ultimately its impact on protein folding [15]. Furthermore, a model translational coupling describes the influence of downstream genes in a single operon on each other's translation initiation rate[16]. Work has also focused on transcriptional modeling, but with a greater focus on creating vast libraries of promoters [17]. However, the factors that control the degradation rates of mRNA has received less attention. Previous work on mRNA degradation has established general principles of degradation, but has not produced a comprehensive predictive model of mRNA half-life [18].

Learn-by-Design Approach

To overcome this challenge, synthetic biologists have used rational design to design synthetic DNA constructs to test a particular hypothesis about a particular genetic function of interest. Instead of testing thousands of random sequences for biological function, this approach

creates focused libraries of rationally designed sequences that systematically vary one aspect of gene expression. Likewise, we do not model existing natural systems to avoid context dependent variables from clouding the reliability of the model. This isolates the particular function of interest and allows for robust modeling. This approach has applicability to evaluate and model the many sub-aspects of gene expression. Overall, these models, if used in conjunction, can design larger and more complex systems.

Goals of this Work

We propose using a learn-by-design approach to build models to predict many of the core aspects of gene expression. In chapter 2, I describe my work determining the core sequence and structural determinants of mRNA decay in *E. coli*. In this set, small individual libraries systematically evaluate the impact of translation initiation rate, the type and location of different RNAse sites in a bacterial operon, and the secondary structure considerations when designing an operon. Furthermore, we create individual models of these sub-components of mRNA decay to explain their function and relative importance in determining the overall decay rate of an mRNA. In chapter 3, we expand our study of mRNA decay by designing 62,120 sequences to systematically test the determinants of mRNA decay. In this study, we generate time course data to directly quantify the decay rate of each mRNA construct. Employing the power of Next Generation Sequencing (NGS), we carry out these experiments simultaneously to generate decay rates for the entire sequence library. This data allows us to evaluate particular hypotheses of the features of decay in *E. coli*. Additionally, we calculate the particular features predicted to cause changes in the decay rate. These features help train a series of machine learning models to predict the mRNA decay rate. Overall, these models require only a designed DNA sequence for the model to predict the mRNA lifetime in *E. coli*.

In chapter 4, I use a similar NGS approach to build non-repetitive promoter libraries in *E. coli* and *Clostridium autoethanogenum*. In this work, the core aspects of a promoter avoid too great of similarity so as to avoid homologous recombination or other negative effects that can cause genetic systems to fail. In the case of *C. autoethanogenum*, few promoters previously existed for use in designing genetic systems. The library design had two goals (1) non-repetitive sequence and (2) cover a broad range of transcription initiation rates. This allows genetic system tuning for the correct level of expression.

Chapter 2

Systematic Quantification of Sequence and Structural Determinants Controlling mRNA stability in Bacterial Operons

mRNA degradation is a central process that affects all gene expression levels, and yet the determinants that control mRNA decay rates remain poorly characterized. Here, we applied a synthetic biology, learn-by-design approach to elucidate the sequence and structural determinants that control mRNA stability in bacterial operons. We designed, constructed, and characterized 82 operons in *E. coli*, systematically varying RNase binding site characteristics, translation initiation rates, and transcriptional terminator efficiencies in the 5' UTR, intergenic, and 3' UTR regions, followed by measuring their mRNA levels using RT-qPCR assays during exponential growth. We show that introducing long single-stranded RNA into 5' UTRs reduced mRNA levels by up to 9.4-fold and that lowering translation rates reduced mRNA levels by up to 11.8-fold. We also found that RNase binding sites in intergenic regions had much lower effects on mRNA levels. Surprisingly, changing transcriptional termination efficiency or introducing long single-stranded RNA into 3' UTRs had no effect on upstream mRNA levels. From these measurements, we developed and validated biophysical models of ribosome protection and RNase activity with excellent quantitative agreement. We also formulated design rules to rationally control a mRNA's stability, facilitating the automated design of engineered genetic systems with desired functionalities.

Fundamental mRNA Decay Processes

Engineering sophisticated genetic systems requires the development of more comprehensive biophysical models that can predict how changes to sequence affect gene expression levels, taking into account transcription, translation, and mRNA degradation [19-22]. mRNA decay rates are an important contributor to genetic circuit function, altering the circuit's dynamic and steady-state gene expression levels as well as controlling "turn off" times in response to changing transcriptional programs [23-24]. While significant focus has been given to measuring and modeling transcriptional initiation rates [13, 25-28], transcriptional control [29-30], and translation initiation rates [31-33], previous work on mRNA stability has primarily focused on discovering and characterizing the proteins and pathways responsible for mRNA decay [34-38]. Known RNase sites have also been inserted into operons to modulate expression[39].

A key challenge to modeling and predicting mRNA decay rates arises from the many participating enzymes and the high degree of coupling between transcription, translation, and RNase activity [40-41]. Predicting mRNA decay rates in multi-cistronic operons is further complicated by the potential for differential decay rates across coding sequences, called polarity [42-43]. Here, we apply a learn-by-design approach to quantitatively measure how a bacterial mRNA's sequence and structural properties determine its stability through the several mechanisms of mRNA decay. From these results, we formulate quantitative design rules for controlling a mRNA's stability.

Several enzymes work together in a multi-pathway process to determine how quickly bacterial mRNAs are degraded [37-38], including endonucleases (*E. coli* RNases E, G, III), exonucleases (*E. coli* RNase II, PNPase, RNase R), and the helper enzymes that modify the ends of mRNAs (*E. coli* RppH, DapF, PAPI) (**Figure 2-1**). These pathways are intimately coupled to both transcription and translation. Beginning with transcription initiation, the 5' ends of all

bacterial mRNAs contain a triphosphate group that inhibits RNase activity. However, over time, RppH hydrolyzes these 5' ends, yielding a monophosphate group that has higher affinity to RNase E [44-45]. RppH activity is promoted by the protein DapF and depends on the sequence and structure of at least the first three nucleotides of the transcript (**Figure 2-1A**) [46-48]. As transcription elongation proceeds, co-transcriptional folding determines the types of mRNA structures that form and the portions that remain unstructured and accessible, which is also affected by RNA polymerase's transcriptional elongation rate [49-50]. Furthermore, additional proteins (e.g. CsrA and Hfq) may bind nascent mRNA transcripts and can impact site accessibility and structure[51-52]. During this time, as start codons are transcribed, ribosomes will bind and begin to translate the mRNA, following RNA polymerase and covering the mRNA with protective ribosomes. Once the mRNA is fully transcribed, the translation initiation and elongation rates of its coding sequences will determine the number of protective ribosomes bound to the transcript (**Figure 2-1B**).

As transcription and translation proceeds, RNases will interact with all accessible mRNA regions, binding and cleaving where possible. RNase E and RNase G cleave mRNAs at either their 5' ends via end-dependent entry (**Figure 2-1C**) or at internal sites that remain accessible and unstructured (direct entry) [53-55] (**Figure 2-1D**). Other RNases (e.g. RNase III) bind to specific types of RNA structures, initiating their cleavage activity [56-57]. Notably, an RNA endonuclease's first cleavage event will create two new ends in the transcript, a mono-phosphate 5' end and a 3' end, that can then acted upon by the RNA degradosome [58], a loosely associated multi-protein complex, that combines both exonucleases and endonucleases (e.g. oligo-RNase, RNase R, PNPase) to rapidly reduce the transcript to short oligonucleotides and mononucleotides [59-60]. These activities include endonucleases that bind nearby monophosphate 5' ends and cleave internally, releasing oligonucleotides [53] as well as exonucleases that bind to structurally accessible 3' ends and chew them back in a 3' to 5' direction, releasing mononucleotides [61]. Helper enzymes

(e.g. PAPI) improve the accessibility of 3' ends by adding unstructured polyA tails [59], accelerating exonuclease activity. Altogether, after the first cleavage event takes place, large regions of the mRNA transcript can be rapidly processed to mononucleotides. Overall, the rate of the first RNase cleavage event is the slowest, making it a rate-limiting step in determining a mRNA's overall stability [62]. In particular, a first cleavage event inside the 5' UTR or intergenic regions is often rapidly followed by destruction of start codons in nearby coding regions, eliminating protein expression.

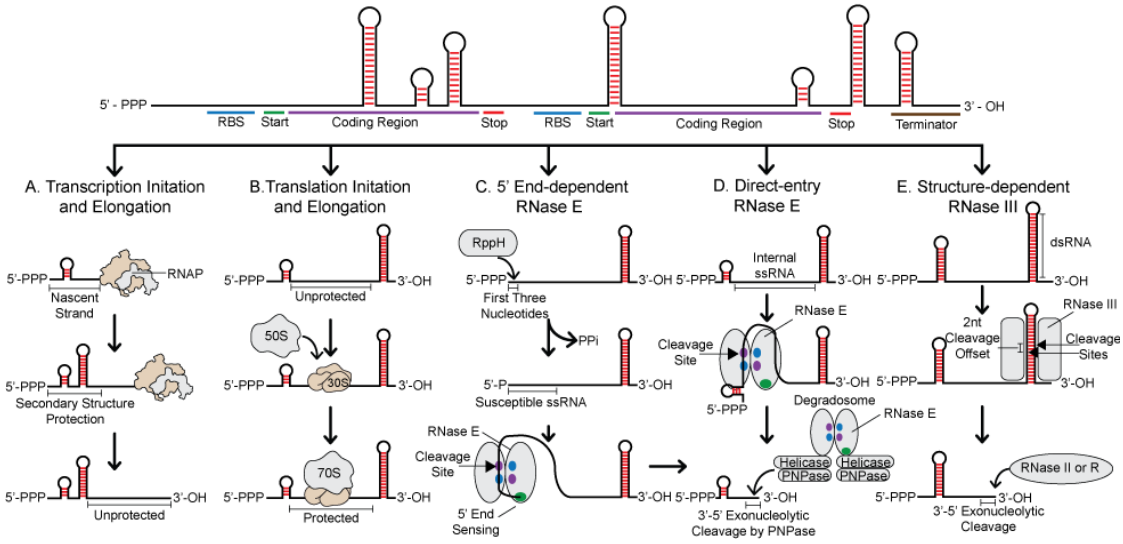

Figure **2-1**: mRNA sequence determinants control mRNA Stability through coupling of transcription, translation, and RNase activity. (A) Transcription elongation rates control mRNA structure and accessibility at 5' ends. (B) Translation initiation and elongation rates control the number of protective ribosomes bound to mRNA. (C) Both mRNA structure and the first three nucleotides at mRNA 5' ends controls RppH and RNase E activity via the end-dependent decay pathway. (D) RNase E binds and cleaves mRNA at single-stranded internal sites unprotected by structure. After initial cleavage, the RNA degradosome processively degrades mRNA. (E) RNase III targets specific mRNA structures for cleavage. PNPase, RNase II, and RNase R degrade mRNA at 3' ends using 3' to 5' exonuclease activity.

Previous studies have qualitatively observed several factors that alter a mRNA's decay rate, including (i) mRNA sequence motifs that may specifically bind RNases [63-64]; (ii) specific sequences at a mRNA's 5' end that alter RppH binding [47, 65]; (iii) the co-transcriptional, temperature-sensitive formation of mRNA structures that block RNase binding and increase mRNA stability [66]; (iv) the binding of ribosomes in 5' UTRs and coding regions that block RNase

binding, called protective ribosomes [42, 67-68]; (v) RNA structures at transcript 3' ends that block exonuclease activities [69]; and (vi) changes in environmental growth conditions that differentially alter RNase and helper enzyme levels (e.g. stress responses) [70]. RNase III is also known to cleave long RNA duplexes [71-72], while specific motifs, such as RAUUW [64], RNWUU [73], and RNAU [18, 53], were suggested to specifically bind RNase E. Overall, *E. coli* natural mRNAs have half-lives of about 1 to 10 minutes with the potential to vary protein expression levels by at least 10-fold [49, 74].

However, the precise sequence determinants that alter mRNA stability have not been quantitatively elucidated. Prior studies suggesting specific binding motifs have relied on a small number of characterized mRNAs, while the relationship between RppH activity, translation rates, transcriptional termination efficiency, and transcript stability remains poorly characterized. Because natural mRNAs are often attacked by RNases at multiple entry sites, it remains challenging to interpret natural sequence-stability assays and identify quantitative cause-effect relationships. Instead, we use a synthetic biology, learn-by-design approach to systematically alter a mRNA transcript's properties and measure its mRNA levels during sustained exponential growth conditions. Through this strategy, we designed, constructed, and characterized the mRNA levels from 82 mono- and bi-cistronic operons in *E. coli*, systematically varying the mRNAs' sequence and structural properties to modulate RppH, RNase E/G, and RNase III activities. We also systematically varied the mRNAs' translation initiation rates and transcriptional termination efficiencies to quantitatively determine how they protect mRNA transcripts from endonuclease or exonuclease activities. Altogether, our dataset and biophysical modeling provide a quantitative understanding of how mRNA sequence controls mRNA stability through several interactions.

Results

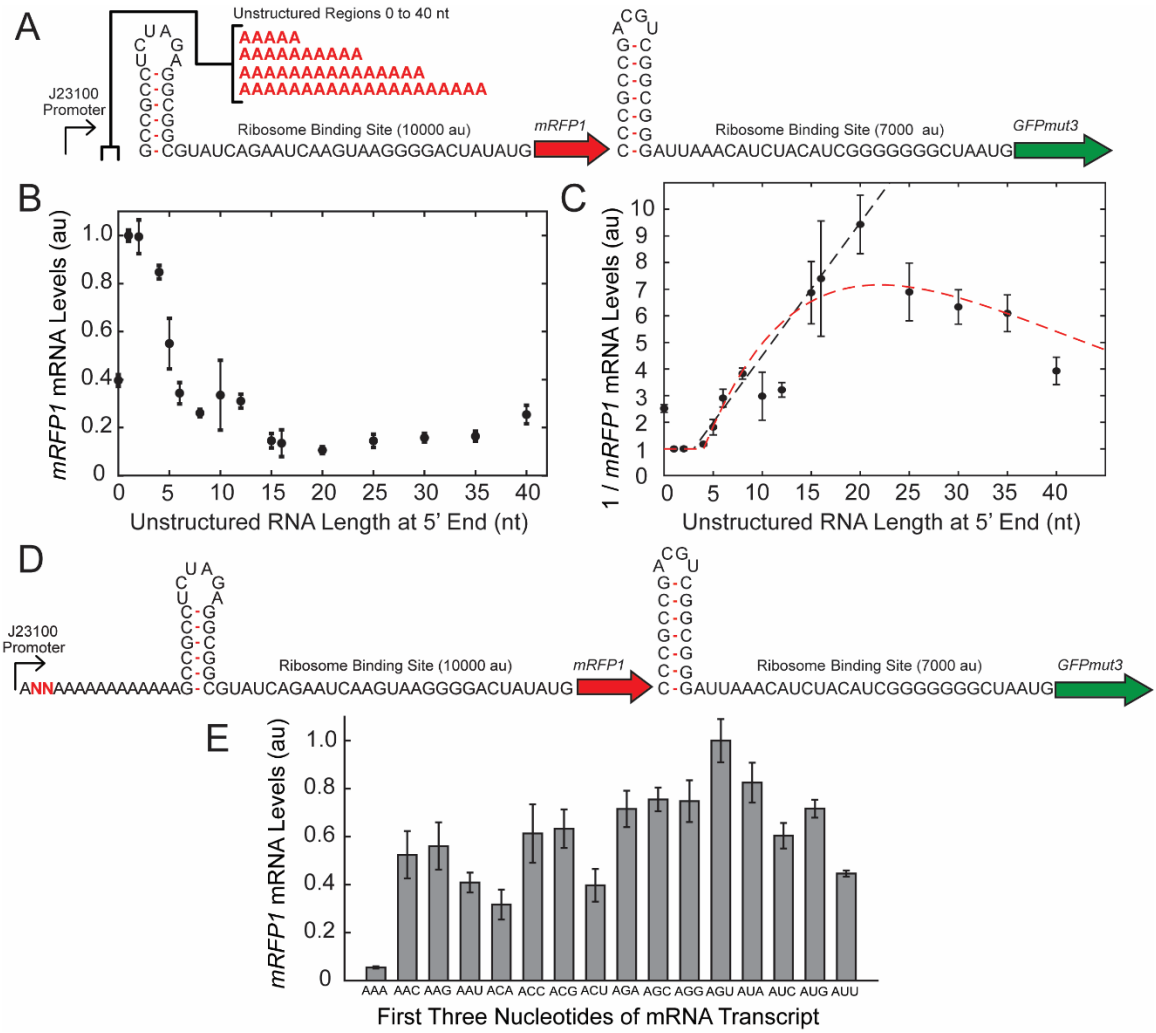

Figure **2-2**: Sequence determinants of mRNA stability at 5' ends. (A) Schematic showing a mRFP1-sfGFP operon with systematically varied 5' UTR sequences with between 0 to 40 nucleotides of single-stranded RNA, inserted upstream of an insulating hairpin. (B) RT-qPCR measurements of *mRFP1* mRNA levels show a quantitative relationship between single-stranded RNA length and mRNA stability. Dots and error bars are the mean and standard deviation of 2 biological and 6 technical replicates. RNA levels are plotted relative to an unstructured 5' UTR with one 'A'. (C) RT-qPCR measurements are replotted showing the relationship between the inverse mRNA level (proportional to decay rate) and single-stranded RNA length. Dashed black **(Equation 2-2)** and red **(Equation 2-3)** lines show predictions from two biophysical models. **Equation 2-2** calculates the change in a mRNA's specific decay rate based on the minimum RNase E/G binding site and the length of the ssRNA region. **Equation 2-3** builds on **Equation 2-2** by taking into account the RNA persistence length, which controls its propensity to form intramolecular loops. (D) Schematic showing a mRFP1-sfGFP operon with systematically varied dinucleotides at the 5' end. (E) RT-qPCR measurements of *mRFP1* RNA levels show how varied dinucleotides affected mRNA

stability. Bar heights and error bars are the mean and standard deviation of 2 biological and 6 technical replicates. RNA levels are plotted relative to the most abundant RNA 'AGU'.

Unstructured RNA in the 5' Untranslated Region Controls mRNA Stability

We first interrogated the sequence and structural determinants in the 5' untranslated region that affect a mRNA's stability in *E. coli*, combining systematic design, DNA assembly, and characterization using RT-qPCR and fluorescence measurements. To do this, we designed and constructed plasmid-encoded, bi-cistronic operons that utilize a constitutive sigma70 promoter (J23100), a rationally designed synthetic ribosome binding site, and two codon-optimized fluorescent protein reporters (mRFP1 and GFPmut3) (**Methods**, **Figure 2-2A**). The ribosome binding sites were designed by the RBS Calculator v2.1 to have moderate translation initiation rates (7000 to 10000 au). Using a sequence constraint, the RBS designs included a fast-folding, insulating mRNA hairpin ($\Delta G_{folding}$ = -12.61 kcal/mol) that prevents upstream sequence changes from changing the mRNA's translation initiation rate. Rationally designed upstream sequences were then inserted to replace the 5' untranslated region, beginning with the transcriptional start site and ending past the coding region's start codon. We then carried out long-time cultures of transformed strains, maintaining them in the exponential growth phase with periodic serial dilutions, followed by mRNA level measurements using RT-qPCR assays with TaqMan probes specific to the internal regions of reporter coding sequences (**Methods**). All sequences and measurements are located in **Appendix A-1**. All TaqMan probe sequences and measured efficiencies are located in **Appendix A-2**. The MIQE information for these experiments is located in **Appendix A-3**. By keeping the promoter unchanged and maintaining cells in the exponential growth phase for a long time period, changes in mRNA level can be primarily attributed to changes in mRNA stability, although we discuss potentially confounding factors below.

In the first series of 16 operons, we designed and inserted 5' UTR sequences that contained between 0 to 40 polyA nucleotides, creating systematically longer single-stranded RNA (ssRNA) regions upstream of the ribosome binding site's insulating hairpin (**Figure 2-2A**). Using Vienna RNAfold [75], we verified that these polymeric A sequences likely do not fold into mRNA structures that could potentially block RNase endonuclease activity. After characterizing these operons, we found a clear, quantitative relationship between the length of the ssRNA region and the resulting *mRFP1* mRNA level (**Figure 2-2B**). The highest mRNA levels occurred when the 5' UTR contained very short polyA sequences (2 or 3 As), followed by a precipitous decrease as the ssRNA region was lengthened. mRNA levels were reduced by about 2-fold when a ssRNA region was 5 nucleotides long. At its lowest, mRNA levels were reduced by 9.4-fold when 20 ssRNA nucleotides were added to the 5' UTR. However, further lengthening of the ssRNA region reversed the trend; we found that mRNA levels consistently increased when the ssRNA region was 20 to 40 nucleotides long. Surprisingly, without a polymeric A region (0 As), the presence of an insulating hairpin at the transcriptional start site actually reduced mRNA levels by about 2.5-fold.

With further analysis, we then developed biophysical models to explain how lengthening the ssRNA region could alter mRNA levels in such a non-linear fashion. When cultures are maintained in the exponential phase of growth, reaching steady-state conditions, mRNA levels will be inversely proportional to their decay or degradation rates, according to a balance equation:

$$\frac{d[mRNA]}{dt} = r_{TX} - k_{decay}[mRNA] = 0 \rightarrow [mRNA]_{SS} = \frac{r_{TX}}{k_{decay}}$$

Equation 2-1

Where $[mRNA]_{SS}$ is the steady-state mRNA concentration (level), r_{TX} is the promoter's transcription rate, and k_{decay} is the first-order kinetic constant quantifying the mRNA's specific decay rate.

We therefore replotted our data to show how the mRNA specific decay rates – the inverse of mRNA levels – are related to the lengths of the ssRNA region (**Figure 2C**). The trends in this relationship immediately suggested two types of biophysical models. In both models, RNases E/G bind and cleave mRNA anywhere where is a minimally sized landing pad containing an unstructured ssRNA region. If the unstructured region is too short, RNases E/G cannot bind and cleave mRNA. As the unstructured region is lengthened, the number of potential binding sites also increases, resulting in a proportional increase in RNase activity.

In the first model, we calculate the change in a mRNA's specific decay rate $\Delta\delta_{mRNA}$ by only considering the length of the smallest possible binding site for RNases E/G (N), the length of a ssRNA region (L), and an activity coefficient (C), according to the equation:

$$\Delta\delta_{mRNA} = C\,(L - N)$$

Equation 2-2

With a minimum binding site of 2 nucleotides (N = 2 nt) and an activity coefficient of about one half (C = 0.528 1 / nt), **Equation 2-2** can explain how mRNA specific decay rates depend on ssRNA regions from 1 to 20 nucleotides long (Pearson $R^2 = 0.92$, mean absolute error = 0.473) (**Figure 2-2C,** black dashed line). However, it's clear that this linear relationship is no longer true after the ssRNA region is more than 20 nucleotides long.

In the second model, we additionally take into account that untethered ssRNA molecules can dynamically form transient loop structures that limit their accessibility, and therefore may prevent RNases from binding to them. From the wormlike chain model of polymer theory, the likelihood that ssRNA with length L will *not* form an intramolecular loop is proportional to exp(-L / P), where P is the persistence length of ssRNA, which is a measure of its bending stiffness. As the length of a ssRNA region exceeds the persistence length, it becomes highly likely that the ssRNA will fold into a disordered structure. We therefore incorporated this effect to calculate the change in a mRNA's specific decay rate according to the equation:

$$\Delta\delta_{mRNA} = C\,(L - N)\exp(-L / P)$$

Equation 2-3

The persistence length of a ssRNA region depends on several properties, including RNA sequence, temperature, and the solution's salt composition, with previous measurements ranging from 1 to 7 nm or roughly 3 to 21 nucleotides [76-79]. Here, with a minimum binding site of 2 nucleotides (N = 2 nt), an activity coefficient of about one (C = 0.975 1/nt), and a persistence length of 21 (P = 21 nt), we found that **Equation 2-3** can explain how the mRNA's specific decay rate depended on ssRNA region's length from 1 to 40 nucleotides (Pearson $R^2 = 0.874$, mean absolute error = 0.767) (**Figure 2-2C**, red dashed line). Overall, the second model better predicts these mRNAs' specific decay rates across the entire dataset. However, when there is no single-stranded region (L = 0 nt), we do observe a discontinuity in both model predictions that could be caused by a confounding effect by another interaction, for example, a reduction in transcription rate.

First Three Nucleotides in the 5' Untranslated Region Control a mRNA's Stability

Previous *in vitro* measurements have shown that the first three nucleotides of a mRNA transcript have an effect on the rate of 5' end dephosphorylation by RppH or RppH-DapF complex [44], which will affect RNase E/G's ability to initiate mRNA decay via the end-dependent pathway. Here, we carried out *in vivo* measurements to quantify how the first three nucleotides of a mRNA transcript altered its mRNA level. We designed and constructed a series of 16 bi-cistronic operons that express mRFP1 and GFPmut3 where the 5' UTR contains a 'ANN' 5' end, followed by a polymeric 15A region, an insulating RNA hairpin, and a designed ribosome binding site with a translation initiation rate of about 10000 au on the RBS Calculator v2.1 scale (**Figure 2-2D**). The polyA region greatly reduces the likelihood of the formation of mRNA

structures regardless of the first three nucleotides in the transcript, which was verified by using Vienna RNAfold [75]. After maintaining cell cultures in the exponential growth phase, we carried out RNA extractions and RT-qPCR to measure *mRFP1* mRNA levels. We found that *mRFP1* mRNA levels varied by up to 18.2-fold when varying just the first three nucleotides of the transcript with the 5'-AAA-3' and 5'-AGU-3' trinucleotides have the most and least effect, respectively (**Figure 2-2E**). Even when excluding the most potent variant 5'-AAA-3', mRNA levels still varied by 3.2-fold, which is a potent effect for such a small sequence region. These results provide further support that the 5' end sequence affects how RppH and DapF interact with mRNA to accelerate end-dependent mRNA decay.

mRNA Translation Initiation Rate Controls mRNA Stability via Ribosome Protection

Previous studies have suggested that ribosomes can protect mRNA transcripts from RNase activity, possibly by covering RNase binding sites and limiting their accessibility [67, 80-83] (**Figure 2-3A**). However, the causal relationship between a mRNA's translation rate and its specific decay rate has never been systematically varied and measured. To determine such a relationship, we designed and constructed 18 mono-cistronic operons expressing either mRFP1 or sfGFP reporter, systematically varying their translation initiation rates by up to 1300-fold (**Figure 2-3B**). To do this, we inserted optimized RBS libraries into their 5' UTRs, each designed by the RBS Library Calculator to have an insulating hairpin, followed by a small number of pinpoint mutations that would greatly change the reporters' translation initiation rates [84]. We then characterized the operons' expression levels, measuring both their mRNA and protein levels using RT-qPCR and flow cytometry, respectively (**Methods**).

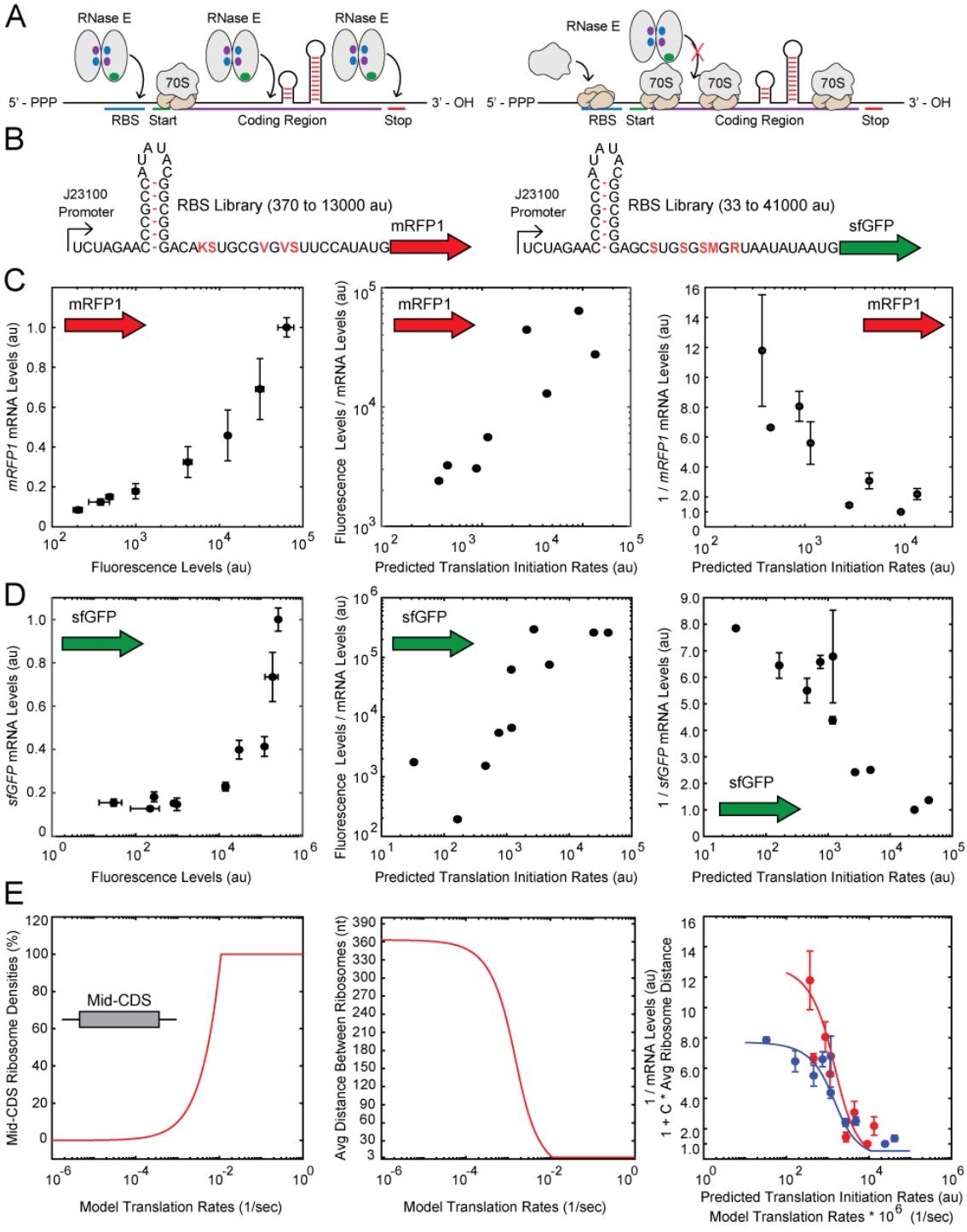

Figure **2-3**: Translation rate and mRNA stability are coupled via ribosome protection. (A) Schematics illustrating how a mRNA's translation rate controls the accessibility of RNase binding sites in an operon's coding regions. (B) Schematics showing the RBS library designs inside operons expressing either *mRFP1* or *sfGFP* reporter proteins. Highlighted nucleotides indicate the pinpoint mutations in RBS libraries. (C) Characterization of the *mRFP1* operons, showing (left) *mRFP1*

mRNA levels compared to mRFP1 fluorescence levels, (middle) the apparent mRFP1 translaton rates (the ratio of mRFP1 fluorescence over *mRFP1* mRNA levels) compared to the *mRFP1* predicted translation initiation rates, and (right) the inverse *mRFP1* mRNA levels compared to the *mRFP1* predicted translation initiation rates. (D) Characterization of the *sfGFP* operons, showing (left) *sfGFP* mRNA levels compared to sfGFP fluorescence levels, (middle) the apparent *sfGFP* translation rates (the ratio of sfGFP fluorescence over *sfGFP* mRNA levels) compared to the *sfGFP* predicted translation initiation rates, and (right) the inverse *sfGFP* mRNA levels compared to the *sfGFP* predicted translation initiation rates. (E) Calculations from a biophysical model of ribosome protection show (left) how translation rate controls ribosome density, (middle) how translation rate controls the average distance between adjacent ribosomes, and (right) a comparison between the average distance between ribosomes and the characterized mRNA decay rates (inverse mRNA levels) for (red) *mRFP1* and (blue) *sfGFP*. Dots and error bars showing mRNA level measurements are the mean and standard deviation from 2 biological and 6 technical replicates. Dots and error bars showing fluorescence level measurements are the mean and standard deviation from 2 to 4 biological replicates. Predicted translation initiation rates use the RBS Calculator v2.1 proportional scale.

We found that systematically increasing the mRNAs' translation initiation rates resulted in corresponding increases in both their mRNA and protein levels. For example, increasing the *mRFP1* translation initiation rate from 370 to 13000 au on the RBS Calculator v2.1 scale resulted in a 11.8-fold increase in *mRFP1* mRNA level and a 313-fold increase in mRFP1 fluorescence level (**Figure 2-3C**, left). Likewise, increasing the sfGFP translation initiation rate from 33 to 41000 au resulted in a 7.85-fold increase in sfGFP mRNA level and a 8760-fold increase in sfGFP fluorescence level (**Figure 2-3D**, left).

Here, the higher protein levels originated from both higher translation initiation rates and higher mRNA levels. To distinguish these sources, we calculated the apparent translation rates by taking the ratio of fluorescence level over mRNA level, and compared them to the RBS Calculator's predicted translation initiation rates (**Figure 2-3CD**, middle). We confirmed excellent quantitative agreement between the predicted and measured translation rates for both *mRFP1* and *sfGFP* with strong statistical significance (*mRFP1*: Pearson $R^2 = 0.782$, $p = 0.0035$; *sfGFP*: Pearson $R^2 = 0.706$, $p = 0.0023$). We then replotted the same predictions and measurements to determine how the mRNAs' specific decay rates – the inverse of our mRNA level measurements – were affected by the mRNA's translation initiation rates (**Figure 2-3CD**,

right). As expected, the lowest translation initiation rates yielded the highest mRNA specific decay rates, qualitatively consistent with a ribosome protection mechanism. However, quantitatively, our measurements could equally support a log-linear or sigmoidal relationship between translation rate and mRNA decay. To investigate further, we developed a biophysical model of ribosome protection that provides support for a sigmoidal relationship.

A Biophysical Model of Ribosome Protection Explains Translation-Decay Coupling

Consider a mRNA transcript that contains a protein coding sequence (CDS) with L amino acids. The key question is to calculate how many of these nucleotides remain unprotected by elongating ribosomes. Here, we designate α as the ratio of the ribosome's translation initiation rate over its elongation rate. We also designate F as the physical footprint of each ribosome in units of trinucleotides (amino acids). Based on prior measurements, we specify that the ribosome's footprint F is about 10 trinucleotides (30 nt)[33]. According to a TASEP (totally asymmetric exclusion process) model of ribosome dynamics that includes the ribosome's footprint on the mRNA [85], designated as an "extended object", a single equation can be used to approximately calculate how the translation initiation rate controls the ribosome's average density. When translation initiation is the rate-limiting step (α is less than one), the steady-state ribosome density ρ_r is:

$$\rho_r = \min\left(\frac{F\alpha}{1 + (F-1)\alpha}, 1\right)$$

Equation 2-4

when averaged over the middle portion of the CDS, away from the start and stop codons. The density of protected mRNA is $F\rho_r$, which includes the ribosome's footprint. By definition, the maximum number of bound ribosomes is L/F with a density of one (100% covered). The

relationship defined by **Equation 2-4** illustrates how increasing a mRNA's translation initiation rate results in higher ribosome densities, up until the maximum possible value (**Figure 2-3E**, left).

We next consider the unprotected distance D between adjacent ribosomes, which is the number of trinucleotides between the end of one ribosome and the beginning of the next ribosome, and is called the "headway distance". Here, the absence of a bound ribosome at a position along the mRNA is typically called a "hole". From the TASEP model with extended objects [86], the probability that a bound ribosome has *m* free trinucleotides in front of it is related to both the ribosome density ρ_r and hole density ρ_h according to:

$$P(D = m) \propto \frac{\rho_r}{\rho_r + \rho_h}\left(\frac{\rho_h}{\rho_r + \rho_h}\right)^m$$

Equation 2-5

where the hole density is determined by $\rho_h = 1 - F\rho_r$. The first part of **Equation 2-5** calculates the probability that a mRNA position is bound by a ribosome whereas the second part calculates the probability that the next *m* adjacent positions on the mRNA all contain a hole. We then substitute **Equation 2-4** and the definition of the hole density into **Equation 2-5**, yielding the probability distribution for the ribosomes' headway distances in terms of the ribosome density and footprint:

$$P(D = m) \propto \frac{\rho_r(1 - F\rho_r)^m}{(1 + \rho_r - F\rho_r)^{m+1}}$$

Equation 2-6

We then determine the ribosomes' average headway distance by calculating the first moment of the probability distribution in **Equation 2-6**, using:

$$\langle D \rangle = \frac{1}{Z}\sum_{m=1}^{m=L} P(D = m)\, m$$

Equation 2-7

where $Z = \sum_{m=1}^{m=L} P(D = m)$ to enforce the requirement that the integral over the probability distribution must always be one.

Altogether, for each selected translation initiation rate, we calculate the ribosome density using **Equation 2-4**, substitute it into **Equation 2-6** across a range of potential distances, and use this distribution and **Equation 2-7** to calculate the average distance between ribosomes (**Figure 2-3E**, middle). The overall relationship is a sigmoidal curve; at the highest translation initiation rates, the average distance plateaus to the smallest possible value (1 trinucleotide), whereas at very low translation initiation rates ($\rho_r \to 0$), the maximum possible average distance plateaus to:

$$\langle D \rangle_{max} = \frac{1}{L} \sum_{m=1}^{m=L} m = (L+1)/2$$

Equation 2-8

Perhaps counter-intuitively, $\langle D \rangle_{max}$ is not equal to L. To explain why, visualize what happens at very low translation initiation rates; the likelihood that a ribosome is bound to any position along the mRNA is about the same. Therefore, when a ribosome initiates translation at the start codon, the distance to the next ribosome on the mRNA is also uniformly distributed; any headway distance from 1 to L trinucleotides is equally likely.

Finally, according to a ribosome protection mechanism, we anticipated that the average distance between ribosomes – the amount of unprotected mRNA – is proportional to the mRNAs' decay rates. We therefore compared our inverse mRNA level measurements to the average distances between ribosomes across the 18 operons expressing *mRFP1* or *sfGFP* at varying translation initiation rates (**Figure 3E**, right) and found strong quantitative agreement with statistical significance (Pearson $R^2 = 0.823$, $p < 10^{-6}$), using proportionality constants $C_{mRFP1} = 0.10$ and $C_{sfGFP} = 0.058$. Here, the proportionality constant combines several factors, including the RNase E/G activity per unprotected mRNA, the likelihood that unprotected mRNA forms protective structures, and the scale of the mRNA level measurements. The developed biophysical

model of ribosome protection shows that the relationship between a mRNA's translation initiation rate and its specific decay rate is expected to be a sigmoidal curve, which is well supported by our measurements.

Location of RNase Binding Sites Controls mRNA Stability in Bicistronic Operons

We next investigated how the locations of RNase binding sites – inside either the 5' UTR, intergenic region, or 3' UTR regions – affected the mRNA levels of individual cistrons within a bicistronic operon. To do this, we designed and constructed 6 bicistronic operons expressing *mRFP1* and *GFPmut3*, inserting either a 20A ssRNA region (an RNase E/G site) or a 25 bp RNA hairpin (a RNase III site) into these three separate locations (**Figure 2-4A**). We applied Vienna RNA folding calculations to verify that the 20A region does not likely fold into a mRNA structure after being inserted into these locations. We performed similar calculations to verify that the 25 bp RNA hairpin does not fold into an alternative structure when inserted. We also constructed baseline operons that did not contain inserted RNase binding sites, which we used as the reference for comparisons. Otherwise, all operons utilized the same promoters, ribosome binding sites, and coding regions. Based on prior studies of gene expression polarity [69, 87-89], we anticipated the possibility that *mRFP1* and *GFPmut3* mRNA levels could be differentially affected by these inserted RNase binding sites.

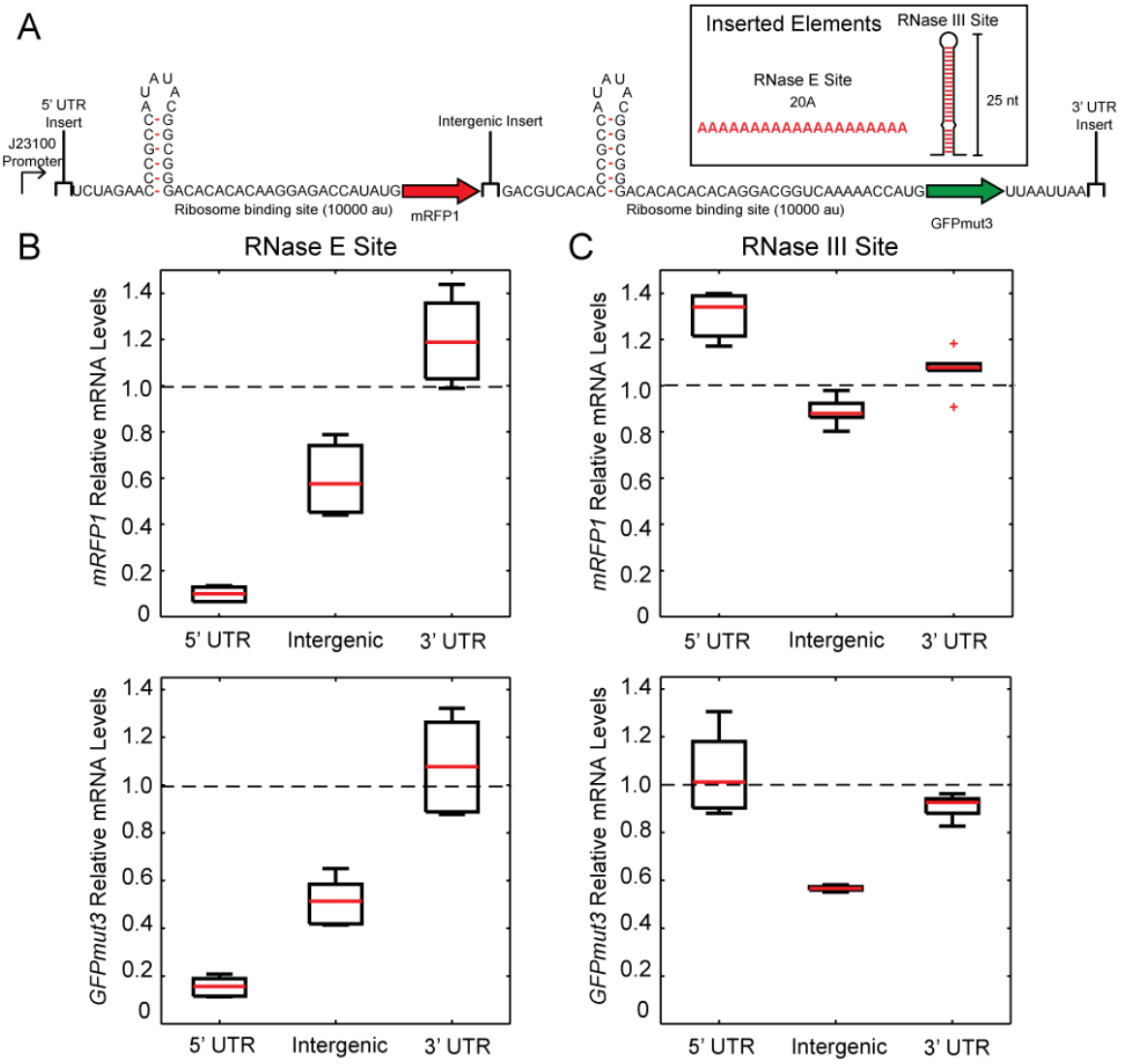

Figure **2-4**: Positions of RNase sites control mRNA stability. (A) Schematic illustrating the bistronic operons and the locations where RNase E/G or RNase III sites were inserted (boxed). The two types of inserted RNase binding sites are shown: either a 20A single-stranded RNA site or a 25 bp RNA hairpin structure. RNA levels were measured relative to the base operon without inserted RNase sites. (B) Characterization of the operons, showing how the positions of inserted RNase E sites affected the *mRFP1* and *sfGFP* mRNA levels. (C) Characterization of the operons, showing how the positions of inserted RNase III sites affected the *mRFP1* and *sfGFP* mRNA levels. Dots and error bars are the mean and standard deviation from 2 biological and 6 technical replicates.

As before, we characterized these operons by extracting total RNA from cells maintained in the exponential growth phase and applying RT-qPCR assays to separately measure *mRFP1* and

GFPmut3 mRNA levels (**Methods**). Consistent with our prior results (**Figure 2-2B**), we found that inserting a 20A ssRNA region into the 5' UTR greatly reduced the *mRFP1* mRNA levels (by 11-fold) with a concomitant decrease in *GFPmut3* mRNA levels (by 7-fold) (**Figure 2-4B**). However, inserting the same 20A ssRNA region into the intergenic region had a much lower effect; *mRFP1* and *GFPmut3* mRNA levels were reduced by only 1.7-fold and 2-fold, respectively. Notably, inserting the 20A ssRNA region into the 3' UTR region had no appreciable effect on the mRNA levels of either cistron. Overall, we found that RNase E/G sites had a potent effect on a mRNA's stability with a clear position-dependent trend; upstream sites accelerated mRNA decay more so than downstream sites. However, we did not observe any polarity effects; each RNase E/G site similarly affected the mRNA levels of the operon's *mRFP1* and *GFPmut* cistrons.

In contrast, only one of the inserted RNase III sites had an appreciable effect on the operons' mRNA levels, though it was accompanied by evidence of polarity (**Figure 2-4C**). Inserting a 25 bp RNA hairpin into the intergenic region did not appreciably affect the *mRFP1* mRNA level (1.13-fold change), but did decrease the *GFPmut3* mRNA level by 1.8-fold. Inserting the same 25 bp RNA hairpin into either the 5' UTR or 3' UTR had no appreciable effect on either mRNA level. Overall, these results suggest that the directionality and processivity of mRNA decay depend on the RNase that carries out the first cleavage event.

Unstructured RNA in the Intergenic Region Modestly Affects mRNA Stability

Following the previous results, we carried out a systematic investigation to quantitatively determine how single-stranded RNA (ssRNA) sites inside the intergenic region control mRNA stability. We designed and constructed 16 bicistronic operons expressing *mRFP1* and *GFPmut3*, inserting polyA ssRNA sites inside their untranslated intergenic regions, ranging from 0 to 40 nt

long (**Figure 2-5A**). We then carried out RNA extractions on cells maintained in the exponential growth phase, and measured both *mRFP1* and *GFPmut3* mRNA levels using RT-qPCR assays. Consistent with our previous results, we found that unstructured RNA inside the intergenic region has an appreciable, but modest, effect on mRNA stability, affecting both upstream and downstream cistrons similarly (**Figure 2-5B**). Increases in ssRNA length (3 to 40 nt) decreased mRNA levels by modest amounts, reaching a minimum of 1.66-fold lower *mRFP1* and 1.43-fold lower *GFPmut3* mRNA levels.

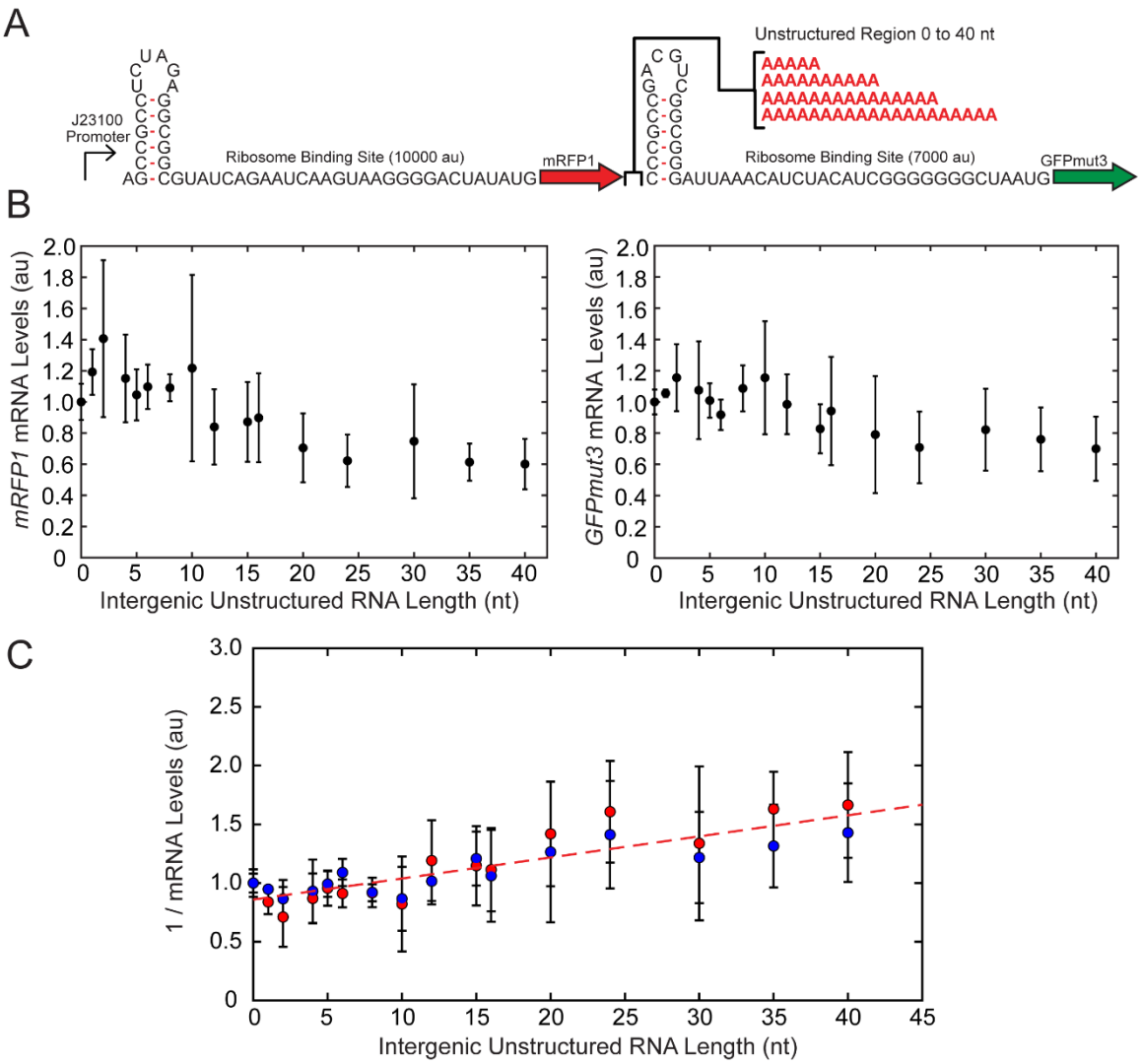

Figure **2-5**: Sequence determinants of mRNA stability at intergenic sites. (A) Schematic illustrating bicistronic operons and the insertion of single-stranded RNA (RNase E/G binding sites) with varying lengths into the intergenic region. RNA levels were measured relative to the base operon without inserted RNase sites. (B) Characterization of the operons, showing (left) *mRFP1* mRNA levels compared to the single-stranded RNA lengths and (right) *GFPmut3* mRNA levels compared to the single-stranded RNA lengths. (C) The inverse mRNA levels for (red) *mRFP1* and (blue) *GFPmut3* are compared to the single-stranded RNA lengths alongside a (red dashed line) linear relationship with slope C = 0.018. Dots and error bars are the mean and standard deviation from 2 biological and 6 technical replicates.

As before, we replotted our measurements to illustrate how the inverse *mRFP1* and *GFPmut3* mRNA levels – proportional to the mRNA's specific decay rate – are controlled by the ssRNA length within intergenic regions (**Figure 2-5C**). Similar to 5' UTRs, we found that the ssRNA region must be at least 2 nt long before we observe any increase in mRNA decay. However, unlike 5' UTRs, we observed only a linear quantitative relationship between ssRNA length and mRNA decay with a small slope (activity coefficient C = 0.018). Overall, these results show that RNase E/G is not able to strongly bind and cleave intergenic regions even with large ssRNA regions, resulting in much smaller changes in mRNA decay rates. Moreover, we do not observe appreciable amounts of differential mRNA decay across upstream and downstream cistrons (indicative of polarity) or attenuation of mRNA decay at higher ssRNA lengths (indicative of transient loop formation).

Transcriptional Terminator Efficiency Has No Effect on mRNA Stability

For our last dataset, we investigated how the efficiency of an operon's transcriptional terminator affected its mRNA stability. To do this, we designed and constructed 8 operons expressing *mRFP1*, inserting different intrinsic transcriptional terminators into their 3' UTRs (**Figure 2-6A**). Each terminator was selected from a toolbox of well-characterized synthetic terminators to have varying termination efficiencies from 8.3% to 99.7% [90]. We also constructed a control operon that does not have a transcriptional terminator at the insertion position (0%

efficiency). As an additional test, we also searched for any transcriptional terminator sequences downstream of the insertion site. Using the FindTerm terminator finder [91], the closest transcriptional terminator candidate was at least 230 base pairs downstream of the *mRFP1* stop codon. Otherwise, all operons have identical promoters, ribosome binding sites, and protein coding sequences.

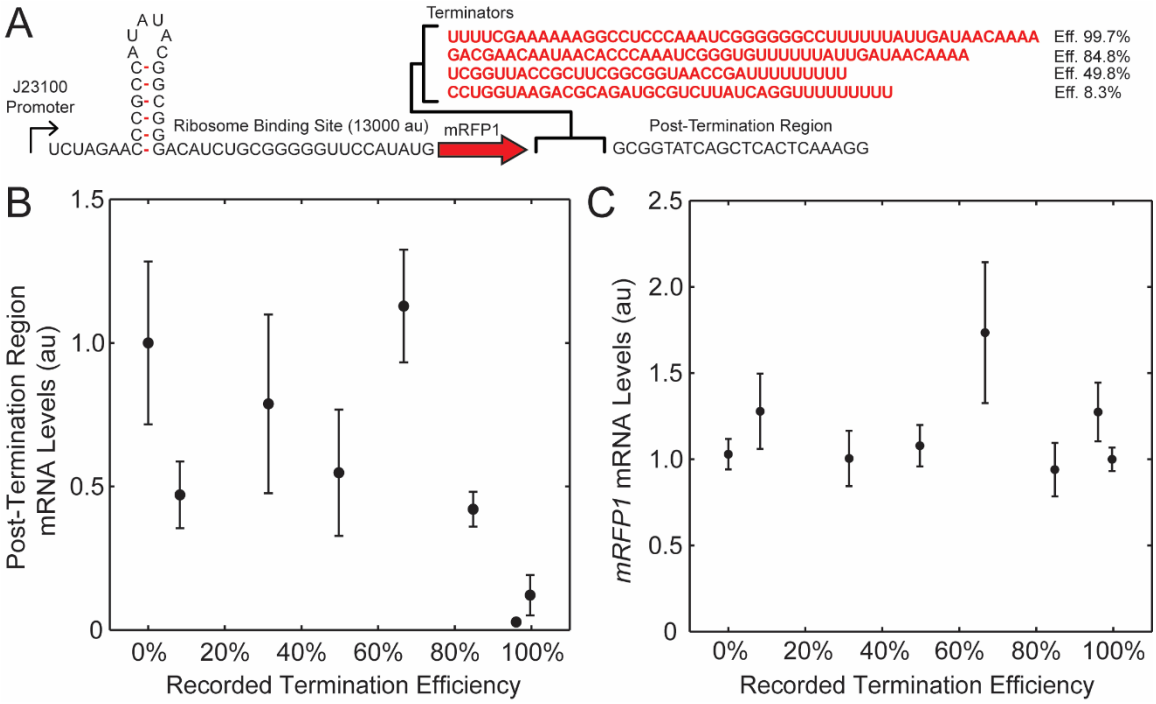

Figure **2-6**: Transcriptional terminator efficiency does not impact mRNA stability. (A) Schematic illustrating *mRFP1* operons containing transcriptional terminators with varied efficiencies. (B) Characterization of the post-terminator mRNA levels using SYBR Green RT-qPCR assays, compared to prior measurements of the transcriptional terminators' efficiencies. RNA levels were measured relative to the base operon without an inserted terminator. (C) Characterization of the *mRFP1* mRNA levels using TaqMan RT-qPCR assays as compared to the transcriptional terminators' efficiencies. Dots and error bars are the mean and standard deviation from 2 biological and 6 technical replicates.

As before, we characterized the mRNA levels of these operons during exponential growth. For our first measurement, we applied SYBR Green RT-qPCR to measure the mRNA level of the region downstream of the inserted transcriptional terminators and to quantify the

efficiencies of the transcriptional terminators (**Figure 2-6B**). We found that the levels of the post-terminator mRNA region were only greatly reduced once the previously recorded terminator efficiencies exceeded 80%. At the same time, we applied TaqMan RT-qPCR to measure the *mRFP1* mRNA levels. Surprisingly, we found no appreciable differences in *mRFP1* mRNA level across the full range of terminator efficiencies (**Figure 2-6C**). The *mRFP1* mRNA levels were similar whether a highly efficient terminator, a low efficiency terminator, or a non-terminating control sequence were inserted downstream of the *mRFP1* coding sequence. For the variants with little or no termination efficiency at the insertion site, it is likely that RNA polymerase continues to transcribe the mRNA for at least 230 nucleotides past the *mRFP1* stop codon, due to the presence of putative terminators and interactions with the ColE1 origin of replication, which would greatly lengthen the 3' UTR. These results suggest that the length and structure of the 3' UTR has little appreciable effect on the mRNA's decay rate, which is consistent with the prior insertion of unstructured RNA into the 3' UTR region (**Figure 2-4**).

Discussion

We designed and characterized 82 operons to quantify the sequence and structural determinants controlling mRNA stability in *E. coli*. We introduced rationally designed sequences into the 5' UTR, intergenic, and 3' UTR regions within mono-cistronic or bi-cistronic operons to measure how they affected the operons' mRNA levels. Through this learn-by-design process, we systematically investigated how these determinants affected each mRNA decay pathway (**Figure 2-1**), including (i) the amount of unstructured, single-stranded RNA (ssRNA) in the 5' UTR, which controls the rate of RNase E/G cleavage via the end-dependent pathway (**Figure 2-2B**); (ii) the first three nucleotides of the transcript, which controls RppH-DapF dephosphorylation at the 5' end (**Figure 2-2E**); (iii) the translation rate of coding sequences, which controls the amount of

ribosome protection (**Figure 2-3**); (iv) the types and locations of RNase binding sites, which controls decay rate and polarity (**Figure 2-4**); (iv) the amount of single-stranded RNA in the intergenic region, which controls the rate of RNase E/G cleavage via the direct entry pathway **(Figure 2-5**); and (v) the efficiency of the transcriptional terminator, which controls the sequence, structure, and length of the 3' UTR (**Figure 2-6**).

Overall, modifying the 5' UTR of a mRNA transcript had the largest effect on its steady state level. Inserting a 20 nucleotide single-stranded RNA (ssRNA) region into the 5' UTR led to a 9.4-fold decrease in mRNA level. Lowering the translation initiation rate of the first coding sequence by 35-fold – through redesigning its ribosome binding site – led to a 11.8-fold decrease in mRNA level. Modifying just the first three nucleotides of the transcript could lead to 18-fold change in mRNA level. In contrast, changing the intergenic region's sequence and structure had much less potent effects. Inserting a 20 nucleotide ssRNA region into the intergenic region reduced mRNA levels by up to 2-fold, similarly affecting both upstream and downstream genes in a bi-cistronic operon (no polarity). Inserting a long RNA hairpin into the intergenic region mainly affected the downstream coding sequence's mRNA level (by 1.8-fold). Finally, we were genuinely surprised that our substantial changes to the 3' UTR sequence, structure, and length had no appreciable effect on the transcript's mRNA level. Based on our measurements, the transcriptional terminator does not alter a mRNA's specific decay rate and only facilitates decoupling of cistrons in adjacent operons. Overall, our results show that end-dependent decay – mediated by RNase E/G and RppH-DapF – is the predominant pathway controlling mRNA stability. RNase E/G likely bind internal unstructured sites via the direct entry pathway with much lower affinity, requiring long stretches of unstructured RNA (e.g. entire coding regions) to achieve high cleavage rates.

Leveraging our measurements, we developed several biophysical models to explain how these sequence determinants controlled mRNA specific decay rates. Each model contains a small number of parameters, but can explain the observed quantitative trends. Previously, RNase E/G were suggested to bind to specific binding motifs (e.g. RAUUW or RNWUU). Here, we propose that RNase E/G simply bind unstructured RNA regions with a 2 nucleotide landing pad. Longer unstructured regions therefore provide a proportionally larger number of potential binding sites for RNase E/G with an expected increase in cleavage rate (**Equation 2-2**). When this model is applied to the 5' UTR, it fits our mRNA decay rate measurements extremely well for when the ssRNA length is varied from 1 to 20 nucleotides (**Figure 2-2C**). However, when the 5' UTR ssRNA region is very long, we see an attenuation of this effect, suggesting that the region is no longer fully unstructured. We therefore turned to polymer theory, which provides well-established equations for calculating the likelihood that a polymer forms transient, non-specific structures (ie. wormlike chains), which determines the fraction of polymer that remains accessible. By incorporating polymer theory (**Equation 2-3**), a single model was able to explain how ssRNA length affects mRNA decay rates across the larger range of lengths from 1 to 40 nucleotides, using a persistence length that is consistent with prior measurements.

We applied the same biophysical modeling (**Equation 2-2**) to the intergenic region with a similarly consistent explanation. Longer ssRNA regions led to proportionally higher mRNA specific decay rates (**Figure 2-5**). However, the proportionality constant in this model was much lower, consistent with our observations that the direct entry pathway is much slower than the end-dependent pathway, when acting on the same amount of ssRNA. Notably, we did not observe a plateau in specific decay rate on very long unstructured intergenic regions, which is expected because the formation of transient wormlike structures requires at least one freely movable (untethered) end. While the 5' end of the transcript can freely move, both ends of the intergenic region are physically constrained by the surrounding ribosomes engaged in translation elongation.

We also developed a biophysical model of ribosome protection that shows how systematically varying a coding sequence's translation rate controls its mRNA level (**Figure 2-3**), applying equations to calculate a mRNA transcript's ribosome density and ribosome-to-ribosome headway distance (**Equations 2-4 to 2-8**). The model provides a mechanistic basis for the observed sigmoidal-like relationship in our measurements with expected plateaus in mRNA level at either very low or very high translation rates, and with sharp changes in between. Here, we treated the CDS region as being equally accessible to RNase E/G activity regardless of their RNA structure. We also assumed that all codons have the same translation elongation rates. In return, the resulting equations can be analytically solved and remain remarkably simple to evaluate. However, the observed differences in proportionality constants for *mRFP1* and *sfGFP* could arise from these assumptions.

Collectively, our biophysical models suggest straight-forward rules for designing operons with controlled mRNA stabilities. For maximum stability, 5' UTRs should begin with 5'-AGN ends and support moderate-to-high translation initiation rates (10000 or higher on the RBS Calculator v2.1 scale), while limiting the amount of single-stranded RNA outside the ribosome binding site. For multi-cistronic operons, intergenic regions should also contain ribosome binding sites that support moderate-to-high translation initiation rates, though the amount of single-stranded RNA will not appreciably affect their stability. Our design rules here do not consider the presence of self-cleaving ribozymes, which introduce a 5′-hydroxyl group that slows down end-dependent mRNA decay [45]. To achieve lower mRNA stabilities, transcripts can be designed to have 5'-AAN ends as well as single-stranded RNA regions in their 5' UTR regions, upstream of the ribosome binding site. Perhaps most surprisingly, the design of the 3' UTR does not affect the operon's expression levels and primarily controls read-through transcription into downstream operons. Altogether, our measurements and models provide the quantitative means to controlling mRNA stability in operons.

Methods

Plasmid Design and Cloning

A series of libraries of pFTV1-derived plasmids with a ColE1 origin of replication and Chloramphenicol resistance (Cm^R) were designed and constructed to express fluorescent protein reporters with the objective of testing specific design motifs' effect on mRNA stability. In **Appendix A-1**, the dataset name and sequence information are listed for each construct. For the datasets mRFP1 Expression Library and Terminator Efficiency, a base mono-cistronic mRFP1 pFTV1-derived plasmid was used that contained the σ^{70} constitutive promoter J23100 and a double terminator. To create the mRFP1 Expression Library, the 5' UTR was engineered using RBS Library Calculator to design a set of ribosome binding sites that varied the expression of mRFP1 1000-fold. The degenerate oligonucleotide (IDT) designed by RBS Calculator containing these designs was flanked by a 5' XbaI and 3' NdeI restriction site and primer binding sites, which allowed for PCR amplification. The plasmid was then assembled using cut and paste cloning. Similarly, the Terminator Efficiency set used construct 6 from the mRFP1 Expression Library as the base plasmid. The terminators tested were assembled by annealing two complementary oligonucleotides (IDT) with 5' FseI and 3' SphI restriction overhangs. The annealed oligonucleotides were ligated into the base plasmid replacing the existing terminators. Likewise, for the datasets sfGFP Expression Library and sfGFP Coupled Expression, the base plasmid contained the σ^{70} constitutive promoter J23100 and a double terminator. RBS Libraries were designed and cloned in for both using a similar method as for the mRFP1 Expression Library, except that the restriction sites employed were 5' XbaI and 3' AgeI. For the datasets, 5' Variable Length Poly A, Intergenic Variable Length Poly A, RNase Polarity, and RppH Set, a base bi-cistronic codon optimized mRFP1 and GFPmut3B pFTV1-derived plasmid was used that

contained the σ^{70} constitutive promoter J23100 and a double terminator. At the 5' UTR, RBS Calculator v2.1 was used to design a ribosome binding site sequence with an XbaI restriction site on the 5' end and with a translation initiation rate of 9395 au on RBS Calculator's proportional scale. The newly designed 5' UTR and J23100 promoter was assembled by annealing two complementary oligonucleotides (IDT) with 5' BamHI and 3' SacI restriction overhangs and ligating them into the base plasmid. Likewise, RBS Calculator v2.1 was used to design a ribosome binding site sequence with an AatII restriction site on the 5' end with a translation initiation rate of 7083 au on RBS Calculator's proportional scale. The newly designed intergenic region was assembled by annealing two complementary oligonucleotides (IDT) with 5' EcoRI and 3' XhoI restriction overhangs and ligating them into the base plasmid. Additions to the base 5' UTR were made using a 5' BamHI site upstream of the promoter and the newly added 3' XbaI site. Sequence additions to the base intergenic UTR was made using an upstream 5' EcoRI site and the newly added 3' AatII site. For additions to the 3' UTR, a 5' PacI site and 3' SpeI site were used. All additions were made by annealing complementary oligonucleotides (IDT) with the correct restriction overhangs for the 5' UTR, intergenic region, and 3' UTR. All plasmids were transformed into *Escherichia coli* DH10B cells, followed by sequence verification of isolated clones.

Strains, Growth and Characterization

All measurements were conducted using *Escherichia coli* DH10B cells containing plasmids, cultured using M9 minimal media, and maintained in the exponential growth phase for at least 20 hours. For each construct, isogenic colonies were used to inoculate overnight cultures in 500 ul LB media supplemented with 50 ug/mL Chloramphenicol (Cm) in a 96-well deep-well plate. Overnight cultures were diluted 100-fold by diluting 2 uL of culture into 198 ul of M9

minimal media with Cm using a 96-well microtiter plate, and incubated at 37 ^{0}C with high orbital shaking inside a Spark spectrophotometer (TECAN). OD_{600} absorbance was taken every 10 minutes until the OD_{600} reached 0.15-.20, indicating the cells' entry into the mid-exponential phase of growth. At this time, a subsequent 96-well microtiter plate was prepared by serial dilution of the culture from the first plate into M9 minimal media with Cm maintaining the cells in the exponential phase of growth. For each culture, single-cell red florescent protein (RFP) or green florescent protein (GFP) fluorescence measurements were performed by collecting 10 ul from the end of the second dilution, transferring to a microtiter plate with 200 ul PBS solution with 2 mg/ml kanamycin, and recording 100,000 single-cell fluorescence levels with a Fortessa flow cytometer (BD Biosciences). All single-cell fluorescence distributions were unimodal. The mean of the distributions is calculated, and the background autofluorescence of *Escherichia coli* DH10B cells subtracted. All reported fluorescence levels are the average of at least two biological replicates, which are listed in **Appendix A-1**.

mRNA level measurements were performed on selected strains by inoculating a 5 mL culture of Cm supplemented LB media and incubated at 37 ^{0}C with 300 RPM shaking. Once cells reached an OD_{600} of 1.0, measured using a cuvette-based spectrophotometer (NanoDrop 2000C), they were diluted to an OD_{600} of 0.05 in a 5 mL culture of Cm supplemented M9 minimal media and incubated at 37 ^{0}C with 300 RPM shaking. Once cells reached an OD_{600} of between 0.4-0.6, they were diluted to OD 0.001 in a 5 mL culture of Cm supplemented M9 minimal media and incubated at 37 ^{0}C with 300 RPM shaking. Cells were harvested once they reached an OD_{600} of between 0.15-0.25 and their total RNA extracted using the Total RNA Purification kit (Norgen Biotek), followed by non-specific degradation of contaminant DNA using the Turbo DNAse kit (Ambion). Following extraction, cDNA was prepared using the High Capacity cDNA Reverse Transcription kit (Applied Biosystems). Taqman based qPCR was performed using an ABI Step One Plus real-time thermocycler (Applied Biosystems), utilizing a Taqman probe targeting the

constructs of interest mRFP1, codon optimized mRFP1, sfGFP, codon optimized GFPmut3B. Additionally, a custom 16S rRNA TaqMan probe was used as an endogenous control and was used to calculate relative mRNA levels from ΔC_t numbers. TaqMan probes sequence and validation are listed in Supplementary Data 2. Likewise, SYBR Green based qPCR was performed in a similar manner for the determination of the relative mRNA level.

Chapter 3

Comprehensive modeling and design of 5' UTRs for RNA stability using a 62,000 unique 5' UTR Library

mRNA degradation is a central process that affects all gene expression levels, and yet, the determinants that control mRNA decay rates remain poorly characterized. To better predict gene expression in both natural and synthetic systems, we applied a synthetic biology, learn-by-design approach to elucidate the sequence and structural determinants that control mRNA stability in bacterial operons. We designed 62,120 unique 5' UTRs to further explore the effect of translation initiation rate, secondary and tertiary structure, sequence composition, and RNA binding protein recognition sites on mRNA decay. Pairing rifampicin-based time course RNA extraction with next generation sequencing (NGS), we measured the degradation rate for each member of the sequence library. Using this comprehensive library, we employed machine learning tools to construct a predictive model of mRNA decay rate for the design and analysis of bacterial operons.

Introduction: Modeling mRNA Decay Processes

Accurately predicting overall gene expression from DNA sequence alone combined with having the ability to craft sequences to tune the expression of a gene or set of genes remains one of the chief goals of synthetic biology. While significant work and progress has been achieved on predicting transcription[92-96] and translation rates[14, 32, 97], far less focus has been given to predicting RNA decay rates[98-99]. The continued development and myriad applications of complex genetic circuits and multi-enzyme metabolic pathways would benefit greatly from greater precision in mRNA half-life design, allowing for tighter control of expression. Controlling mRNA stability would be particularly useful for circuit tuning; transcripts could be designed to rapidly decay, allowing for fast circuit transitions or designed to persist[100]. Additionally, purposefully designing

and knowing the mRNA half-life of multi-enzyme metabolic pathways could help balance enzyme concentrations to create more efficient metabolic flux and avoid concentration of toxic intermediates[101-102]. Overall, having the ability to determine if a novel genetic construct will have undesirable mRNA stability before construction, will be broadly useful.

Over the past decades, extensive work has been completed on determining the enzymes responsible for controlling mRNA decay. *E. coli* endonucleases such as RNase E[103-104], RNase G[62, 103], and RNase III[72, 105] along with exonucleases such as RNase II, RNase R, and PNPase[59-60] govern the vast majority of mRNA turnover. Additional helper enzymes, such as RppH[44, 47] and Poly(A) polymerase[59], chemically modify the mRNA to alter its susceptibility to RNase attack. Similarly, RNA biding proteins such as CsrA[51] and Hfq[106] physically bind to the mRNA transcript and alter the translation rate and mRNA decay rate. Hammerhead ribozymes can be appended to the 5' UTR, which catalyzes the self-cleavage of the 5' UTR and the formation of a more stable 5' UTR hairpin[107]. Furthermore, past studies have shown that many of these enzymes have unique sequence and RNA secondary structure preferences that impact both the binding location and catalytic activity of the enzyme[108-109]. The fundamental challenge of modeling RNA decay is that a broad combination of enzymes act in concert on multiple sequence and structural elements in the mRNA to degrade the mRNA molecule.

This fundamental information has been used to model some of the individual processes of degradation or identify specific locations in the *E. coli* transcriptome that are susceptible to attack by each of the enzymes. The *E. coli* transcriptome has been sequenced and many of the half-lives of the 4288 genes organized into 2584 operons has been determined[42, 110-111]. However, while the results from these experiments are greatly informative, 2584 transcripts are not sufficient to explore the possible design space available for designing mRNAs. Furthermore, since the context of each gene is not controlled, it is difficult to elucidate concrete rules governing stability or determine how different sequences and structures act together to produce an overall turnover rate.

Furthermore, the scale of the design space possible in the 5' UTR makes it difficult to generalize from a few sequences. In particular, a 30 nucleotide 5' UTR has the potential for 4^{30} different possible UTR designs.

To answer this question, a pool of 62,120 unique 5' UTRs was computationally designed to explore the sequence and structure design space of mRNA in *E. coli.* We previously showed that the 5' UTR is the most important region in controlling the decay rate of mRNA and focused this study on that region[98]. The same promoter and coding region (sfGFP) was used to isolate the impact of the 5' UTR. The entire library was not designed by inserting random sequences, but instead rationally designed to test individual aspects of decay and combine them together to test combinatorial effects. It was designed to contain many different structure and sequence motif such as every combination of the first four nucleotides in the transcript to test the impact of RppH on stability, the full range of ribosome translation initiation rates, different nucleotide composition combinations of unstructured regions of the mRNA, unstructured regions of the mRNA flanked between two hairpins, sequences of differing length, hairpins of different length, diverse secondary structures such as hairpins that contain internal loops and bulges, G-quadruplexes, i-motifs, known RNase E binding sites, CsrA binding sites, and lastly hammerhead ribozymes. These different design criteria were paired together to explore the different combinations of interactions that impact mRNA decay. The entire pool's mRNA decay rates were measured in a time-course Illumina Novaseq NGS experiment. These results were analyzed to determine the quantitative impact of the major trends of mRNA decay.

While other models have looked at some basic components of mRNA decay, our model looks to combine all of the interactions both large and small that influence the decay of mRNA. We first verified that our dataset shows the same major trends in mRNA decay that have been shown previously. However, to develop the model to be more predictive, we wanted to capture the impact of the many weak interactions that contribute to the overall decay rate. To accomplish

this, we chose to pair our rational design of the library with a machine learning approach. These results were fed into the machine learning algorithm LightGBM[112], which is a gradient boosting regression algorithm. We took each mRNA sequence and calculated 357 biophysical features, allowing us to build a comprehensive model of mRNA decay. The model allows for any 5' UTR to be tested for it predicted stability.

Results

To build our model, we used a high-throughput testing workflow to design, build, and test the sequence and structural features of mRNA decay in *E. coli* **(Figure 3-1A)**. We designed 62,120 sequences that explore the design space of the 5' UTR of *E. coli*. We created a diverse set of sequences that explore a wide range of sequence and structural features of the 5' UTR **(Figure 3-1B)**. We ordered the designs as a set of five oligopools. The oligopools were inserted into our plasmid backbone in such a way that every construct was identical except for the designed 5' UTR and a barcode placed after the reporter protein sfGFP. The genetic context of the 5' UTR was kept constant to minimize decay effects unrelated to the UTRs of interest. The pool of constructed plasmids was transformed into *E. coli* and grown as a single culture. The culture was treated with rifampicin at high concentration and then sampled at 0, 2, 4, 8, and 16-minutes post-addition of rifampicin. Rifampicin arrests RNA Polymerase, but not the decay machinery, which allows for kinetic data to be determined. Additionally, this method further avoids issues related to the context of the genetic construct because the starting amount of mRNA, which could be potentially influenced by the context of the promoter, will not impact the decay rate calculation **(Methods)**. The kinetic information was used to calculate the decay rate where N_0 is the initial amount of mRNA at rifampicin addition, N(t) is the amount of mRNA remaining at time t, and λ is the decay rate **(Equation 3-1)**.

Equation 1: $N(t) = N_0 e^{-\lambda t}$

Equation 3-1

For the model predictions, the RNA levels at each timepoint were normalized by the DNA counts for each member of the library. The machine learning algorithm LightGBM was used to predict the normalized RNA level at each timepoint using 357 calculated features based on the sequence and structure of the RNA. The overarching sequence and structural features of the 5' UTR were broken down into the component features. These features translate general ideas of sequence and structure into quantifiable predictors of decay. The predicted RNA level outcomes from each of the models was then normalized and fitted to calculate the decay rate.

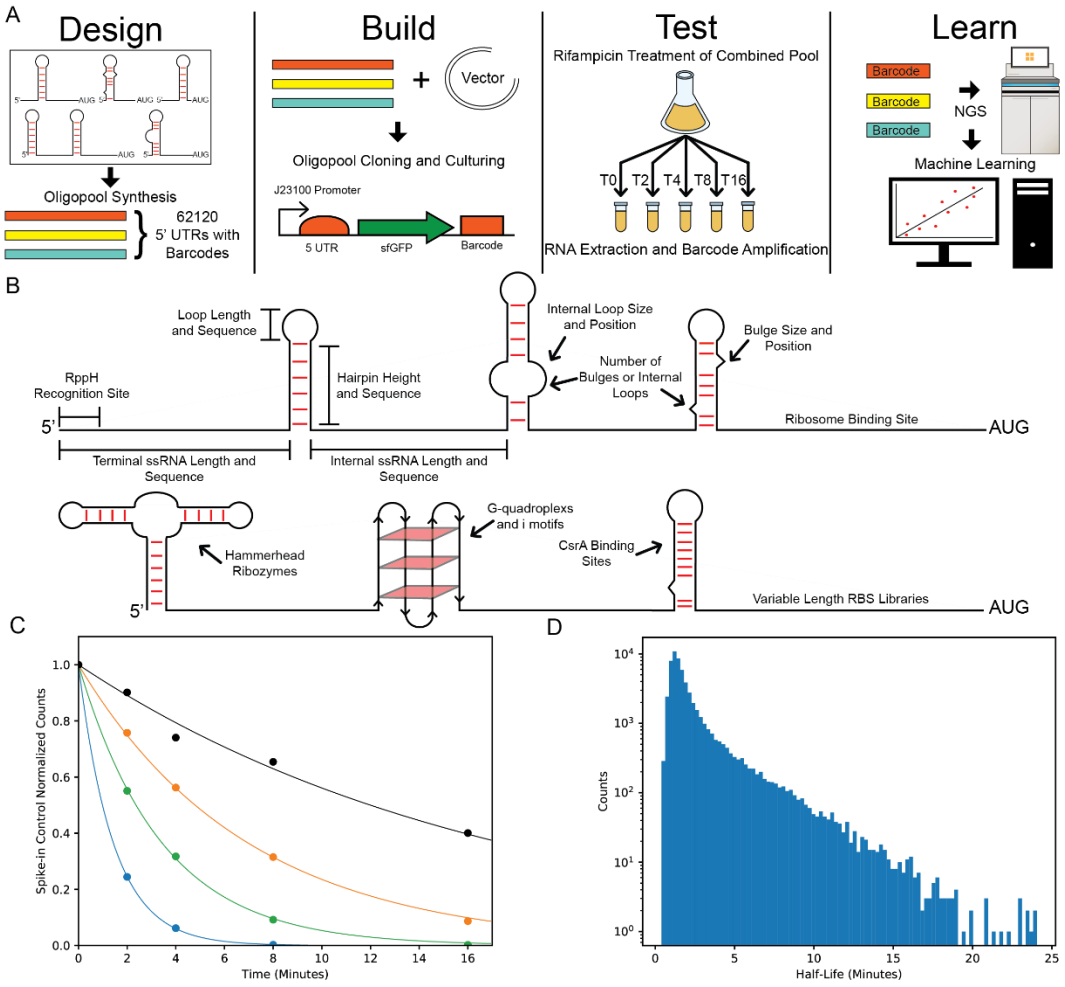

Figure **3-1**: Library Design and Half-life Calculations (A) An NGS workflow was used to design a 62,120-member library. Following design, the library was constructed in a plasmid vector and transformed into E. coli. The sample was treated with Rifampicin to arrest transcription and the culture was sampled at 0,2,4,8, and 16 minutes post addition. Extracted DNA and RNA was run on Novaseq for use to build a machine learning model of RNA decay. (B) Many different sequence and structural considerations were made when designing the library. Library members were designed to systematically vary the sequence, translation initiation rate, and secondary and tertiary structural characteristics. (C) The time course data was used to fit the half-life equation to determine the decay rate. (D) The half-life of all of the library members is plotted. The half-life of most of the sequences is less than 5 minutes.

For each construct the decay rate was determine using the kinetic data determined from the rifampicin experiment. **Figure 3-1C** shows representative data used to fit the half-life to the data. Additionally, the half-life can be determined by the following relationship to the decay rate **(Equation 3-2)**.

$$\text{Equation 2: } t_{1/2} = \ln(2)/\lambda$$

Equation 3-2

Figure 3-1D shows the distribution of half-lives calculated for the entire library of sequences. The calculated half-life distribution falls in line with the timeframe seen in natural *E. coli* systems with most natural *E. coli* sequences decaying in less than 5 minutes[113]. We do see the majority of the sequences having short lifetimes with a few having a half-live greater than 20 minutes.

Major Features of Decay

To further ensure our library is reliable and in agreement with past experiments, we interrogated individual sets of the entire library to determine if the trends expected from literature are observed. We also wanted to ensure that the features determined from the 5' UTRs were reasonable and informative. Furthermore, we investigated some of our own hypothesis to have a quantitative understanding of the major features of mRNA decay.

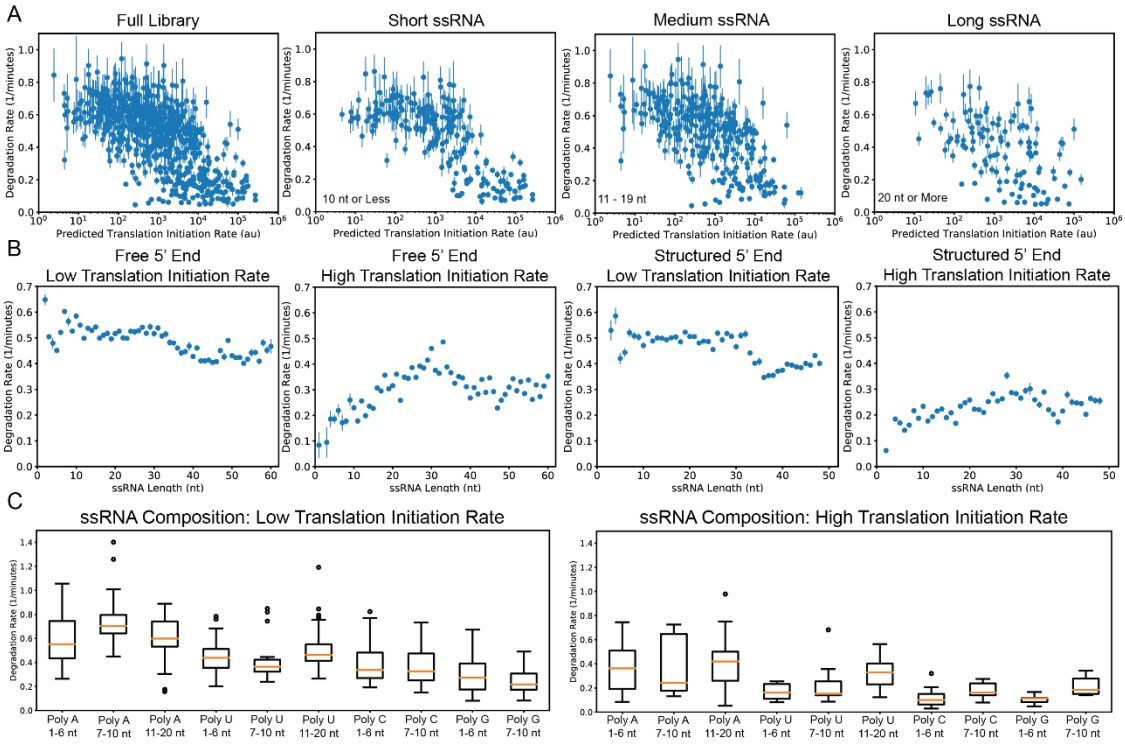

Figure **3-2**: Translation Initiation Rate and ssRNA determinants of RNA Decay (A) RBS Calculator was used to design a series of sequences that have a wide range of translation initation rates. As translation rate increases, the stability of the mRNA increases. The RBS libraries were separately plotted by amount of ssRNA in the 5' UTR in groups of 10 nt or less, 11-19 nt, or 20 nt and greater. As ssRNA increased, the decay rates did not decrease as fast with increasing translation initiation rate. (B) Unstructured RNA at the 5' UTR increases the decay rate. Structure at the very beginning of the 5' UTR provides some protection and reduces the decay rate. However, when the translation initiation rate is less than 5000 au the decay rate of the sequence remains high at all lengths of unstructured RNA. (C) Different lengths of Poly A, U, C, and G sequences were tested to determine the effect of sequence composition. Poly A sequences are the least stable, poly U sequences are moderately stable, and poly C and G sequences are the most stable.

Based on previous finding that show their importance, we first interrogated the impact of the translation initiation rate and regions of unstructured RNA in the 5' UTR on the overall stability of the mRNA. The first set consisted of a series of RBS Libraries spanning more than four orders of magnitude of translation initiation rates **(Figure 3-2A)**. Each member of this library had a different amount of unstructured RNA. Using the Vienna RNA package[114], we determined the secondary structure of each member and calculated the total amount of unstructured RNA. We included unpaired nucleotides that were not within secondary structure.

Unpaired nucleotides that formed bulges or loops within hairpins were excluded from this calculation. As expected, we see that increasing the translation initiation rate indeed increases the stability of the mRNA and decreases the decay rate. Furthermore, we took the entire library and split it into three bins of unstructured RNA length of 0-10 nt, 11-19 nt, and 20 nt or more. While the trend of increasing stability with increasing translation initiation rate holds in all categories, the decay rates do not decrease as quickly with increasing translation initiation rate when the amount of unstructured RNA increases.

To further understand the relationship between the decay rate and the translation initiation rate and unstructured RNA, we interrogated two additional libraries that systematically varied the amount unstructured RNA **(Figure 3-2B)**. In this set all combinations of A, U, C, and G were used. Each construct can be seen in **Appendix B-4**. The first library contained sequences that increase the amount of unstructured RNA at the extreme 5' end. In this library, all sequences had an unstructured 5' end. The structure of the sequences was again calculated and the unstructured RNA outside of secondary structures was determined. The sequences were then binned as having either low translation initiation rate (<5000 TIR) or high translation initiation rate (>5000 TIR). We found that when translation initiation rate is low, the amount of unstructured RNA has little impact on the overall decay rate. In this situation, the decay rate is already so high due to the poor protection of translating ribosomes that additional regions of unstructured RNA have little impact of the decay rate. However, when the translation initiation rate is high, the decay rate increases with increasing length of unstructured mRNA. Past 30 nucleotides, the decay rate plateaus and slightly decreases due to tertiary structure formation as we previously described. The second set is much like the first, except that a terminal hairpin was placed at the extreme 5' end of the mRNA. Like the set without a hairpin at the 5' end, the decay rate was both high and constant with respect to the length of unstructured mRNA. Additionally, with increased translation initiation rate, the decay rate increased with increasing amounts of

unstructured RNA. However, the decay rate did not increase as much with increasing unstructured RNA and plateaued at a lower decay rate. This suggests, that the hairpin disrupted the 5'-end dependent decay pathway for RNase E in *E. coli*.

While unstructured RNA is a strong indicator of the mRNA decay rate, we wanted to understand the impact of the nucleotide composition on the overall decay rate. We designed libraries with increasing amounts of A, U, C, and Gs at the 5' end **(Figure 2C)**. Poly A and U sequence were systematically varied from 1 to 20 nucleotides, while poly C and G sequences were varied from 1 to 10 nucleotides due to the constrains of oligopool synthesis. We again binned the sequences with the same low and high TIR constraints. At low translation initiation rate, as before, the length of unstructured RNA had a small impact on the decay rate. Likewise, at high translation initiation rate, increasing the length of unstructured RNA increased the decay rate. When looking at the contributions of the different nucleotides, poly A sequences showed the highest decay rates, poly U sequences showed medium decay rates, and poly C and G sequences showed the lowest decay rates.

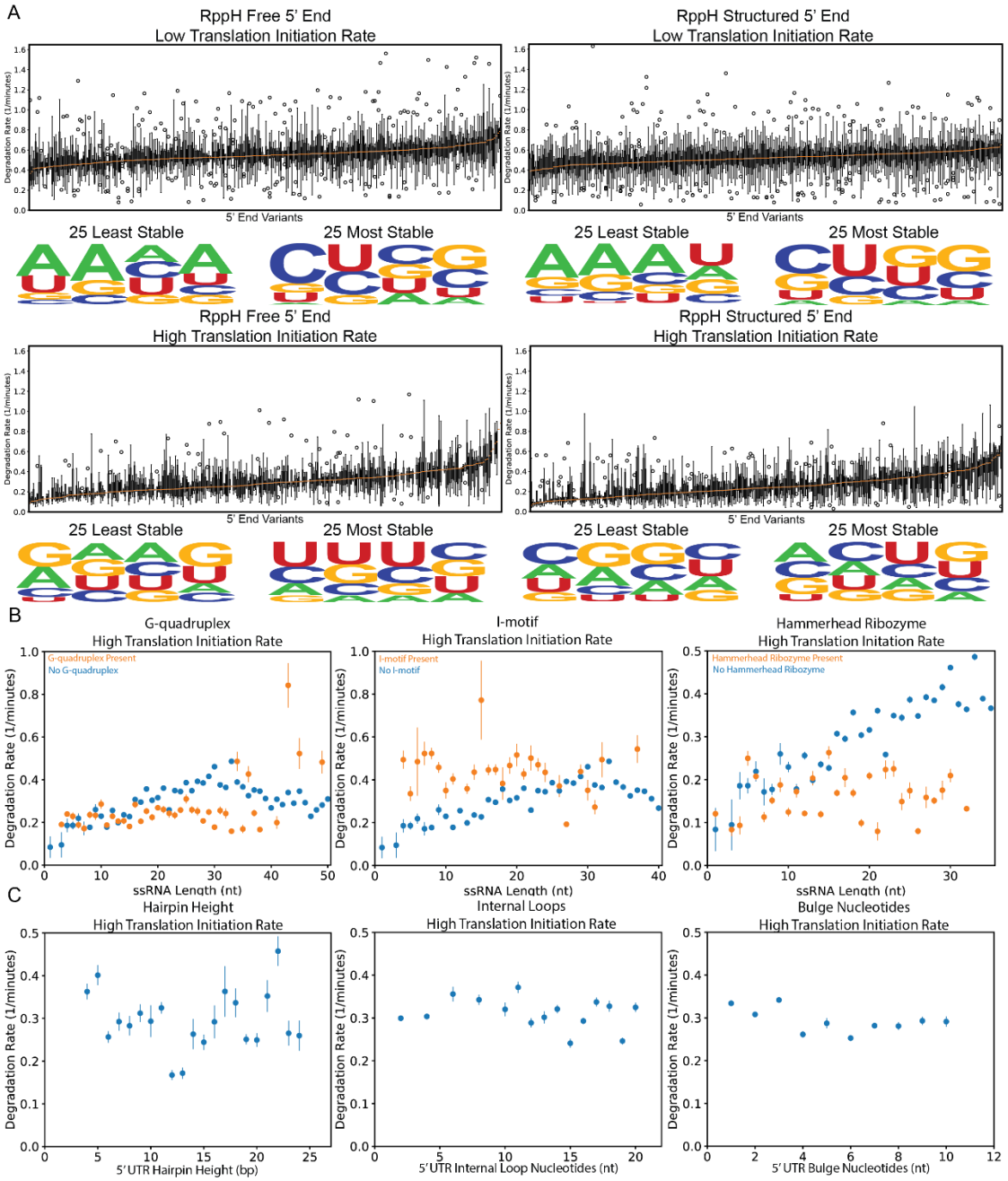

Figure **3-3**: RppH, Tertiary Structure, and dsRNA determinants of RNA Decay (A) Every combination of the first four nucleotides was tested to determine the impact of sequence on RppH binding. The combinations were tested where the first four nucleotides were unstructured or incorporated in a hairpin. Furthermore, the sequences were tested at low and high translation initiation rate. A sequence logo for the 25 least and most stable sequences was created. In all test groups, purines were the least stable, while pyrimidines were the most stable. (B) G-quadruplex, I-motif, and Hammerhead Ribozyme sequences were compared against a library without these structures with the same amount of unstructured RNA. G-quadruplexes and Hammerhead Ribozymes protected the sequences, while I-motif sequences did not provide protection. (C)

Different aspects of secondary structure were tested. Hairpin height and the number of nucleotides in internal loops or bulges in hairpin structures did not have a significant effect on the decay rate.

We additionally wanted to further understand the importance of sequences at the 5' End. Previous in vitro data demonstrated that the composition of the extreme 5' end nucleotides impacted the binding of enzyme RppH and the conversion of the 5' end from a triphosphate to a mono phosphate. We tested all 256 combinations of the first four nucleotides in the 5' UTR **(Figure 3-3A)**. Additionally, we tested the 256 combinations of the first four nucleotides in the 5' UTR within different libraries with sequence contexts. For example, we tested the sequences with both an unstructured 5' end and an end where the first four nucleotides perfectly complemented sequences used to create a hairpin to ensure incorporation into a secondary structure. We plotted the sequences with either low or high translation rate and then took the 25 least and 25 most stable sequence combinations. We created a sequence logo to represent the relative frequency of the different nucleotides represented. Overall, regardless of structure or translation initiation rate, unstable sequences were highly enriched with purines (A and G), while highly stable sequences were enriched with pyrimidines (U and C). Additionally, it appeared that only the first three nucleotides impacted the decay rate. The fourth position in the sequence had little impact and all the different nucleotides were represented equally.

To complete our investigation on the major contributors of decay, we wanted to interrogate the impact of secondary and tertiary structures on the overall decay rate. We looked at complex structures such as G-quadruplexes, I-motifs, and hammerhead ribozymes **(Figure 3-3B)**. We compared these sequences to sets with the same translation initiation rate and unstructured RNA length. With regard to G-quadruplexes, their addition to the 5' UTR increased the stability of the mRNA. However, I-motif sequences, which are less likely to form in physiological conditions in *E. coli* had higher decay rates and offered no protection to the 5' UTR. Lastly, the

addition of hammerhead ribozymes greatly reduced the decay rate of the mRNA sequence and decoupled the decay rate from the amount of unstructured RNA.

Furthermore, we investigated additional features of secondary structure to understand their impact on the decay rate **(Figure 3-3C)**. We created a library with increasing hairpin height. However, we did not see any correlation between the number of consecutive nucleotides in the hairpin and the overall decay rate. Likewise, we looked at both the number of nucleotides in internal loops within a hairpin and the number of nucleotides within a bulge within a hairpin. Here the impact of unstructured RNA with a secondary structure was not predictive of the overall decay rate of the mRNA.

Model Construction and Performance

Using this core understanding of the features of mRNA decay, we wanted to expand the usefulness of our dataset from the identification of the key sequence and structural predictors of decay to a comprehensive model of RNA decay. Additionally, we wanted a model that could use additional features that might have a small, but additive effect on the overall decay rate. To accomplish this goal, we employed the gradient boosting machine learning algorithm LightGBM to create a predictive model of RNA decay.

We began by analyzing each designed 5' UTR and calculating 357 biophysical features of each sequence to create a feature vector to train the LightGBM models. All calculated features and their respective values are available in **Appendix B-5**. Overall, our features fall into the three categories: ribosome coverage, sequence, and structure. For ribosome coverage, we used RBS Calculator v2.1 to calculate the individual delta G terms of the model and the overall translation initiation rate. We also calculated the mean headway distance or the predicted average distance between translating ribosomes on our mRNA construct. To encode sequence related information

into the features, we first identified and used the first four nucleotides at the 5'-end of the mRNA. Additionally, we calculated the A, U, C, and G composition fraction of the entire 5' UTR and the 5' most unstructured region of the 5' UTR, along with the GC composition of the 5' UTR.

Lastly, RNA secondary and tertiary structure provided a rich resource to identify features that could impact the decay rate. To convert structure information into features, we calculated both summary variables of the various structure components present in the 5' UTR along with position dependent vectors that return a value on a per-nucleotide basis. For the summary structure variables, we calculated whether the beginning of the transcript at the 5' end is incorporated into a secondary structure and determined the total number and fraction of nucleotides unpaired in the 5' UTR. To further categorize the unstructured RNA we calculated the total amount and single longest consecutive length of different categories of unstructured RNA. These categories of unstructured RNA are the unstructured RNA outside a secondary structure, unstructured RNA within a structure such as an internal loop or bulge, and unstructured RNA that forms a loop at the terminus of a hairpin structure. Furthermore, we calculated the total number of base paired nucleotides and the percentage of the nucleotides base paired along with the max length of a hairpin. We also looked at additional structures that could form. We determined whether a G-quadruplex, i-motif RNA structure, CsrA sequence and structure, or Hammerhead Ribozyme was present.

We also created a series of position dependent variables that were calculated on a per-nucleotide basis up to 80 nucleotides. The first set identified whether each nucleotide was either base paired, unpaired, or did not exist since the sequence was less than 80 nucleotides. The next set added to this information by classifying the nucleotides in the same manner as the first set, however, returning the number of consecutive nucleotides present in the structure. The purpose of this set was to not only identify structure, but quantify the size of the structures and their position. An additional set was added that quantifies the number of nucleotides present between two paired

nucleotides. The number returned for the interaction is one divided by the distance between the nucleotides. Lastly, the sequence of the 5' UTR was split into its individual A, U, C, or G and the nucleotide at each position was recorded. If there was no nucleotide present, due to the 5' UTR being shorter than 80 nucleotides, it was marked with an x.

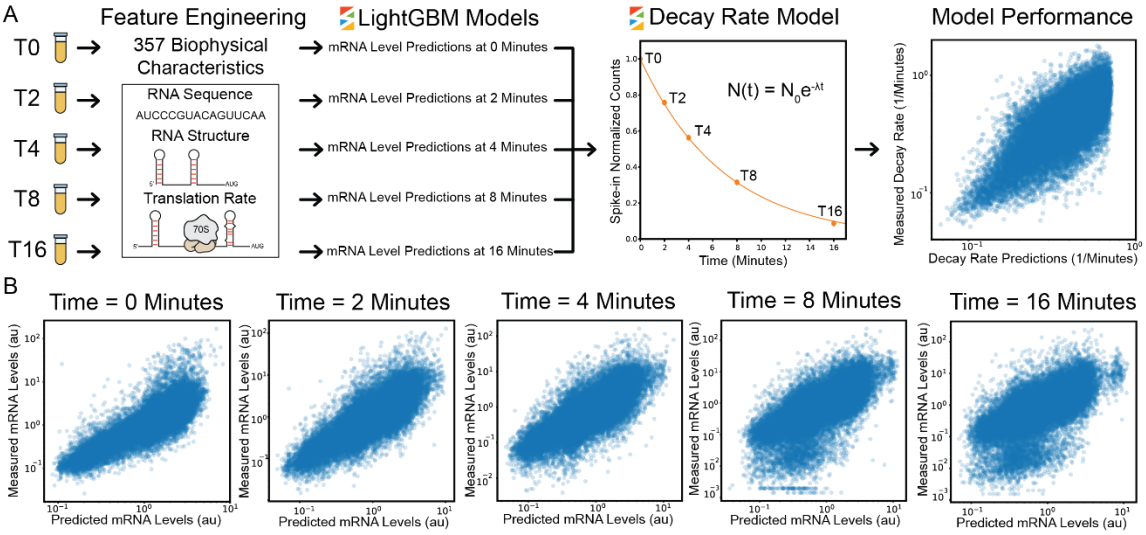

Figure **3-4**: Model Training and Performance (A) The RNA and DNA sequencing results at each timepoint were used to train individual models of RNA level at each timepoint. The gradient boosting algorithm LightGBM was used to build each model along with 357 calculated features covering the sequence, structure, and translation initiation rate of each library member. The output of these five models was used in conjunction with the feature data to build a model of the decay rate. The fitted time course data was used to calculated the degradation rate of each library member to determine model performance. The model performance on the combined train and test data is plotted. (B) The individual models of RNA level are plotted at each timepoint.

This robust set of features was used to train models of the RNA level (RNA/DNA) at each of the measured timepoint. The data set was first filtered to remove RNA decay rates that were predicted with low or moderate confidence. Specifically, if the five timepoints that were used to fit the curve to determine the decay rate did not fit the curve with at least an R^2 of 0.95 they were removed from the set. Additionally, any sequence that registered five or fewer DNA read counts after the Novaseq run was removed. Since the DNA levels are not dependent on the decay rate, any sequence that contained that few DNA reads was not well represented in the overall culture and would have insufficient data to accurately calculate the decay rates. After

filtering the data, 47435 sequences remained out of the original dataset. The data was split into an 80% train set (37435 sequences) and a 20% test set (9486 sequences) to build and determine the performance of the model. The gradient boosting, machine learning algorithm, LightGBM, was chosen to train the model. To avoid overfitting the model, two forms of cross validation were used when building the model with the train set. Only a subset of the training data was used in each successive round of model training. Likewise, only a subset of the features were used in each successive round of model training. These parameters are listed in **Appendix B-6** for each model. The performance of the train and test set were visualized with a learning curve showing the model error at each successive round of training. The train and test performance can be seen in **Appendix B-1**. The models performed well and the overall predictions of the combined dataset can be seen in **Figure 3-4B** and the model performance is listed in **Appendix B-6**. As expected, the models performed the best at the early timepoints before many of the sequences had degraded, while by 16 minutes less of the RNA remained as is more susceptible to noise.

All of the predicted RNA levels at each timepoint were fed into a single additional LightGBM model with the entire feature vector to predict the decay rate for each construct. The decay rates experimentally determined earlier from fitting the half-life equation were used to compare the predictive accuracy of the model. As in the RNA timepoint models, the same constructs were used for the training and test set. Again, cross validation was used in the train set to avoid overfitting by using only a subset of the library members and feature vectors at each round of model training. The learning curves and error distribution can be seen in **Appendix B-2**. Overall, the model performed on the combined data with an R^2 of 0.63 (**Figure 4-4A**) and the error distribution is within two-fold.

Discussion

Overall, we were able to identify the core sequence and structural determinants of mRNA decay in *E. coli* and translate that information into a predictive model of mRNA decay. This model has broad usefulness as a tool to better understand overall gene expression of a synthetic construct of interest. For metabolic engineering applications, it would be useful to have a greater understanding of the turnover rate of the RNA to help balance overall expression. Likewise, in designing genetic circuits, having a clear understanding of the half-life of a gene in a complex genetic circuit to help explain circuit performance and switching speed between states. Lastly, the model has general applicability to help identify potentially problematic synthetic designs that cause low stability before construction. The model will be hosted for use on the Denovo DNA website (https://www.denovodna.com).

To better understand the physical meaning and relative contribution of the different sequence and structural features, we did a feature analysis to rank the relative importance of each of the features used in the model. The complete listing of feature importance can be seen in **Appendix B-3**. As expected, the output of the RNA timepoint models is the most important feature for the prediction of the final decay rates. Beyond the timepoint data predictions, calculations related to the translation initiation rate were very predictive of the overall decay rate. Both output from the RBS Calculation and the mean headway distance calculation were critical predictors of mRNA decay in our model. Likewise with regard to the sequence composition, the A, U, C, and G composition was highly predictive of the overall decay rate. With regard to structure, the total amount of unpaired RNA outside of a secondary structure in the 5' UTR along with the length of the longest consecutive region of unpaired RNA outside of a secondary structure were the most predictive. Additionally, the sequence composition of the first four nucleotides in the 5' UTR was important in determining the overall decay rate. Overall, the model

valued the predictive capability of the translation initiation rate, the nucleotide composition of the 5' UTR, unpaired RNA outside of structures than within structures, and the contribution of the nucleotides at the beginning of the 5' UTR that we suggest is related to RppH. However, this does not mean that the other features do not add to the predictive capability of the model, but they instead show the additive effect of these features.

Methods

Library Design and Cloning

A library of 62,120 unique 5' UTR sequences was designed to explore a wide sequence and structure design space. Each 5' UTR sequence was assigned a unique 15 nt non-repetitive barcode. The designs were split up into five equal sized oligopools that were ordered from GenScript. Additionally, primer binding sites, cloning restriction sites, and padding was added to each design for future cloning and to make all the sequences the same 170 nt in length for improved synthesis. All sequences are provided in **Appendix B-4.** The vector was a pFTV1 derivative containing a ColE1 origin of replication and Chloramphenicol resistance (Cm^R). Each design contained a J23100 promoter positioned in front of the variable 5' UTR. Following the 5'UTR was superfolder GFP (sfGFP) with a barcode at the end of the gene. Downstream, an insulating hairpin was used along with a moderate strength ribosome binding site 9973 au in front of mRFP1. In all constructs, every part of the design was constant except for the 5' UTR and the barcode at the end of the coding region. The vector sequence data can be found in **Appendix B-4**.

The entire pool of variants was constructed by cloning each oligopool into the vector and pooling the resultant transformants for characterization. The cloning was done in two parts, the first being inserting the oligopool and the second being inserting the sfGFP coding region

between the 5’ UTRs and barcodes. For the first round of cloning, each oligopool was PCR amplified for a total of 22 cycles and was gel purified using a 3% agarose gel. Afterwards, the amplified oligopools flanked by 5’ AvrII and 3’ AatII sites along with the vector containing these sites was digested for 37 C for 6 hours. The digested vector was gel purified using a 1 % agarose gel. T4 ligase was then used to perform two ligations for each digested oligopool to the vector. 300 fmoles of digested oligopool and 30 fmoles of digested vector were used in the ligation reaction. Additionally, two controls were performed where the ligation reaction did not contain any digested oligopool, but in one contained T4 ligase, while the other did not. The goal of these controls was to determine the amount of background in the cloning. Using DH5α cells from NEB, the ligated plasmids were transformed. Using serial dilutions and plating, each oligopool contained at least one million transformants. The cells were grown and then plasmid extracted using the plasmid DNA mini kit (Omega Biotek).

For the second round of cloning, a gblock containing sfGFP and flanked by EcoRI and restriction sites was PCR amplified for 30 cycles and then gel purified using a 1.5% agarose gel. It was then digested with NEB EcoRI-HF and SacI-HF for 6 hours at 37 C. Each pool of plasmids containing the oligopool inserted into the vector was also digested with EcoRI-HF and SacI-HF (NEB) in addition to shrimp alkaline phosphatase (rSAP) to prevent re-ligation of the vector. Both the digested sfGFP and digested pool of plasmids containing the oligopool were gel purified using a 1.5 % agarose gel. T4 ligase was then used to perform two ligations for each oligopool containing vector. 90 fmoles of digested sfGFP and 25 fmoles of oligopooling vector were ligated. Additionally, two controls were performed where the ligation reaction did not contain any digested sfGFP, but in one contained T4 ligase, while the other did not. The goal of these controls was to determine the amount of background in the cloning. Using DH5α cells (NEB), the ligated plasmids were transformed. Using serial dilutions and plating, each oligopool contained at least ten million transformants. The transformants for each of the five oligopools were mixed and

added to 50 mL of LB CM50 and were grown for 16 hours. The cell material was pelleted and used to make cryostocks and do plasmid purification to send for a MiSeq run to ensure oligopool was cloned in properly.

DNA and RNA Characterization

For DNA and RNA extractions, one entire cryostock was resuspended in 100 mL of LB media supplemented with 50 ug/mL Chloramphenicol (Cm) in a 1 L Erlenmeyer flask and incubated at 37 C, 300 RPM agitation for 12 hours. Afterwards, it was diluted to an OD_{600} of 0.05 in 50 mL EZRich media (Teknova) supplemented with 50 ug/mL Chloramphenicol (Cm) in a 1 L Erlenmeyer flask and incubated at 37 C, 300 RPM agitation for 2 hours until the OD_{600} reached 0.3. Again, the culture was diluted to an OD600 of 0.05 in 250 mL EZRich media supplemented with 50 ug/mL Chloramphenicol (Cm) in a 1 L Erlenmeyer flask and incubated at 37 C, 300 RPM agitation for 1.5 hours until the OD_{600} reached 0.25 and the cells were harvested. All characterization steps were done in triplicate. First, three 50 mL aliquots were taken and used to perform a plasmid extraction using the plasmid DNA mini kit (Omega Biotek). Additionally, three 4 mL samples of the culture were taken and mixed with 8 mL of RNAprotect Bacteria Reagent (Qiagen) to fix the cells and stop RNA degradation. Next, Rifampicin was added to the remaining bacterial culture to a final concentration of 500 ng/uL to stop transcription. Three 4 mL samples were taken at 2, 4, 8, and 16 minutes post Rifampicin addition and mixed with 8 mL RNAprotect Bacteria Reagent (Qiagen). Fixed cultures were then centrifuge pelleted at 5000 x g for 10 minutes and RNA extracted using the Total RNA Purification kit (Norgen Biotek), followed by non-specific degradation of contaminant DNA using the Turbo DNAse kit (Ambion). Additionally, an RNA spike-in control was produced using a HiScribe T7 High Yield RNA Synthesis kit (NEB) using a gblock (Integrated DNA Technologies) template designed with a T7

promoter and sequence the same design as the experimental samples, but includes a unique intergenic region. Both the purified RNA samples and T7 RNA spike-in control were measured with the dye-based Quant-iT RNA Assay kit (Thermo Fisher). Samples were measured in a 96-well microtiter plate with an excitation/emission maxima 644/673 nm in a Spark spectrophotometer (TECAN). Based on the concentrations, spike-in control RNA was added at the ratio of 10 attomoles of control to 1 ug of total RNA. Each sample was then rRNA depleted using the NEBNext rRNA depletion kit with (NEB). The depleted RNA samples were then reverse transcribed to cDNA using the SuperScript IV First-Strand Synthesis System (Invitrogen). A specific reverse primer flanking the barcode region of the mRNA transcript was used to specifically reverse transcribe the experimental mRNAs of interest. Following cDNA production, both the cDNA and plasmid samples were PCR amplified for 25 cycles using primers flanking the barcode region. The PCR product was gel purified using a 3.0% agarose gel. Samples were then sent to Genewiz to perform NovaSeq (Illumina) next generation sequencing.

Modeling Methods

LightGBM models were trained for each RNA timepoint. Hyperparameters were optimized for each model and listed in **Appendix B-6**. Additionally, a LightGBM model was trained for using the output of these models combined with the full feature vector. The hyperparameters used in the model are listed in **Appendix B-6**. The complete feature vector for all constructs used in the model is listed in **Appendix B-5**.

Chapter 4

Constructing Nonrepetitive Promoters Libraries for *E. coli* and *C. autoethanogenum*

The forward progress of synthetic biology has led to the increasingly ambitious construction of genetic circuits and metabolic pathways of greater size and complexity. However, for many of these large-scale projects, this requires the expression of many genes with multiple promoters, ribosome-binding sites, and terminators. For many organisms, there does not exist a sufficient number of these parts to build these systems. As a result, a standard small set of parts are frequently used repeatedly when building genetic systems. This has led to many of the engineered DNA sequences and engineered organisms having regions of high similarity. Unfortunately, repeated use of identical parts or regions of DNA in a genetic construct greatly increases the risk of failure. In most unmodified organisms, homologous recombination is a major system of DNA repair. When potentially catastrophic double stranded DNA breaks occur, homologous recombination rejoins the fragments at regions of similarity. In natural systems, this works well as it repairs the strand and returns it back to the original function. However, in engineered systems with multiple regions of homology, homologous recombination can cause deletions or other changes to the sequence since there are multiple regions that can be joined together due to the homology.

Previous work has determined the degree and size of homology required to be affected by homologous recombination. The longest repetitive sequence is a key determinant of a genetic system's stability, called the maximum shared repeat length (L_{max}). A 21-bp repetitive sequence is sufficient to trigger homologous recombination in *E. coli*, which can excise the DNA in between the two parts and break the system function[115]. In other organisms such as *Saccharomyces*

cerevisiae and Bacillus subtilis, a 12- to 18-bp repeat is long enough for strand invasion and homologous recombination[116-117].

In this work, we chose to focus on improving the number of promoters available in both *E. coli* and *C. autoethanogeum* in conjunction with LanzaTech, a commercial leader in engineering bacteria to produce useful chemicals from waste gas feedstocks[118]. In this work we looked to solve two problems with the promoters previously available. First, we wanted to develop a set of nonrepetitive promoters that could all be used in conjunction without causing homologous recombination. Second, we wanted to develop a library of promoters with a wide range of transcription initiation rates. This would allow greater tuning of expression for the particular application.

E. coli promoter library

To accomplish this goal, we developed the Nonrepetitive Parts Calculator to rapidly generate thousands of highly nonrepetitive genetic parts from specified design constraints. We applied the Nonrepetitive Parts Calculator to design 4,350 highly nonrepetitive σ^{70} promoters, followed by experimentally characterizing their transcription rates in *E. coli*. We first formulated a reference design constraint for σ^{70} promoters that includes a core promoter sequence for constitutive transcription and a downstream sgRNA binding site for $dCas9_{SP}$-mediated transcriptional regulation **(Figure 4-1A).** The core promoter sequence consists of a 20-bp upstream region, a consensus −35 hexamer, a 17-bp spacer, a consensus −10 hexamer and a 6-bp discriminator, whereas the sgRNA binding site contains a non-template protospacer adjacent motif (PAM) (CCN) and a 20-bp sgRNA target site. The sgRNA binding site also serves as the initial transcribed region for these promoters. We then added model-based constraints to prevent the formation of cryptic promoters, improve DNA synthesis and generate promoter sequences

with genome-matched GC content. Finally, we specified an L_{max} of 10 bp to ensure that all promoters could be used in the same genetic system without introducing genetic instability or DNA synthesis failures **(Figure 4-1B).**

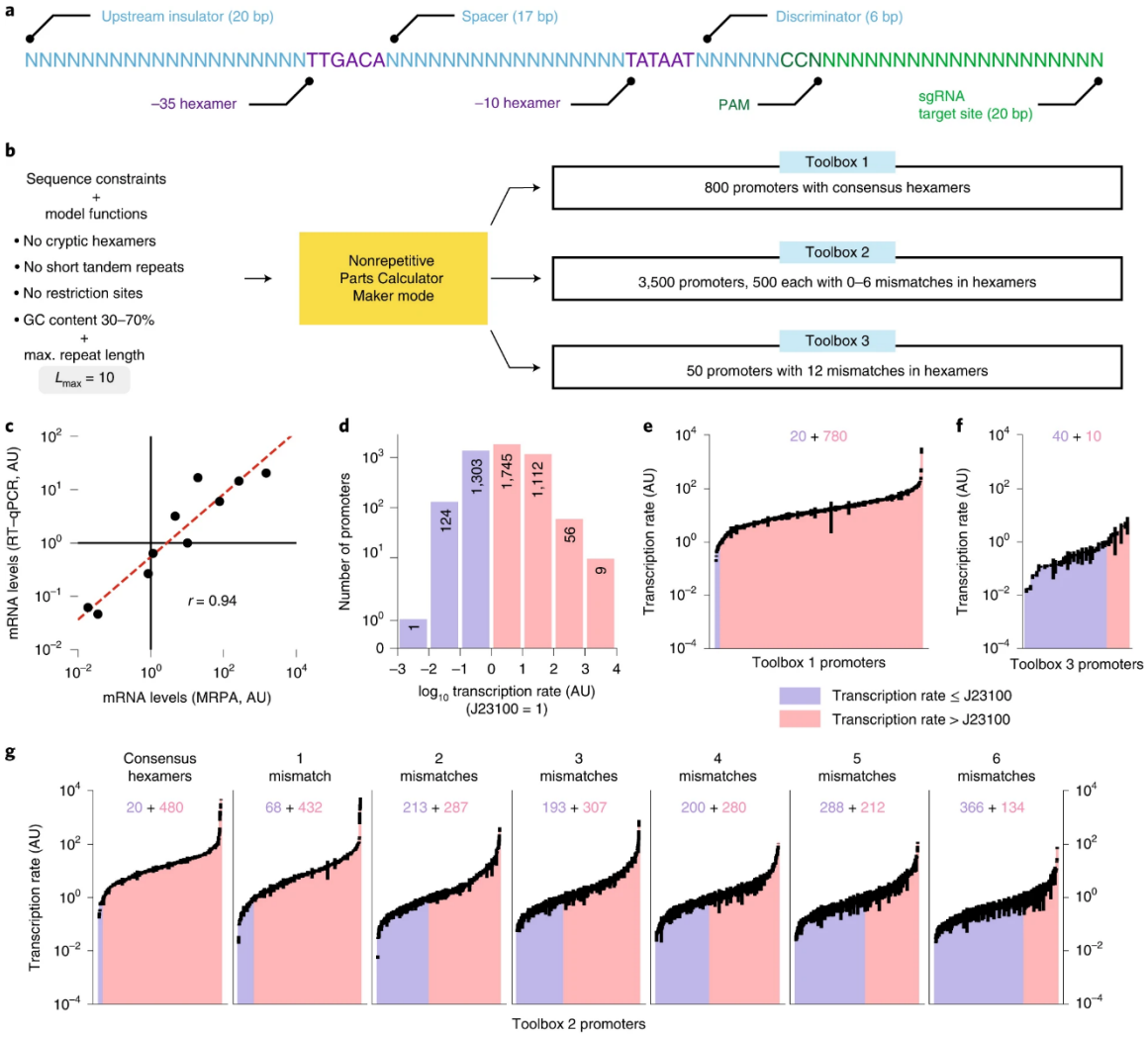

Figure **4-1**: (A) The core sequence constraint for the promoter design. (B) The specifications of the three sets of promoters designed. (C) Comparison between a subset of the designed promoters with regard to their mRNA levels measured with RT-qPCR and massively parallel reporter assay. (D) A histogram describing the distribution of promoters with transcription initiation rates above and below J23100. (E) The complete distribution of transcription initiation rates of the first toolbox of promoters. (F) The complete distribution of transcription initiation rates of the third toolbox of promoters. (G) The complete distribution of transcription initiation rates of the second toolbox of promoters.

We then constructed and characterized the 4,350 highly nonrepetitive promoters, combining barcoding, oligopool synthesis, library-based cloning and NGS to measure their transcription rates in *E. coli*. Overall, all promoter variants were covered by at least 21 DNA-seq and 40 RNA-seq counts with high reproducibility. We determined the promoters' transcription rates by taking the ratio between their RNA-seq and DNA-seq read counts and dividing by the J23100 promoter's transcription rate for normalization. These massively parallel measurements were then independently confirmed by selecting 10 nonrepetitive promoters, separately cloning them into the same expression vector, including their sgRNA-binding sites, and measuring their mRNA levels using RT-qPCR (**Figure 4-1C**).

The 4,350 highly nonrepetitive bacterial promoters have an exceptionally high breadth and depth of transcription rates that vary across an 820,000-fold range (**Figure 4-1D**); 2,922 promoters are stronger than the common J23100 promoter, with up to a 4,731-fold higher transcription rate. The weakest promoter has a 173-fold lower transcription rate than J23100. On average, there are 621 nonrepetitive promoters within each 10-fold increment in transcription rate. As expected, the first toolbox has the highest number of strong promoters whereas the third toolbox has predominantly weak promoters (**Figure 4-1EF**). As designed, promoters in the second toolbox have systematically varied transcription rates, depending on the number of mismatches in their hexamer sequences (**Figure 4-1G**). Across all three toolboxes, 3,550 promoters may be used simultaneously with an L_{max} of 10 bp. When using all 4,350 promoters simultaneously, the L_{max} is 19 bp.

C. autoethanogenum promoter library

Working with LanzaTech, we designed 30,909 promoters for *Clostridium autoethanogenum*, an industrial bacterial species that can grow using the industrial feedstock of

syngas. Due to the relative niche applications of *C. autoethanogenum* very few promoters were designed previously for use in the organism, reducing the scale of genetic engineering that could be done in the organism. Using the nonrepetitive parts calculator, eight promoter toolboxes were designed with the constrains listed **(Table 4-1)**. With this toolbox, over 1000 genes can be simultaneously expressed in *Clostridia* with tunable transcriptional control across a 1,000,000-fold range **(Figure 4-2)**. Overall, this non-repetitive toolbox of promoters enables a breadth of metabolic engineering applications in an important, non-model industrial organism.

Non-Repetitive Promoter Toolbox Specifications								
Toolbox	Max Repeat L_{MAX} bp	GC% range	UPS region	-35 hexamer consensus	spacer region	-10 hexamer consensus	CRISPRi PAM site	orthogonal sgRNA binding sites
A	12	0, 33.5	20 N	TTGACA	17 N	TATAAT	NCC	20 N
B	15	0, 33.5	20 N	TTGACA	17 N	TATAAT	NCC	20 N
C	12	0, 50.8	20 N	TTGACC	17 N	TATAAT	NCC	20 N
D	15	0, 50.8	20 N	TTGACC	17 N	TATAAT	NCC	20 N
E	12 to 32	0, 20.9	20 N	WWWWWW	17 W	WWWWWW	NCC	20 N
F	12	0, 33	20 N	TATG	22 N	TAACTT	NCC	20 N
K	17	0, 35	20 N	TTGACA	17 N	TATAAT	NCC	20 N
M	15	0, 100	20 N	TTGACA	11 to 23 N	TATAAT	NCC	20 N

Table **4-1**: Constrains used to design the promoter toolboxes.

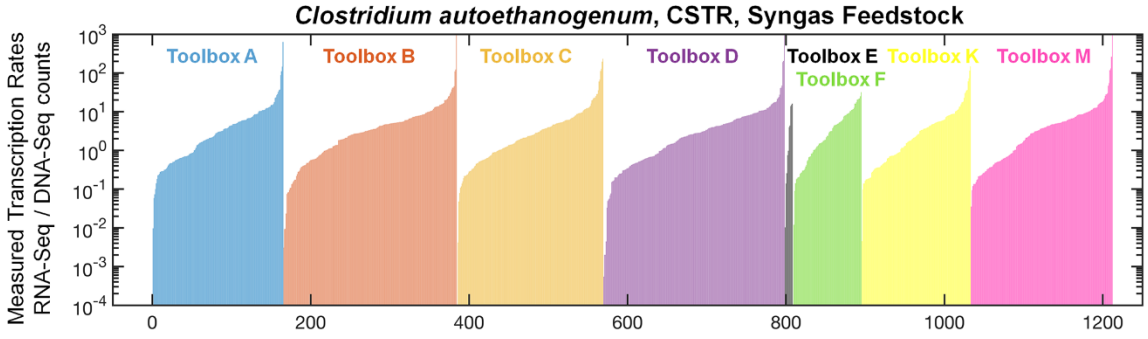

Figure **4-2**: Library performance in *Clostridium autoethanogenum* using Syngas as a feedstock

Methods

DNA-seq and RNA-seq

For the *E. coli* promoters, total RNA from 5 ml of the 50-ml cultures with an OD_{600} of 0.2 were extracted using Norgen Biotek Total RNA Purification Kit. Extracted total RNA was treated with Invitrogen TURBO DNA-free Kit. RNA integrity and DNA removal were verified by running an analytical agarose gel. Total RNA was then concentrated to 10 µl using the Qiagen RNeasy MinElute Cleanup Kit. Total RNA samples (5 µg) were depleted of ribosomal RNA using Illumina Ribo-Zero Removal Kit and cleaned up using the Qiagen RNeasy MinElute Cleanup Kit. Complementary DNA was synthesized from the rRNA-depleted samples using Invitrogen SuperScript IV First-Strand Synthesis System. The cDNA was treated with RNase H to degrade all RNA hybridized to the cDNA. Plasmid DNA from the same 50-ml cultures was harvested using QIAprep Spin Miniprep Kit following the manufacturer's instructions. Promoters from the extracted plasmids were PCR amplified and the barcodes from cDNA were PCR amplified. All amplicon lengths were verified on an analytical agarose gel. Two DNA and RNA biological replicates (from separate cultures) were sequenced on an Illumina HiSeq (Quintara Biosciences). Complete library coverage and precision measurements were ensured, obtaining at least 4 million transformants per biological replicate, culturing cells for at least 6 h in exponential growth conditions, and acquiring at least 64 million DNA-seq reads and at least 95 million RNA-seq reads per biological replicate. Each promoter was covered by at least 21 DNA-seq and 40 RNA-seq read counts with median counts of 5,861 and 3,804, with an overall mapping efficiency of 89% and 96%, respectively. Read count reproducibility was extremely high across biological duplicates (Pearson's $r = 0.99$ and 0.98, respectively).

Isogenic bacterial promoter RT–qPCR measurements

Selected nonrepetitive promoter sequences, including their CRISPR interference sgRNA-binding sites, were separately cloned into a flexible test vector pFTV expressing mRFP1 reporter protein. *E. coli* NEB-5α cells containing these sequence-verified plasmids were cultured in 5 ml minimal M9 medium supplemented with 0.4% glucose, 0.5 mM leucine and antibiotic, maintained in exponential growth conditions for at least 6 h and then harvested. Total RNA was extracted using a Total RNA Purification kit (Norgen Biotek), followed by Turbo DNAse treatment (Ambion). The cDNA was prepared using a High Capacity cDNA Reverse Transcription kit (Applied Biosystems). The qPCR was carried out using an ABI Step One Plus real-time thermocycler (Applied Biosystems) with customized Taqman probes that bind the mRFP1 coding sequence and the 16S rRNA as an endogenous control. Three biological replicates (separate cultures) and three technical replicates per biological replicate were performed.

Chapter 5

Conclusions and Future Work

Conclusions

This body of work set out to create enabling technology to help improve the ability to design genetic systems. Building biophysical models of the fundamental processes of the central dogma of biology offers a clear path to a greater understanding of the underlying biology, while also allowing for efficient biological system engineering. As the field of synthetic biology moves forward, constructing complex genetic circuits and metabolic pathways will require more sophisticated models. The size of the design space increases exponentially with the number of genes involved in the system. This large design space exceeds the capacity of using random design and mass screening, necessitating models of gene expression.

To aid in the improvement of models of gene expression, we explored the fundamental processes of RNA decay and transcription initiation. In chapter 2, we leveraged the fundamental biochemical knowledge of RNA decay pathways to build small libraries that tested individual features that modulate the mRNA decay rate. We relied on past scientific work that identified, but did not quantify, the key pathways of RNA turnover. We designed these sequence libraries to explore the limits of these pathways. We used RT-qPCR to quantify the changes in RNA level caused by these sequence and secondary structure modifications. As opposed to the intergenic region and 3' UTR, we found that the 5' UTR had the greatest impact on the mRNA transcript in determining the stability of the mRNA. Likewise, we learned that transcription initiation, secondary structure, and the RppH binding site in the 5' UTR had the largest quantitative impact and determined the decay rate.

Using the findings of Chapter 2, in Chapter 3 we looked to expand on our understanding of the 5' UTR by building a 62,120 member library to investigate with higher precision how the sequence, secondary structure, and translation initiation rate impacts the stability of the mRNA. We used a next generation sequencing (NGS) method to simultaneously measure the decay rate of the entire library. Using a Rifampicin-based, time-course method, we directly calculated the decay rate of the mRNA sequences. We next calculated 357 features of each library member and used them to build a model of mRNA decay. The model first used the features to calculated the steady state RNA levels at each timepoint and then, in a subsequent model, predicted the final decay rate. The model well predicted mRNA decay rate, which no model had done previously.

In chapter 4, we used a similar next generation sequencing (NGS) method to explore transcription. Previously, the number of promoters available for use in one system limited design. The high sequence similarity between many of the existing promoters makes them incompatible due to the deletions and other mutations caused when the highly similar or identical promoters trigger homologous recombination. Likewise, the previously existing promoters did not contain a broad range of transcription initiation rates, limiting the ability to properly tune large systems. We successfully created a library of 4,350 E. coli promoters and 30,909 C. autoethanogeneum promoters for simultaneous use in a large genetic system. Furthermore, they have a broad range of transcription initiation rates allowing for tight control of transcription. In addition to providing a useful tool to synthetic biologists, these promoters demonstrate the exciting opportunity to develop additional nonrepetitive parts that can avoid the problem of homologous recombination in genetic systems.

Lastly, the experimental workflows of chapter 3 and 4 demonstrate the power of massively parallel reporter assays to test a large number of sequences simultaneously. This approach allows for faster and more comprehensive testing of many design ideas simultaneously, which will push the field of Synthetic Biology forward.

Future Work

This work helped improve the usability of RNA decay and transcription initiation rate in genetic system design. However, numerous opportunities exist to build on this work and continue to improve the predictive capability of the models. In particular, additional work could not only improve prediction of gene expression in *E. coli*, but also expand to other organisms to make it easier to engineer other organisms.

With respect to the model of mRNA decay in *E. coli*, additional experiments could improve the predictive capability of the model. Pairing the library of 5' UTRs, described in chapter 4, with promoters that have a broad range of transcription initiation rates could help determine the interaction between transcription level and the RNA decay rate. Possibly, the high levels of mRNA produced by strong promoters could saturate one or more of the mRNA decay pathways. Likewise, while all of the work completed in chapters 2, 3, and 4 occurred while the cells grew in exponential growth phase, an additional set of experiments could explore the decay rate with the cells in different growth states. This would help improve understanding of how the decay rate changes in response to a decrease in the cellular growth rate. In a similar vein, the relative contribution of features of the designed sequences in determining the decay rate could change under different cell stress conditions such as cold, heat, or nutrient deficiency. While overall transcription and RNA levels would fall during many types of cell stress, the particular sequences attacked could change depending on the type of stress.

An additional set of experiments could help improve the accuracy of the mRNA decay models with respect to RNA binding proteins present in *E. coli*. Knock-out or knock-down experiments with RNA binding proteins such as CsrA and Hfq could identify the sequences most likely to interact with each of the proteins and have the most impact on the decay rate. RNA binding proteins have the potential to modulate the translation rate, modify the secondary

structure of the mRNA, and physically block attack by RNases. Understanding the relationship between the position, sequence, and secondary structure of regions of the mRNA that the RNA proteins bind could allow for tight control over the decay rate along with improving the accuracy of the translation initiation rate.

Many of the enzymes responsible for RNA decay in *E. coli* have homologs in other prokaryotic and eukaryotic species. However, we would not expect the activity of these enzymes to remain constant in all of the different organisms[119-121]. Therefore, we could use the same approach taken to measure decay rates in *E. coli*, but instead look at the decay rates in a wide set of organisms. Understanding the differences in decay rate in a Gram-positive organism such as *Bacillus subtilis,* or in a similar to *E. coli* Gram-negative organism such as *Pseudomonas fluorescens,* would help determine if one model or a set of models would suffice for mRNA decay in bacteria. Furthermore, a eukaryotic organism such as *Saccharomyces cerevisiae* could test the changes in the decay rate between prokaryotic and eukaryotic organisms, which likely have large differences due to the large change in physiology.

The work described in chapter 4 regarding transcription initiation rate and promoters could expand in a similar way as the RNA decay work to encompasses more cell states and species. All the promoters developed interacted with the σ^{70} transcription factor, which works in conjunction with RNA polymerase to produce an RNA transcript. However, in *E. coli* alone, seven sigma factors exist that control the transcription of different promoters[122]. These sigma factors act during different cell states, particularly in response to various cell stresses. Creating nonrepetitive promoters that work for each of the different sigma factors could unlock opportunities for engineering genetic circuits and pathways that are active during cell stress. Furthermore, while we successfully made nonrepetitive promoters in *E. coli* and *C. autoethanogeneum*, creating and testing promoters for additional industrially useful strains would expand the tools available for working in those organisms.

Appendix A

Supplementary Information for Chapter 2

The supplementary material listed here is from the publication "Systematic Quantification of Sequence and Structural Determinants Controlling mRNA stability in Bacterial Operons." It is also available at ACS Synthetic Biology using the link (https://pubs.acs.org/doi/full/10.1021/acssynbio.0c00471).

Supplementary Tables

Table **A-1**: This table contains a subset of the complete dataset available through ACS Synthetic Biology. This subset contains the sequence name, dataset name, and relative RNA level determined from qPCR.

Sequence Name	Dataset	Reporter ID 1	Relative mRNA Level	Reporter ID 2	Relative mRNA Level
8	mRFP1 Expression Library	mRFP1	1.000	N/A	N/A
7	mRFP1 Expression Library	mRFP1	0.691	N/A	N/A
6	mRFP1 Expression Library	mRFP1	0.458	N/A	N/A

5	mRFP1 Expression Library	mRFP1	0.325	N/A	N/A
4	mRFP1 Expression Library	mRFP1	0.179	N/A	N/A
3	mRFP1 Expression Library	mRFP1	0.151	N/A	N/A
2	mRFP1 Expression Library	mRFP1	0.124	N/A	N/A
1	mRFP1 Expression Library	mRFP1	0.085	N/A	N/A
10	sfGFP Expression Library	sfGFP	1.000	N/A	N/A
9	sfGFP Expression Library	sfGFP	0.735	N/A	N/A
8	sfGFP Expression Library	sfGFP	0.414	N/A	N/A
7	sfGFP Expression Library	sfGFP	0.399	N/A	N/A
6	sfGFP Expression Library	sfGFP	0.229	N/A	N/A
5	sfGFP Expression Library	sfGFP	0.147	N/A	N/A
4	sfGFP Expression Library	sfGFP	0.152	N/A	N/A

3	sfGFP Expression Library	sfGFP	0.182	N/A	N/A
2	sfGFP Expression Library	sfGFP	0.127	N/A	N/A
1	sfGFP Expression Library	sfGFP	0.155	N/A	N/A
RNase III Control	RNase Polarity	mRFP1 (opt)	1.000	GFPmut3 (opt)	1.000
5' UTR RNase III	RNase Polarity	mRFP1 (opt)	1.306	GFPmut3 (opt)	1.036
Intergenic RNase III	RNase Polarity	mRFP1 (opt)	0.886	GFPmut3 (opt)	0.566
3' UTR RNase III	RNase Polarity	mRFP1 (opt)	1.065	GFPmut3 (opt)	0.909
RNase E Control	RNase Polarity	mRFP1 (opt)	1.000	GFPmut3 (opt)	1.000
5' UTR RNase E	RNase Polarity	mRFP1 (opt)	0.093	GFPmut3 (opt)	0.152
Intergenic RNase E	RNase Polarity	mRFP1 (opt)	0.577	GFPmut3 (opt)	0.508
3' UTR RNase E	RNase Polarity	mRFP1 (opt)	1.185	GFPmut3 (opt)	1.069
0nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.397	N/A	N/A

1nt_5p	5' Variable Length Poly A	mRFP1 (opt)	1.000	N/A	N/A
2nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.995	N/A	N/A
4nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.848	N/A	N/A
5nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.550	N/A	N/A
6nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.344	N/A	N/A
8nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.261	N/A	N/A
10nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.336	N/A	N/A
12nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.311	N/A	N/A
15nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.146	N/A	N/A
16nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.135	N/A	N/A
20nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.106	N/A	N/A
25nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.145	N/A	N/A

30nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.158	N/A	N/A
35nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.164	N/A	N/A
40nt_5p	5' Variable Length Poly A	mRFP1 (opt)	0.254	N/A	N/A
0nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.000	GFPmut3 (opt)	1.000
1nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.192	GFPmut3 (opt)	1.055
2nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.406	GFPmut3 (opt)	1.155
4nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.151	GFPmut3 (opt)	1.075
5nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.045	GFPmut3 (opt)	1.009
6nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.097	GFPmut3 (opt)	0.918
8nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.091	GFPmut3 (opt)	1.087
10nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	1.217	GFPmut3 (opt)	1.155
12nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.839	GFPmut3 (opt)	0.984

15nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.872	GFPmut3 (opt)	0.828
16nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.898	GFPmut3 (opt)	0.942
20nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.705	GFPmut3 (opt)	0.791
24nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.622	GFPmut3 (opt)	0.708
30nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.747	GFPmut3 (opt)	0.822
35nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.613	GFPmut3 (opt)	0.760
40nt_I	Intergenic Variable Length Poly A	mRFP1 (opt)	0.601	GFPmut3 (opt)	0.700
RppH_AAA	RppH Set	mRFP1 (opt)	0.055	N/A	N/A
RppH_AAC	RppH Set	mRFP1 (opt)	0.524	N/A	N/A
RppH_AAG	RppH Set	mRFP1 (opt)	0.561	N/A	N/A
RppH_AAT	RppH Set	mRFP1 (opt)	0.409	N/A	N/A
RppH_ACA	RppH Set	mRFP1 (opt)	0.317	N/A	N/A

RppH_ACC	RppH Set	mRFP1 (opt)	0.613	N/A	N/A
RppH_ACG	RppH Set	mRFP1 (opt)	0.634	N/A	N/A
RppH_ACT	RppH Set	mRFP1 (opt)	0.397	N/A	N/A
RppH_AGA	RppH Set	mRFP1 (opt)	0.716	N/A	N/A
RppH_AGC	RppH Set	mRFP1 (opt)	0.755	N/A	N/A
RppH_AGG	RppH Set	mRFP1 (opt)	0.748	N/A	N/A
RppH_AGT	RppH Set	mRFP1 (opt)	1.000	N/A	N/A
RppH_ATA	RppH Set	mRFP1 (opt)	0.825	N/A	N/A
RppH_ATC	RppH Set	mRFP1 (opt)	0.604	N/A	N/A
RppH_ATG	RppH Set	mRFP1 (opt)	0.716	N/A	N/A
RppH_ATT	RppH Set	mRFP1 (opt)	0.446	N/A	N/A
Term 1 (99.67 % Eff)	Terminator Efficiency	mRFP1	1.000	N/A	N/A

Term 2 (96.03 % Eff)	Terminator Efficiency	mRFP1	1.274	N/A	N/A
Term 3 (84.79 % Eff)	Terminator Efficiency	mRFP1	0.940	N/A	N/A
Term 4 (66.68 % Eff)	Terminator Efficiency	mRFP1	1.735	N/A	N/A
Term 5 (49.75 % Eff)	Terminator Efficiency	mRFP1	1.079	N/A	N/A
Term 6 (31.38 % Eff)	Terminator Efficiency	mRFP1	1.005	N/A	N/A
Term 7 (8.28 % Eff)	Terminator Efficiency	mRFP1	1.278	N/A	N/A
Term 8 (No Term)	Terminator Efficiency	mRFP1	1.029	N/A	N/A

Table **A-2**: This table contains the probe and primer sequences along with the amplification efficiency used to validate the use of custom TaqMan probes and primers and SYBR Green primers.

TaqMan Efficiency					
Target	Probe Sequence	Forward Primer	Reverse Primer	Amplicon Length (nt)	Primer Efficiency
mRFP1	ACCTTCCATACGAACTTT	ACGTTATCAAAGAGTTCATGCGTTTC	CGATTTCGAACTCGTGACCGTTAA	72	97.4%
codon optimized mRFP1	ACCGCCTTTGGTAACTT	GTACCCAGACCGCTAAACTCA	TCCCAAGCGAACGGTAAGG	58	104.6%
sfGFP	AACCGGCAGTTTACC	AAGGTGATGCCACGAATGGTAAA	CCATAGGTCAGGGTGGTAACCA	100	92.2%
codon optimized GFPmut3	ACGGGCACAAGTTTAG	CAATTCTGGTAGAATTAGACGGTGATGT	GCCCTCTCCGCTGACG	62	96.5%
16S	CTAGGCGACGATCCCT	CCCAGATGGGATTAGCTAGTAGGT	TGGCTGGTCATCCTCTCAGA	81	93.8%
SYBR Efficiency					
RFP-End	N/A	GCTTACAAAACCGACATCAAACTGG	GACCTTCAGCACGTTCGTACTG	85	99.24%
Post-Terminator	N/A	GCGGTATCAGCTCACTCAAAGG	GCCTTTTTACGGTTCCTGGC	113	102.70%
16S	N/A	CTCTTGCCATCGGATGTGC	GTGGCTGGTCATCCTCTCA	100	96.90%

Table **A-3**: This table contains the MIQE Data for RT-qPCR Reproducibility

MIQE	
Experimental Design	
Definition of experimental and control groups	See Supplementary Data 1 for listing of experimental and control groups
Number within each group	See Supplementary Data 1 for listing of biological and technical replicates
Assay carried out by the core or investigator's laboratory?	Assay performed by investigator's laboratory
Acknowledgment of authors' contributions	See Contributions Statement
Sample	
Description	See growth conditions
Volume/mass of sample processed	Peleted 5 mL Culture
Microdissection or macrodissection	Not Applicable to bacterial work
Processing procedure	Not Applicable to bacterial work
If frozen, how and how quickly?	Sample was not frozen
If fixed, with what and how quickly?	No fixing was performed
Sample storage conditions and duration (especially for FFPEb samples)	Sample was immediately processed
Nucleic Acid Extraction	
Procedure and/or instrumentation	Norgen Biotek Total RNA Purification Protocol

Name of kit and details of any modifications	Norgen Biotek Total RNA Purification Kit Cat. # 17200
Source of additional reagents used	Lysozyme, from chicken egg white, Sigma-Aldrich, Cat. # L6876-5G; B-mercaptoethanal, G-Biosciences, Cat. # BC98
Details of DNase or RNase treatment	Dnase Treatment: Turbo DNA-free kit, Cat. # AM1907; RNase Decontamination: RNase AWAY, VWR, 17810-494
Contamination assessment (DNA or RNA)	1 % Agarose gel was used to visualize post-Dnase treatement to check for DNA contamination
Nucleic acid quantification	Spectrophotometry
Instrument and method	NanoDrop
RNA integrity: method/instrument	1 % Agarose gel was used to visualize 23S and 16S rRNA to check for RNA degradation
Reverse Transcription - Complete Reaction Conditions	
Complete reaction conditions	1x RT Buffer, 1x RT Random Primers, 4 mM dNTP mix, MultiScribe Reverse Transcriptase 2.5 U/uL
Amount of RNA and reaction volume	5 ng/uL Extracted RNA, 12.5 ng/uL Yeast tRNA, 30 uL reaction volume

Priming oligonucleotide (if using GSP) and concentration	10X RT Random Primers
Reverse transcriptase and concentration	MultiScribe Reverse Transcriptase, Stock 50 U/uL, Working Concentration 2.5 U/uL
Temperature and time	37 C for 120 minutes
Manufacturer of reagents and catalogue numbers	Applied Biosystems High-Capacity cDNA Reverse Transcription Kit, Cat. # 4368814
Storage conditions of cDNA	cDNA stored at -20 C
qPCR target information	
Gene symbol	See Supplementary Data 2 for listing of genes targeted
Sequence accession number	Not Applicable
Location of amplicon	See Supplementary Data 2
Amplicon length	See Supplementary Data 2
In silico specificity screen (BLAST, and so on)	Not Applicable
Pseudogenes, retropseudogenes, or other homologs?	None
Location of each primer by exon or intron (if applicable)	All primers are located in exons
What splice variants are targeted?	There are no splice varients of the gene
qPCR oligonucleotides	

Primer sequences	See Supplementary Data 2
RTPrimerDB identification number	None
Probe sequences	See Supplementary Data 2
Location and identity of any modifications	None
Manufacturer of oligonucleotides	ThermoFisher Scientific
Purification method	Standard Desalting
qPCR Protocol-Complete Reaction Conditions	
Reaction volume and amount of cDNA/DNA	3.5 uL cDNA, 20 uL total reaction volume
Primer, (probe), Mg2+, and dNTP concentrations	Proprietary Applied Biosystems and QuantaBio Information
Polymerase identity and concentration	AccuStart™ Taq DNA Polymerase, concentration proprietary
Buffer/kit identity and manufacturer	PerfeCTa® qPCR SuperMixes, QuantaBio Cat# 101414-128
Exact chemical composition of the buffer	Proprietary QuantaBio Information
Additives (SYBR Green I, DMSO, and so forth)	Passive Reference Dye: ROX
Manufacturer of plates/tubes and catalog number	ThermoGrid™ PCR Plates, Thomas Scientific, cat# 1158U20
Complete thermocycling parameters	95 C 10 min, 40 cycles of 95 C for 15 sec and 60 C for 1 min
Reaction setup (manual/robotic)	Manual

Manufacturer of qPCR instrument	Applied Biosystems StepOne Plus Real-Time PCR System
qPCR validation	
Evidence of optimization (from gradients)	
Specificity (gel, sequence, melt, or digest)	Primers were designed specifically for their gene target
For SYBR Green I, Cq of the NTC	Not Applicable
Calibration curves with slope and y intercept	See Supplementary Data 2
PCR efficiency calculated from slope	See Supplementary Data 2
r2 of calibration curve	See Supplementary Data 2
Linear dynamic range	Dilutions spanned five orders of magnitude
Cq variation at LOD	N/A
Evidence for LOD	N/A
If multiplex, efficiency and LOD of each assay	Not Muliplexed
Data Analysis	
qPCR analysis program (source, version)	Life Technologies Corporation StepOne Software version 2.3
Method of Cq determination	Auto Threshold
Outlier identification and disposition	None
Justification of number and choice of reference genes	16S rRNA was used as a reference since its expression is consistent in log-phase E. coli

Description of normalization method	Each sample was normalized to the endogenous 16S rRNA reference, followed by normalization to the no-hairpin control
Number and concordance of biological replicates	See Supplementary Data 1 for number of biological replicates
Number and stage (reverse transcription or qPCR) of technical replicates	See Supplementary Data 1 for number of technical replicates
Repeatability (intraassay variation)	See Supplementary Data for standard deviation between biological replicates
Statistical methods for results significance	Standard Deviation
Software (source, version)	Excel 2016

Appendix B

Supplementary Information for Chapter 3

Contained within this appendix is the supplementary data for "Comprehensive modeling and design of 5' UTRs for RNA stability using a 62,000 unique 5' UTR Library." Supplementary Figure 1 shows the train and test performance of the five RNA level models. Supplementary Figure 2 shows the train and test performance of the mRNA decay rate model. Supplementary Figure 3 shows the relative importance of the features used to build the mRNA decay rate model.

Supplementary Figures

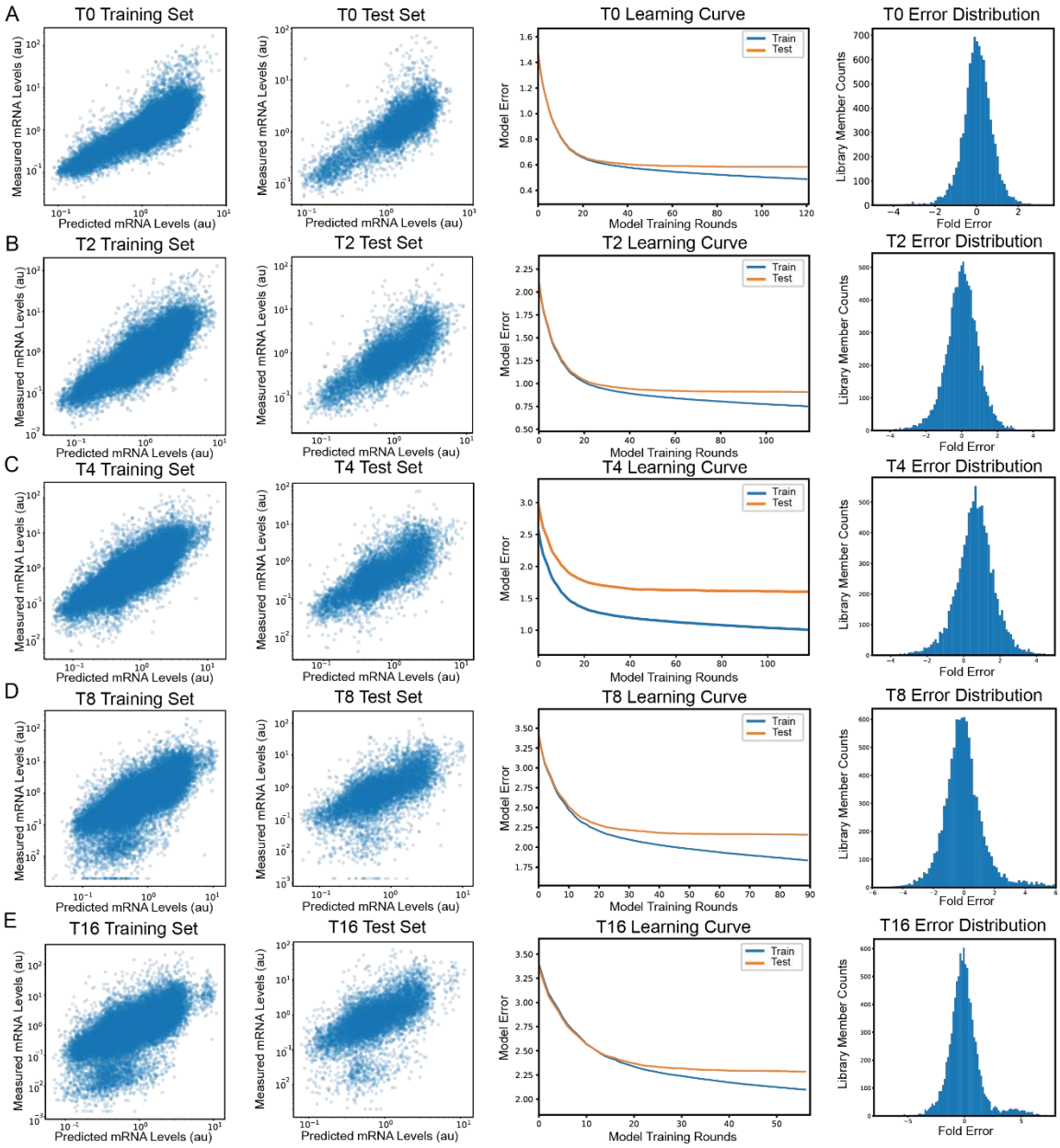

Figure **B-1**: Train and test performance of each RNA level model at each timepoint. Each row A-E corresponds to timepoint T0-T16. The first column shows the LightGBM model predictions vs the actual values in the training set. The second column shows the LightGBM model predictions vs the actual values in the test set. The third column shows the learning curve performance of both the train (blue) and test (orange) set. The fourth column shows the error distribution.

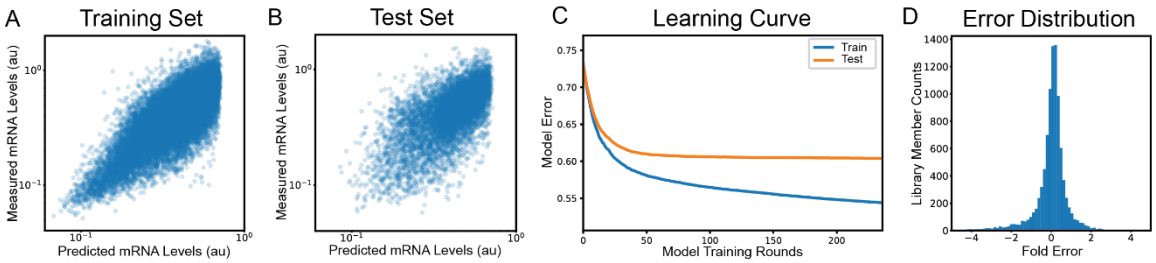

Figure **B-2**: Train and test performance of the RNA decay rate model. (A) The LightGBM model predictions vs the actual values in the training set. (B) The LightGBM model predictions vs the actual values in the test set. (C) The learning curve performance of both the train (blue) and test (orange) set. (D) The error distribution of the model.

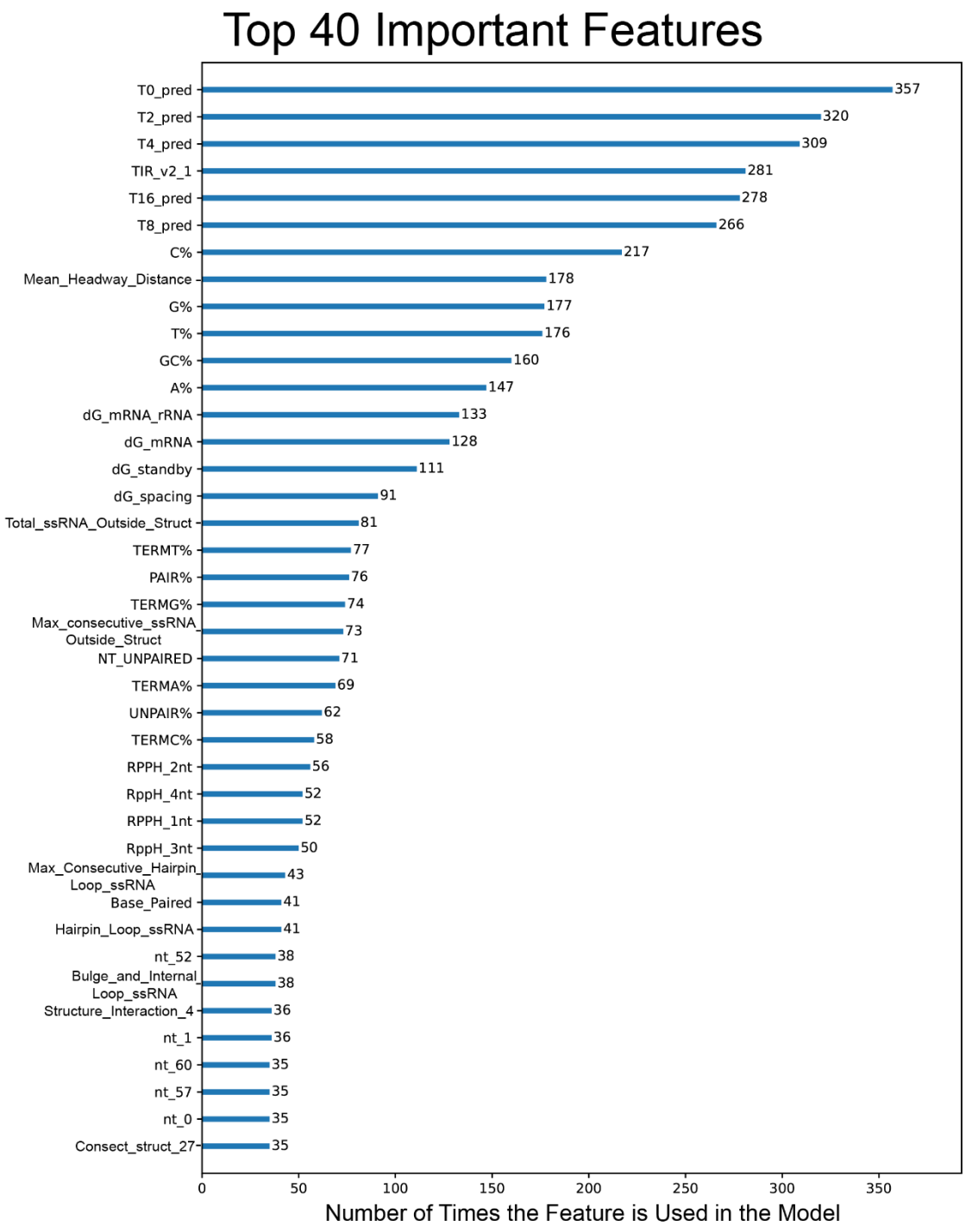

Figure **B-3**: Ranked performance of the top 40 features used in the mRNA decay rate model.

Supplementary Tables

Table **B-4**: This table contains a subset 10 example sequences out of the complete library of 62,100 sequences showing the construction of each sequence.

Index	Designed 5' UTR Sequence	15 nt Barcode	Complete 170 nt Oligo
1	AAAAAAAAAAAAA AAAAAAACCCGCC ATATGGCGGGAAA AACCCCGGCAAAG CAGAGCAGCCAGC C	GTTGTGTT TAGGATG	CTCACTCCATCCTAGGTACAGTGCT AGCAAAAAAAAAAAAAAAAAAAAC CCGCCATATGGCGGGAAAAACCCCG GCAAAGCAGAGCAGCCAGCCATGG AGCTCATACAGTATTGCGGATATAG AATTCGTTGTGTTTAGGATGGACGT CAAATCCCGTCACATACACCAG
2	AAAAAAAAAAAAA AAAAAAACCCGCC ATATGGCGGGAGG CGGACCGAGGGAG CAGCACGACGGGA C	TTCGCCTC ACACGGA	CTCACTCCATCCTAGGTACAGTGCT AGCAAAAAAAAAAAAAAAAAAAAC CCGCCATATGGCGGGAGGCGGACCG AGGGAGCAGCACGACGGGACATGG AGCTCATACAGTATTGCGGATATAG AATTCTTCGCCTCACACGGAGACGT CAAATCCCGTCACATACACCAG
3	AAAAAAAAAAAAA AAAAAAACCCGCC ATATGGCGGGGGG GACCCAGCAAAGA AGGGGGGGAACAA C	AGGCACA CCCGTCGT	CTCACTCCATCCTAGGTACAGTGCT AGCAAAAAAAAAAAAAAAAAAAAC CCGCCATATGGCGGGGGGGGACCCAG CAAAGAAGGGGGGGAACAACATGG AGCTCATACAGTATTGCGGATATAG AATTCAGGCACACCCGTCGTGACGT CAAATCCCGTCACATACACCAG
4	AAAAAAAAAAAAA AAAAAAACCCGCC ATATGGCGGGCGG CCCCCAACGAGGA GCCAGGACAAGCA A	AAACATTC CTGTCCC	CTCACTCCATCCTAGGTACAGTGCT AGCAAAAAAAAAAAAAAAAAAAAC CCGCCATATGGCGGGCGGCCCCCAA CGAGGAGCCAGGACAAGCAAATGG AGCTCATACAGTATTGCGGATATAG AATTCAAACATTCCTGTCCCGACGT CAAATCCCGTCACATACACCAG
5	AAAAAAAAAAAAA AAAAAAACCCGCC ATATGGCGGGAAA GAGGACCGCGAAG CGCGGGGACCGGG C	GTTCCTGC ATCCAAT	CTCACTCCATCCTAGGTACAGTGCT AGCAAAAAAAAAAAAAAAAAAAAC CCGCCATATGGCGGGAAAGAGGAC CGCGAAGCGCGGGGACCGGGCATG GAGCTCATACAGTATTGCGGATATA GAATTCGTTCCTGCATCCAATGACG TCAAATCCCGTCACATACACCAG
6	AAAAAAAAAAAAA AAAAAAACCCGCC ATATGGCGGGAGA	GCCCGTGC CCGGTTG	CTCACTCCATCCTAGGTACAGTGCT AGCAAAAAAAAAAAAAAAAAAAAC CCGCCATATGGCGGGAGAGAGGAA

	GAGGAACAAGAAGGGACAAGAAAGGGG		CAAGAAGGGACAAGAAAGGGGATGGAGCTCATACAGTATTGCGGATATAGAATTCGCCCGTGCCCGGTTGGACGTCAAATCCCGTCACATACACCAG
7	AAAAAAAAAAAAAAAAAAAACCCGCCATATGGCGGGGAGAGCACCCCCCACAAAGCAACCGGGGAC	GCTCGCCTGCATGAA	CTCACTCCATCCTAGGTACAGTGCTAGCAAAAAAAAAAAAAAAAAAAACCCGCCATATGGCGGGGAGAGCACCCCCCACAAAGCAACCGGGGACATGGAGCTCATACAGTATTGCGGATATAGAATTCGCTCGCCTGCATGAAGACGTCAAATCCCGTCACATACACCAG
8	AAAAAAAAAAAAAAAAAAAACCCGCCATATGGCGGGCCAGACAGCACAAGCAGCCGACCCCGGC	CGCCTCGACAACAGG	CTCACTCCATCCTAGGTACAGTGCTAGCAAAAAAAAAAAAAAAAAAAACCCGCCATATGGCGGGCCAGACAGCACAAGCACAGCCGACCCCGGCATGGAGCTCATACAGTATTGCGGATATAGAATTCCGCCTCGACAACAGGGACGTCAAATCCCGTCACATACACCAG
9	AAATAAAAAAAAAAAAAAAACCCGCCATATGGCGGGGCGCAGGCAAGAGGAACCACGGCCGGCCAAA	TATCTGGGCGTTAAT	CTCACTCCATCCTAGGTACAGTGCTAGCAAATAAAAAAAAAAAAAAAACCCGCCATATGGCGGGGCGCAGGCAAGAGGAACCACGGCCGGCCAAATGGAGCTCATACAGTATTGCGGATATAGAATTCTATCTGGGCGTTAATGACGTCAAATCCCGTCACATACACCAG
10	AAATAAAAAAAAAAAAAAACCCGCCATATGGCGGGGGGGCCCACACACCGACGGAGGCAACGGGC	CATTTCCATGACTAC	CTCACTCCATCCTAGGTACAGTGCTAGCAAATAAAAAAAAAAAAAAAACCCGCCATATGGCGGGGGGGCCCACACACCGACGGAGGCAACGGGCATGGAGCTCATACAGTATTGCGGATATAGAATTCCATTTCCATGACTACGACGTCAAATCCCGTCACATACACCAG

Table **B-5**: This table contains a subset 4 example sequences out of the complete library of 62,100 sequences demonstrating the features calculated for each sequence.

Calculated Features				
Index	1	2	3	4
TIR_v2_1	668.3455	2173.739	29567.19	3102.329
dG_mRNA	-14.73	-12.81	-13.49	-14.29
dG_start	-2.76	-2.76	-2.76	-2.76
dG_spacing	0.005326	4.032	0	0.005326
dG_standby	0	0	0	0
dG_mRNA_rRNA	-10.7814	-15.5514	-17.9114	-13.6514
Mean_Headway_Distance	90.29068	36.61795	0.0001	23.12839
RPPH_1nt	A	A	A	A
RPPH_2nt	A	A	A	A
RppH_3nt	A	A	A	A
RppH_4nt	A	A	A	A
A%	0.515152	0.454545	0.5	0.484848
T%	0.030303	0.030303	0.030303	0.030303
C%	0.257576	0.212121	0.181818	0.257576
G%	0.19697	0.30303	0.287879	0.227273
GC%	0.454545	0.515152	0.469697	0.484848
TERMLEN	20	20	20	20
TERMA%	1	1	1	1
TERMT%	0	0	0	0
TERMC%	0	0	0	0
TERMG%	0	0	0	0
Unstructured_First_5nt	TRUE	TRUE	TRUE	TRUE
NT_UNPAIRED	44	44	44	44
UNPAIR%	0.666667	0.666667	0.666667	0.666667
Total_ssRNA_Outside_Struct	33	24	24	32
Max_consecutive_ssRNA_Outside_Struct	20	20	20	20
Bulge_and_Internal_Loop_ssRNA	4	8	10	1
Max_Consecutive_bulge_internal_loop	3	2	3	1
Hairpin_Loop_ssRNA	7	12	10	11
Max_Consecutive_Hairpin_Loop_ssRNA	4	8	6	7
Base_Paired_NT	22	22	22	22
PAIR%	0.333333	0.333333	0.333333	0.333333
Max_hairpin_height	6	6	6	6
GQUAD_Present	FALSE	FALSE	FALSE	FALSE
iMOTIF_Present	FALSE	FALSE	FALSE	FALSE
CsrA_Present	FALSE	FALSE	FALSE	FALSE

Ribozyme_Present	FALSE	FALSE	FALSE	FALSE
Position_0	Unpaired	Unpaired	Unpaired	Unpaired
Position_1	Unpaired	Unpaired	Unpaired	Unpaired
Position_2	Unpaired	Unpaired	Unpaired	Unpaired
Position_3	Unpaired	Unpaired	Unpaired	Unpaired
Position_4	Unpaired	Unpaired	Unpaired	Unpaired
Position_5	Unpaired	Unpaired	Unpaired	Unpaired
Position_6	Unpaired	Unpaired	Unpaired	Unpaired
Position_7	Unpaired	Unpaired	Unpaired	Unpaired
Position_8	Unpaired	Unpaired	Unpaired	Unpaired
Position_9	Unpaired	Unpaired	Unpaired	Unpaired
Position_10	Unpaired	Unpaired	Unpaired	Unpaired
Position_11	Unpaired	Unpaired	Unpaired	Unpaired
Position_12	Unpaired	Unpaired	Unpaired	Unpaired
Position_13	Unpaired	Unpaired	Unpaired	Unpaired
Position_14	Unpaired	Unpaired	Unpaired	Unpaired
Position_15	Unpaired	Unpaired	Unpaired	Unpaired
Position_16	Unpaired	Unpaired	Unpaired	Unpaired
Position_17	Unpaired	Unpaired	Unpaired	Unpaired
Position_18	Unpaired	Unpaired	Unpaired	Unpaired
Position_19	Unpaired	Unpaired	Unpaired	Unpaired
Position_20	Paired	Paired	Paired	Paired
Position_21	Paired	Paired	Paired	Paired
Position_22	Paired	Paired	Paired	Paired
Position_23	Paired	Paired	Paired	Paired
Position_24	Paired	Paired	Paired	Paired
Position_25	Paired	Paired	Paired	Paired
Position_26	Unpaired	Unpaired	Unpaired	Unpaired
Position_27	Unpaired	Unpaired	Unpaired	Unpaired
Position_28	Unpaired	Unpaired	Unpaired	Unpaired
Position_29	Unpaired	Unpaired	Unpaired	Unpaired
Position_30	Paired	Paired	Paired	Paired
Position_31	Paired	Paired	Paired	Paired
Position_32	Paired	Paired	Paired	Paired
Position_33	Paired	Paired	Paired	Paired
Position_34	Paired	Paired	Paired	Paired
Position_35	Paired	Paired	Paired	Paired
Position_36	Unpaired	Unpaired	Unpaired	Unpaired
Position_37	Unpaired	Unpaired	Paired	Paired
Position_38	Unpaired	Unpaired	Unpaired	Paired
Position_39	Unpaired	Unpaired	Unpaired	Paired

Position_40	Unpaired	Paired	Unpaired	Paired
Position_41	Unpaired	Unpaired	Paired	Paired
Position_42	Unpaired	Unpaired	Paired	Unpaired
Position_43	Unpaired	Paired	Paired	Unpaired
Position_44	Unpaired	Paired	Unpaired	Unpaired
Position_45	Paired	Paired	Unpaired	Unpaired
Position_46	Paired	Unpaired	Paired	Unpaired
Position_47	Paired	Unpaired	Unpaired	Unpaired
Position_48	Unpaired	Paired	Unpaired	Unpaired
Position_49	Unpaired	Unpaired	Unpaired	Paired
Position_50	Unpaired	Unpaired	Unpaired	Paired
Position_51	Paired	Unpaired	Unpaired	Unpaired
Position_52	Paired	Unpaired	Unpaired	Paired
Position_53	Unpaired	Unpaired	Paired	Paired
Position_54	Unpaired	Unpaired	Unpaired	Paired
Position_55	Unpaired	Unpaired	Unpaired	Unpaired
Position_56	Paired	Unpaired	Paired	Unpaired
Position_57	Paired	Paired	Paired	Unpaired
Position_58	Unpaired	Unpaired	Paired	Unpaired
Position_59	Paired	Unpaired	Unpaired	Unpaired
Position_60	Paired	Paired	Unpaired	Unpaired
Position_61	Paired	Paired	Unpaired	Unpaired
Position_62	Unpaired	Paired	Paired	Unpaired
Position_63	Unpaired	Unpaired	Unpaired	Unpaired
Position_64	Unpaired	Unpaired	Unpaired	Unpaired
Position_65	Unpaired	Paired	Unpaired	Unpaired
Position_66	x	x	x	x
Position_67	x	x	x	x
Position_68	x	x	x	x
Position_69	x	x	x	x
Position_70	x	x	x	x
Position_71	x	x	x	x
Position_72	x	x	x	x
Position_73	x	x	x	x
Position_74	x	x	x	x
Position_75	x	x	x	x
Position_76	x	x	x	x
Position_77	x	x	x	x
Position_78	x	x	x	x
Position_79	x	x	x	x
Consect_struct_0	0	0	0	0

Consect_struct_1	0	0	0	0
Consect_struct_2	0	0	0	0
Consect_struct_3	0	0	0	0
Consect_struct_4	0	0	0	0
Consect_struct_5	0	0	0	0
Consect_struct_6	0	0	0	0
Consect_struct_7	0	0	0	0
Consect_struct_8	0	0	0	0
Consect_struct_9	0	0	0	0
Consect_struct_10	0	0	0	0
Consect_struct_11	0	0	0	0
Consect_struct_12	0	0	0	0
Consect_struct_13	0	0	0	0
Consect_struct_14	0	0	0	0
Consect_struct_15	0	0	0	0
Consect_struct_16	0	0	0	0
Consect_struct_17	0	0	0	0
Consect_struct_18	0	0	0	0
Consect_struct_19	0	0	0	0
Consect_struct_20	6	6	6	6
Consect_struct_21	6	6	6	6
Consect_struct_22	6	6	6	6
Consect_struct_23	6	6	6	6
Consect_struct_24	6	6	6	6
Consect_struct_25	6	6	6	6
Consect_struct_26	0	0	0	0
Consect_struct_27	0	0	0	0
Consect_struct_28	0	0	0	0
Consect_struct_29	0	0	0	0
Consect_struct_30	6	6	6	6
Consect_struct_31	6	6	6	6
Consect_struct_32	6	6	6	6
Consect_struct_33	6	6	6	6
Consect_struct_34	6	6	6	6
Consect_struct_35	6	6	6	6
Consect_struct_36	0	0	0	0
Consect_struct_37	0	0	1	5
Consect_struct_38	0	0	0	5
Consect_struct_39	0	0	0	5
Consect_struct_40	0	1	0	5
Consect_struct_41	0	0	3	5

Consect_struct_42	0	0	3	0
Consect_struct_43	0	3	3	0
Consect_struct_44	0	3	0	0
Consect_struct_45	3	3	0	0
Consect_struct_46	3	0	1	0
Consect_struct_47	3	0	0	0
Consect_struct_48	0	1	0	0
Consect_struct_49	0	0	0	2
Consect_struct_50	0	0	0	2
Consect_struct_51	2	0	0	0
Consect_struct_52	2	0	0	3
Consect_struct_53	0	0	1	3
Consect_struct_54	0	0	0	3
Consect_struct_55	0	0	0	0
Consect_struct_56	2	0	3	0
Consect_struct_57	2	1	3	0
Consect_struct_58	0	0	3	0
Consect_struct_59	3	0	0	0
Consect_struct_60	3	3	0	0
Consect_struct_61	3	3	0	0
Consect_struct_62	0	3	1	0
Consect_struct_63	0	0	0	0
Consect_struct_64	0	0	0	0
Consect_struct_65	0	1	0	0
Consect_struct_66	0	0	0	0
Consect_struct_67	0	0	0	0
Consect_struct_68	0	0	0	0
Consect_struct_69	0	0	0	0
Consect_struct_70	0	0	0	0
Consect_struct_71	0	0	0	0
Consect_struct_72	0	0	0	0
Consect_struct_73	0	0	0	0
Consect_struct_74	0	0	0	0
Consect_struct_75	0	0	0	0
Consect_struct_76	0	0	0	0
Consect_struct_77	0	0	0	0
Consect_struct_78	0	0	0	0
Consect_struct_79	0	0	0	0
Structure_Interaction_0	0	0	0	0
Structure_Interaction_1	0	0	0	0
Structure_Interaction_2	0	0	0	0

Structure_Interaction_3	0	0	0	0
Structure_Interaction_4	0	0	0	0
Structure_Interaction_5	0	0	0	0
Structure_Interaction_6	0	0	0	0
Structure_Interaction_7	0	0	0	0
Structure_Interaction_8	0	0	0	0
Structure_Interaction_9	0	0	0	0
Structure_Interaction_10	0	0	0	0
Structure_Interaction_11	0	0	0	0
Structure_Interaction_12	0	0	0	0
Structure_Interaction_13	0	0	0	0
Structure_Interaction_14	0	0	0	0
Structure_Interaction_15	0	0	0	0
Structure_Interaction_16	0	0	0	0
Structure_Interaction_17	0	0	0	0
Structure_Interaction_18	0	0	0	0
Structure_Interaction_19	0	0	0	0
Structure_Interaction_20	0.066667	0.066667	0.066667	0.066667
Structure_Interaction_21	0.076923	0.076923	0.076923	0.076923
Structure_Interaction_22	0.090909	0.090909	0.090909	0.090909
Structure_Interaction_23	0.111111	0.111111	0.111111	0.111111
Structure_Interaction_24	0.142857	0.142857	0.142857	0.142857
Structure_Interaction_25	0.2	0.2	0.2	0.2
Structure_Interaction_26	0	0	0	0
Structure_Interaction_27	0	0	0	0
Structure_Interaction_28	0	0	0	0
Structure_Interaction_29	0	0	0	0
Structure_Interaction_30	0.2	0.2	0.2	0.2
Structure_Interaction_31	0.142857	0.142857	0.142857	0.142857
Structure_Interaction_32	0.111111	0.111111	0.111111	0.111111
Structure_Interaction_33	0.090909	0.090909	0.090909	0.090909
Structure_Interaction_34	0.076923	0.076923	0.076923	0.076923
Structure_Interaction_35	0.066667	0.066667	0.066667	0.066667
Structure_Interaction_36	0	0	0	0
Structure_Interaction_37	0	0	0.04	0.058824
Structure_Interaction_38	0	0	0	0.066667
Structure_Interaction_39	0	0	0	0.076923
Structure_Interaction_40	0	0.04	0	0.1
Structure_Interaction_41	0	0	0.058824	0.125
Structure_Interaction_42	0	0	0.066667	0
Structure_Interaction_43	0	0.052632	0.076923	0

Structure_Interaction_44	0	0.058824	0	0
Structure_Interaction_45	0.0625	0.066667	0	0
Structure_Interaction_46	0.071429	0	0.142857	0
Structure_Interaction_47	0.083333	0	0	0
Structure_Interaction_48	0	0.111111	0	0
Structure_Interaction_49	0	0	0	0.125
Structure_Interaction_50	0	0	0	0.1
Structure_Interaction_51	0.166667	0	0	0
Structure_Interaction_52	0.25	0	0	0.076923
Structure_Interaction_53	0	0	0.142857	0.066667
Structure_Interaction_54	0	0	0	0.058824
Structure_Interaction_55	0	0	0	0
Structure_Interaction_56	0.25	0	0.076923	0
Structure_Interaction_57	0.166667	0.111111	0.066667	0
Structure_Interaction_58	0	0	0.058824	0
Structure_Interaction_59	0.083333	0	0	0
Structure_Interaction_60	0.071429	0.066667	0	0
Structure_Interaction_61	0.0625	0.058824	0	0
Structure_Interaction_62	0	0.052632	0.04	0
Structure_Interaction_63	0	0	0	0
Structure_Interaction_64	0	0	0	0
Structure_Interaction_65	0	0.04	0	0
Structure_Interaction_66	0	0	0	0
Structure_Interaction_67	0	0	0	0
Structure_Interaction_68	0	0	0	0
Structure_Interaction_69	0	0	0	0
Structure_Interaction_70	0	0	0	0
Structure_Interaction_71	0	0	0	0
Structure_Interaction_72	0	0	0	0
Structure_Interaction_73	0	0	0	0
Structure_Interaction_74	0	0	0	0
Structure_Interaction_75	0	0	0	0
Structure_Interaction_76	0	0	0	0
Structure_Interaction_77	0	0	0	0
Structure_Interaction_78	0	0	0	0
Structure_Interaction_79	0	0	0	0
nt_0	a	a	a	a
nt_1	a	a	a	a
nt_2	a	a	a	a
nt_3	a	a	a	a
nt_4	a	a	a	a

nt_5	a	a	a	a
nt_6	a	a	a	a
nt_7	a	a	a	a
nt_8	a	a	a	a
nt_9	a	a	a	a
nt_10	a	a	a	a
nt_11	a	a	a	a
nt_12	a	a	a	a
nt_13	a	a	a	a
nt_14	a	a	a	a
nt_15	a	a	a	a
nt_16	a	a	a	a
nt_17	a	a	a	a
nt_18	a	a	a	a
nt_19	a	a	a	a
nt_20	c	c	c	c
nt_21	c	c	c	c
nt_22	c	c	c	c
nt_23	g	g	g	g
nt_24	c	c	c	c
nt_25	c	c	c	c
nt_26	a	a	a	a
nt_27	u	u	u	u
nt_28	a	a	a	a
nt_29	u	u	u	u
nt_30	g	g	g	g
nt_31	g	g	g	g
nt_32	c	c	c	c
nt_33	g	g	g	g
nt_34	g	g	g	g
nt_35	g	g	g	g
nt_36	a	a	g	c
nt_37	a	g	g	g
nt_38	a	g	g	g
nt_39	a	c	g	c
nt_40	a	g	a	c
nt_41	c	g	c	c
nt_42	c	a	c	c
nt_43	c	c	c	c
nt_44	c	c	a	a
nt_45	g	g	g	a

nt_46	g	a	c	c
nt_47	c	g	a	g
nt_48	a	g	a	a
nt_49	a	g	a	g
nt_50	a	a	g	g
nt_51	g	g	a	a
nt_52	c	c	a	g
nt_53	a	a	g	c
nt_54	g	g	g	c
nt_55	a	c	g	a
nt_56	g	a	g	g
nt_57	c	c	g	g
nt_58	a	g	g	a
nt_59	g	a	g	c
nt_60	c	c	a	a
nt_61	c	g	a	a
nt_62	a	g	c	g
nt_63	g	g	a	c
nt_64	c	a	a	a
nt_65	c	c	c	a
nt_66	x	x	x	x
nt_67	x	x	x	x
nt_68	x	x	x	x
nt_69	x	x	x	x
nt_70	x	x	x	x
nt_71	x	x	x	x
nt_72	x	x	x	x
nt_73	x	x	x	x
nt_74	x	x	x	x
nt_75	x	x	x	x
nt_76	x	x	x	x
nt_77	x	x	x	x
nt_78	x	x	x	x
nt_79	x	x	x	x
T0_pred	-0.22207	0.317209	1.256764	0.320867
T2_pred	-1.06169	-0.5757	0.903731	-0.70106
T4_pred	-1.9638	-1.30083	0.556079	-1.23217
T8_pred	-2.02192	-1.72642	-0.41986	-1.28041
T16_pred	-1.87001	-1.32733	-0.19897	-1.69223

Table **B-6**: This table contains the hyperparameters used to build the five RNA level models and the RNA decay rate model. The Table also includes the performance of the train and test set for each model.

Hyper Parameters	T0 Timepoint	T2 Timepoint	T4 Timepoint	T8 Timepoint	T16 Timepoint	Decay Rate Model
task	train	train	train	train	train	train
boosting_type	gbdt	gbdt	gbdt	gbdt	gbdt	gbdt
linear_tree	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
force_row_wise	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
force_col_wise	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
objective	l2	l2	l2	l2	l2	poisson
metric	l2	l2	l2	l2	l2	poisson
num_leaves	40	40	40	40	40	40
max_depth	50	50	50	50	50	50
max_bin	1000	1000	1000	1000	1000	1000
n_estimators	121	119	118	90	57	236
learning_rate	0.1	0.1	0.1	0.1	0.1	0.1
verbose	1	1	1	1	1	1
bagging_fraction	0.5	0.5	0.5	0.5	0.5	0.5
bagging_freq	5	5	5	5	5	2

feature_fraction	0.25	0.25	0.25	0.25	0.25	0.2
Model Performance on the Train and Test Set						
	T0 Timepoint	T2 Timepoint	T4 Timepoint	T8 Timepoint	T16 Timepoint	Decay Rate Model
Train R^2	0.70	0.68	0.64	0.50	0.42	0.69
Test R^2	0.65	0.62	0.43	0.41	0.36	0.42

Appendix C

Table of Abbreviations

Table **C-1**: This table contains a description of the abbreviations used.

3’ UTR	Region of an mRNA transcript beyond the coding region
5’ UTR	Region of an mRNA transcript before the start codon
C. autoethanogenum	*Clostridium autoethanogenum* – anaerobic organism capable of using syngas to produce ethanol
Coding region	Region of an mRNA transcript that contains the information for the amino acid sequence of a protein
E. coli	*Escherichia coli* – model Gram-negative bacteria species
GFP	Green Fluorescent Protein
Intergenic region	Region of an mRNA transcript between two protein coding regions
MIQE	Minimum Information for Publication of Quantitative real-time PCR Experiments
mRNA	Messenger RNA - RNA that contains protein coding sequences
NGS	Next Generation Sequencing

RBS	Ribosome Binding Site
RFP	Red Fluorescent Protein
RT-qPCR	Real-Time Quantitative Polymerase Chain Reaction

References

1. Bokinsky, G.; Peralta-Yahya, P. P.; George, A.; Holmes, B. M.; Steen, E. J.; Dietrich, J.; Lee, T. S.; Tullman-Ercek, D.; Voigt, C. A.; Simmons, B. A., Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered Escherichia coli. *Proceedings of the National Academy of Sciences* **2011,** *108* (50), 19949-19954.

2. Celedon, J. M.; Chiang, A.; Yuen, M.; Diaz-Chavez, M. L.; Madilao, L. L.; Finnegan, P. M.; Barbour, E. L.; Bohlmann, J., Heartwood-specific transcriptome and metabolite signatures of tropical sandalwood (Santalum album) reveal the final step of (Z)-santalol fragrance biosynthesis. *The Plant Journal* **2016,** *86* (4), 289-299.

3. Ibraheem, F.; Gaffoor, I.; Tan, Q.; Shyu, C.-R.; Chopra, S., A sorghum MYB transcription factor induces 3-deoxyanthocyanidins and enhances resistance against leaf blights in maize. *Molecules* **2015,** *20* (2), 2388-2404.

4. Lambert, J. M.; Berkenblit, A., Antibody–drug conjugates for cancer treatment. *Annual review of medicine* **2018,** *69*, 191-207.

5. Sandborn, W. J.; Cyrille, M.; Hansen, M. B.; Feagan, B. G.; Loftus Jr, E. V.; Rogler, G.; Vermeire, S.; Cruz, M. L.; Yang, J.; Boedigheimer, M. J., Efficacy and safety of abrilumab in a randomized, placebo-controlled trial for moderate-to-severe ulcerative colitis. *Gastroenterology* **2019,** *156* (4), 946-957. e18.

6. van Dyck, C. H., Anti-amyloid-β monoclonal antibodies for Alzheimer's disease: pitfalls and promise. *Biological psychiatry* **2018,** *83* (4), 311-319.

7. Newman, D. J.; Cragg, G. M., Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. *Journal of Natural Products* **2012,** *75* (3), 311-335.

8. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Müller, R.; Wohlleben, W., antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research* **2015,** *43* (W1), W237-W243.

9. Harbaugh, S. V.; Goodson, M. S.; Dillon, K.; Zabarnick, S.; Kelley-Loughnane, N., Riboswitch-Based Reversible Dual Color Sensor. *ACS synthetic biology* **2017**.

10. Cohen, S. N.; Chang, A. C. Y.; Boyer, H. W.; Helling, R. B., Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences* **1973,** *70* (11), 3240-3244.

11. Jackson, D. A.; Symons, R. H.; Berg, P., Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli. *Proceedings of the National Academy of Sciences* **1972,** *69* (10), 2904-2909.

12. Baltz, R. H., Genetic manipulation of secondary metabolite biosynthesis for improved production in Streptomyces and other actinomycetes. *Journal of industrial microbiology & biotechnology* **2016,** *43* (2-3), 343-370.

13. Brewster, R. C.; Jones, D. L.; Phillips, R., Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. *PLOS Computational Biology* **2012,** *8* (12), e1002811.

14. Espah Borujeni, A.; Channarasappa, A. S.; Salis, H. M., Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic acids research* **2014,** *42* (4), 2646-2659.

15. Boël, G.; Letso, R.; Neely, H.; Price, W. N.; Wong, K.-H.; Su, M.; Luff, J.; Valecha, M.; Everett, J. K.; Acton, T. B., Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **2016,** *529* (7586), 358.

16. Tian, T.; Salis, H. M., A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons. *Nucleic acids research* **2015,** *43* (14), 7137-7151.
17. Kosuri, S.; Goodman, D. B.; Cambray, G.; Mutalik, V. K.; Gao, Y.; Arkin, A. P.; Endy, D.; Church, G. M., Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences* **2013,** *110* (34), 14024-14029.
18. Mackie, G. A., RNase E: at the interface of bacterial RNA processing and decay. *Nature reviews. Microbiology* **2013,** *11* (1), 45.
19. Dods, G.; Gómez-Schiavon, M.; El-Samad, H.; Ng, A. H., Accurate prediction of genetic circuit behavior requires multidimensional characterization of parts. *bioRxiv* **2020**.
20. Brophy, J. A.; Voigt, C. A., Principles of genetic circuit design. *Nature methods* **2014,** *11* (5), 508-520.
21. Taketani, M.; Zhang, J.; Zhang, S.; Triassi, A. J.; Huang, Y.-J.; Griffith, L. G.; Voigt, C. A., Genetic circuit design automation for the gut resident species Bacteroides thetaiotaomicron. *Nature Biotechnology* **2020**, 1-8.
22. Aggarwal, N.; Breedon, A. M. E.; Davis, C. M.; Hwang, I. Y.; Chang, M. W., Engineering probiotics for therapeutic applications: recent examples and translational outlook. *Current Opinion in Biotechnology* **2020,** *65*, 171-179.
23. Cheng, Y.-Y.; Hirning, A. J.; Josić, K. i.; Bennett, M. R., The timing of transcriptional regulation in synthetic gene circuits. *ACS synthetic biology* **2017,** *6* (11), 1996-2002.
24. Kurtz, C. B.; Millet, Y. A.; Puurunen, M. K.; Perreault, M.; Charbonneau, M. R.; Isabella, V. M.; Kotula, J. W.; Antipov, E.; Dagon, Y.; Denney, W. S., An engineered E. coli Nissle improves hyperammonemia and survival in mice and shows dose-dependent exposure in healthy humans. *Science Translational Medicine* **2019,** *11* (475).
25. Xue, X.-c.; Liu, F.; Ou-Yang, Z.-c., A Kinetic Model of Transcription Initiation by RNA Polymerase. *Journal of Molecular Biology* **2008,** *378* (3), 520-529.
26. Tayara, H.; Tahir, M.; Chong, K. T., Identification of prokaryotic promoters and their strength by integrating heterogeneous features. *Genomics* **2020,** *112* (2), 1396-1403.
27. Ireland, W. T.; Beeler, S. M.; Flores-Bautista, E.; Belliveau, N. M.; Sweredoski, M. J.; Moradian, A.; Kinney, J. B.; Phillips, R., Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *bioRxiv* **2020**, 2020.01.18.910323.
28. Kong, L.-H.; Xiong, Z.-Q.; Song, X.; Xia, Y.-J.; Zhang, N.; Ai, L.-Z., Characterization of a Panel of Strong Constitutive Promoters from Streptococcus thermophilus for Fine-Tuning Gene Expression. *ACS Synthetic Biology* **2019,** *8* (6), 1469-1472.
29. Agrawal, D. K.; Tang, X.; Westbrook, A.; Marshall, R.; Maxwell, C. S.; Lucks, J.; Noireaux, V.; Beisel, C. L.; Dunlop, M. J.; Franco, E., Mathematical Modeling of RNA-Based Architectures for Closed Loop Control of Gene Expression. *ACS Synthetic Biology* **2018,** *7* (5), 1219-1228.
30. Westbrook, A.; Tang, X.; Marshall, R.; Maxwell, C. S.; Chappell, J.; Agrawal, D. K.; Dunlop, M. J.; Noireaux, V.; Beisel, C. L.; Lucks, J.; Franco, E., Distinct timescales of RNA regulators enable the construction of a genetic pulse generator. *Biotechnology and Bioengineering* **2019,** *116* (5), 1139-1151.
31. Espah Borujeni, A.; Channarasappa, A. S.; Salis, H. M., Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic acids research* **2013,** *42* (4), 2646-2659.
32. Salis, H. M.; Mirsky, E. A.; Voigt, C. A., Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* **2009,** *27* (10), 946-950.

33. Espah Borujeni, A.; Cetnar, D.; Farasat, I.; Smith, A.; Lundgren, N.; Salis, H. M., Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic acids research* **2017,** *45* (9), 5437-5448.
34. Misra, T. K.; Apirion, D., RNase E, an RNA processing enzyme from Escherichia coli. *Journal of Biological Chemistry* **1979,** *254* (21), 11154-11159.
35. Robertson, H. D.; Webster, R. E.; Zinder, N. D., Purification and properties of ribonuclease III from Escherichia coli. *Journal of Biological Chemistry* **1968,** *243* (1), 82-91.
36. Donovan, W. P.; Kushner, S. R., Polynucleotide phosphorylase and ribonuclease II are required for cell viability and mRNA turnover in Escherichia coli K-12. *Proceedings of the National Academy of Sciences* **1986,** *83* (1), 120-124.
37. Rauhut, R.; Klug, G., mRNA degradation in bacteria. *FEMS Microbiology Reviews* **1999,** *23* (3), 353-370.
38. Hui, M. P.; Foley, P. L.; Belasco, J. G., Messenger RNA Degradation in Bacterial Cells. *Annual Review of Genetics* **2014,** *48* (1), 537-559.
39. Smolke, C. D.; Carrier, T. A.; Keasling, J. D., Coordinated, Differential Expression of Two Genes through Directed mRNA Cleavage and Stabilization by Secondary Structures. *Applied and Environmental Microbiology* **2000,** *66* (12), 5399-5405.
40. Pfleger, B. F.; Pitera, D. J.; Smolke, C. D.; Keasling, J. D., Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nature Biotechnology* **2006,** *24* (8), 1027-1032.
41. Hanson, G.; Coller, J., Codon optimality, bias and usage in translation and mRNA decay. *Nature reviews Molecular cell biology* **2018,** *19* (1), 20-30.
42. Dar, D.; Sorek, R., Extensive reshaping of bacterial operons by programmed mRNA decay. *PLoS genetics* **2018,** *14* (4), e1007354.
43. Belasco, J. G.; Biggins, C. F., Mechanisms of mRNA decay in bacteria: a perspective. *Gene* **1988,** *72* (1-2), 15-23.
44. Vasilyev, N.; Serganov, A., Structures of RNA complexes with the Escherichia coli RNA pyrophosphohydrolase RppH unveil the basis for specific 5'-end-dependent mRNA decay. *Journal of Biological Chemistry* **2015,** *290* (15), 9487-9499.
45. Celesnik, H.; Deana, A.; Belasco, J. G., Initiation of RNA decay in Escherichia coli by 5' pyrophosphate removal. *Molecular cell* **2007,** *27* (1), 79-90.
46. Gao, A.; Vasilyev, N.; Serganov, A.; Luciano, D. J.; Richards, J.; Belasco, J. G.; Levenson-Palmer, R.; Traaseth, N. J.; Marsiglia, W. M., Structural and kinetic insights into stimulation of RppH-dependent RNA degradation by the metabolic enzyme DapF. *Nucleic Acids Research* **2018,** *46* (13), 6841-6856.
47. Gao, A.; Vasilyev, N.; Kaushik, A.; Duan, W.; Serganov, A., Principles of RNA and nucleotide discrimination by the RNA processing enzyme RppH. *Nucleic acids research* **2020,** *48* (7), 3776-3788.
48. Foley, P. L.; Hsieh, P.-k.; Luciano, D. J.; Belasco, J. G., Specificity and evolutionary conservation of the Escherichia coli RNA pyrophosphohydrolase RppH. *Journal of Biological Chemistry* **2015,** *290* (15), 9478-9486.
49. Chen, H.; Shiroguchi, K.; Ge, H.; Xie, X. S., Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Molecular Systems Biology* **2015,** *11* (1), 781.
50. Makarova, O. V.; Makarov, E. M.; Sousa, R.; Dreyfus, M., Transcribing of Escherichia coli genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with

polymerase speed. *Proceedings of the National Academy of Sciences* **1995,** *92* (26), 12250-12254.
51. DUBEY, A. K.; BAKER, C. S.; ROMEO, T.; BABITZKE, P., RNA sequence and secondary structure participate in high-affinity CsrA–RNA interaction. *RNA* **2005,** *11* (10), 1579-1587.
52. Zhang, A.; Wassarman, K. M.; Rosenow, C.; Tjaden, B. C.; Storz, G.; Gottesman, S., Global analysis of small RNA and mRNA targets of Hfq. *Molecular Microbiology* **2003,** *50* (4), 1111-1124.
53. Garrey, S. M.; Blech, M.; Riffell, J. L.; Hankins, J. S.; Stickney, L. M.; Diver, M.; Hsu, Y.-H. R.; Kunanithy, V.; Mackie, G. A., Substrate binding and active site residues in RNases E and G: the role of the 5'-sensor. *Journal of Biological Chemistry* **2009**.
54. Clarke, J. E.; Kime, L.; Romero A., D.; McDowall, K. J., Direct entry by RNase E is a major pathway for the degradation and processing of RNA in Escherichia coli. *Nucleic Acids Research* **2014,** *42* (18), 11733-11751.
55. Spickler, C.; Stronge, V.; Mackie, G. A., Preferential cleavage of degradative intermediates of rpsT mRNA by the Escherichia coli RNA degradosome. *Journal of bacteriology* **2001,** *183* (3), 1106-1109.
56. Aristarkhov, A.; Mikulskis, A.; Belasco, J. G.; Lin, E. C., Translation of the adhE transcript to produce ethanol dehydrogenase requires RNase III cleavage in Escherichia coli. *Journal of Bacteriology* **1996,** *178* (14), 4327-4332.
57. Gordon, G. C.; Cameron, J. C.; Pfleger, B. F., RNA Sequencing Identifies New RNase III Cleavage Sites in Escherichia coli and Reveals Increased Regulation of mRNA. *mBio* **2017,** *8* (2), e00128-17.
58. Miczak, A.; Kaberdin, V. R.; Wei, C. L.; Lin-Chao, S., Proteins associated with RNase E in a multicomponent ribonucleolytic complex. *Proceedings of the National Academy of Sciences* **1996,** *93* (9), 3865-3869.
59. Cheng, Z.-F.; Deutscher, M. P., An important role for RNase R in mRNA decay. *Molecular cell* **2005,** *17* (2), 313-318.
60. Spickler, C.; Mackie, G. A., Action of RNase II and Polynucleotide Phosphorylase against RNAs Containing Stem-Loops of Defined Structure. *Journal of Bacteriology* **2000,** *182* (9), 2422-2427.
61. Deutscher, M. P.; Reuven, N. B., Enzymatic basis for hydrolytic versus phosphorolytic mRNA degradation in Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences* **1991,** *88* (8), 3277-3280.
62. Jourdan, S. S.; McDowall, K. J., Sensing of 5’ monophosphate by Escherichia coli RNase G can significantly enhance association with RNA and stimulate the decay of functional mRNA transcripts in vivo. *Molecular Microbiology* **2008,** *67* (1), 102-115.
63. Kim, K.; Sim, S.-H.; Jeon, C. O.; Lee, Y.; Lee, K., Base substitutions at scissile bond sites are sufficient to alter RNA-binding and cleavage activity of RNase III. *FEMS Microbiology Letters* **2011,** *315* (1), 30-37.
64. Ehretsmann, C. P.; Carpousis, A. J.; Krisch, H. M., Specificity of Escherichia coli endoribonuclease RNase E: in vivo and in vitro analysis of mutants in a bacteriophage T4 mRNA processing site. *Genes & development* **1992,** *6* (1), 149-159.
65. Luciano, D. J.; Vasilyev, N.; Richards, J.; Serganov, A.; Belasco, J. G., A novel RNA phosphorylation state enables 5’ end-dependent degradation in Escherichia coli. *Molecular cell* **2017,** *67* (1), 44-54. e6.

66. Del Campo, C.; Bartholomäus, A.; Fedyunin, I.; Ignatova, Z., Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS genetics* **2015,** *11* (10), e1005613-e1005613.
67. Deana, A.; Belasco, J. G., Lost in translation: the influence of ribosomes on bacterial mRNA decay. *Genes & development* **2005,** *19* (21), 2526-2533.
68. Eriksen, M.; Sneppen, K.; Pedersen, S.; Mitarai, N., Occlusion of the Ribosome Binding Site Connects the Translational Initiation Frequency, mRNA Stability and Premature Transcription Termination. *Frontiers in Microbiology* **2017,** *8* (362).
69. Emory, S. A.; Bouvet, P.; Belasco, J. G., A 5'-terminal stem-loop structure can stabilize mRNA in Escherichia coli. *Genes & Development* **1992,** *6* (1), 135-148.
70. Zhang, Y.; Zhang, J.; Hoeflich, K. P.; Ikura, M.; Qing, G.; Inouye, M., MazF Cleaves Cellular mRNAs Specifically at ACA to Block Protein Synthesis in Escherichia coli. *Molecular Cell* **2003,** *12* (4), 913-923.
71. Court, D. L.; Gan, J.; Liang, Y.-H.; Shaw, G. X.; Tropea, J. E.; Costantino, N.; Waugh, D. S.; Ji, X., RNase III: Genetics and Function; Structure and Mechanism. *Annual Review of Genetics* **2013,** *47* (1), 405-431.
72. Zhang, H.; Kolb, F. A.; Jaskiewicz, L.; Westhof, E.; Filipowicz, W., Single processing center models for human Dicer and bacterial RNase III. *Cell* **2004,** *118* (1), 57-68.
73. Chao, Y.; Li, L.; Girodat, D.; Förstner, K. U.; Said, N.; Corcoran, C.; Śmiga, M.; Papenfort, K.; Reinhardt, R.; Wieden, H.-J.; Luisi, B. F.; Vogel, J., In Vivo Cleavage Map Illuminates the Central Role of RNase E in Coding and Non-coding RNA Pathways. *Molecular Cell* **2017,** *65* (1), 39-51.
74. Li, G.-W.; Burkhardt, D.; Gross, C.; Weissman, Jonathan S., Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* **2014,** *157* (3), 624-635.
75. Lorenz, R.; Bernhart, S. H.; Zu Siederdissen, C. H.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L., ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **2011,** *6* (1), 26.
76. Chen, H.; Meisburger, S. P.; Pabit, S. A.; Sutton, J. L.; Webb, W. W.; Pollack, L., Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proceedings of the National Academy of Sciences* **2012,** *109* (3), 799-804.
77. Doose, S.; Barsch, H.; Sauer, M., Polymer properties of polythymine as revealed by translational diffusion. *Biophys J* **2007,** *93* (4), 1224-1234.
78. Laurence, T. A.; Kong, X.; Jäger, M.; Weiss, S., Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (48), 17348-17353.
79. Caliskan, G.; Hyeon, C.; Perez-Salas, U.; Briber, R.; Woodson, S.; Thirumalai, D., Persistence length changes dramatically as RNA folds. *Physical review letters* **2005,** *95* (26), 268303.
80. Braun, F.; Le Derout, J.; Régnier, P., Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of Escherichia coli. *The EMBO Journal* **1998,** *17* (16), 4790-4797.
81. Yarchuk, O.; Jacques, N.; Guillerez, J.; Dreyfus, M., Interdependence of translation, transcription and mRNA degradation in the lacZ gene. *Journal of Molecular Biology* **1992,** *226* (3), 581-596.
82. Dreyfus, M., Killer and protective ribosomes. *Progress in molecular biology and translational science* **2009,** *85*, 423-466.

83. Sharp, J. S.; Bechhofer, D. H., Effect of translational signals on mRNA decay in Bacillus subtilis. *Journal of Bacteriology* **2003,** *185* (18), 5372-5379.
84. Farasat, I.; Kushwaha, M.; Collens, J.; Easterbrook, M.; Guido, M.; Salis, H. M., Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Molecular Systems Biology* **2014,** *10* (6), 731.
85. Shaw, L. B.; Zia, R. K. P.; Lee, K. H., Totally asymmetric exclusion process with extended objects: A model for protein synthesis. *Physical Review E* **2003,** *68* (2), 021910.
86. Sharma, A. K.; Chowdhury, D., Stochastic theory of protein synthesis and polysome: Ribosome profile on a single mRNA transcript. *Journal of Theoretical Biology* **2011,** *289*, 36-46.
87. Luciano, D. J.; Hui, M. P.; Deana, A.; Foley, P. L.; Belasco, K. J.; Belasco, J. G., Differential Control of the Rate of 5'-End-Dependent mRNA Degradation in *Escherichia coli*. *Journal of Bacteriology* **2012,** *194* (22), 6233-6239.
88. Mackie, G. A., Ribonuclease E is a 5'-end-dependent endonuclease. *Nature* **1998,** *395* (6703), 720-724.
89. Mackie, G. A., Stabilization of circular rpsT mRNA demonstrates the 5'-end dependence of RNase E action in vivo. *Journal of Biological Chemistry* **2000,** *275* (33), 25069-25072.
90. Chen, Y.-J.; Liu, P.; Nielsen, A. A. K.; Brophy, J. A. N.; Clancy, K.; Peterson, T.; Voigt, C. A., Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods* **2013,** *10*, 659.
91. Salamov, V. S. A.; Solovyevand, A., Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and its applications in agriculture, biomedicine and environmental studies. Hauppauge: Nova Science Publishers* **2011**, 61-78.
92. Urtecho, G.; Tripp, A. D.; Insigne, K. D.; Kim, H.; Kosuri, S., Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry* **2019,** *58* (11), 1539-1551.
93. Ireland, W. T.; Beeler, S. M.; Flores-Bautista, E.; McCarty, N. S.; Röschinger, T.; Belliveau, N. M.; Sweredoski, M. J.; Moradian, A.; Kinney, J. B.; Phillips, R., Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *Elife* **2020,** *9*, e55308.
94. Wilson, E. H.; Groom, J. D.; Sarfatis, M. C.; Ford, S. M.; Lidstrom, M. E.; Beck, D. A. C., A Computational Framework for Identifying Promoter Sequences in Nonmodel Organisms Using RNA-seq Data Sets. *ACS Synthetic Biology* **2021,** *10* (6), 1394-1405.
95. Yu, T. C.; Liu, W. L.; Brinck, M. S.; Davis, J. E.; Shek, J.; Bower, G.; Einav, T.; Insigne, K. D.; Phillips, R.; Kosuri, S.; Urtecho, G., Multiplexed characterization of rationally designed promoter architectures deconstructs combinatorial logic for IPTG-inducible systems. *Nature Communications* **2021,** *12* (1), 325.
96. Warman, E.; Forrest, D.; Wade, J. T.; Grainger, D. C., Widespread divergent transcription from prokaryotic promoters. *bioRxiv* **2020**, 2020.01.31.928960.
97. Espah Borujeni, A.; Cetnar, D.; Farasat, I.; Smith, A.; Lundgren, N.; Salis, H. M., Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Research* **2017,** *45* (9), 5437-5448.
98. Cetnar, D. P.; Salis, H. M., Systematic quantification of sequence and structural determinants controlling mRNA stability in bacterial operons. *ACS Synthetic Biology* **2021,** *10* (2), 318-332.
99. Zhang, Q.; Ma, D.; Wu, F.; Standage-Beier, K.; Chen, X.; Wu, K.; Green, A. A.; Wang, X., Predictable control of RNA lifetime using engineered degradation-tuning RNAs. *Nature Chemical Biology* **2021,** *17* (7), 828-836.

100. Nielsen, A. A.; Der, B. S.; Shin, J.; Vaidyanathan, P.; Paralanov, V.; Strychalski, E. A.; Ross, D.; Densmore, D.; Voigt, C. A., Genetic circuit design automation. *Science* **2016,** *352* (6281).
101. George, K. W.; Thompson, M. G.; Kang, A.; Baidoo, E.; Wang, G.; Chan, L. J. G.; Adams, P. D.; Petzold, C. J.; Keasling, J. D.; Lee, T. S., Metabolic engineering for the high-yield production of isoprenoid-based C 5 alcohols in E. coli. *Scientific reports* **2015,** *5* (1), 1-12.
102. Ghosh, I. N.; Martien, J.; Hebert, A. S.; Zhang, Y.; Coon, J. J.; Amador-Noguez, D.; Landick, R., OptSSeq explores enzyme expression and function landscapes to maximize isobutanol production rate. *Metabolic engineering* **2019,** *52*, 324-340.
103. Garrey, S. M.; Blech, M.; Riffell, J. L.; Hankins, J. S.; Stickney, L. M.; Diver, M.; Hsu, Y.-H. R.; Kunanithy, V.; Mackie, G. A., Substrate binding and active site residues in RNases E and G. *Journal of Biological Chemistry* **2009,** *284* (46), 31843-31850.
104. Clarke, J. E.; Kime, L.; Romero A, D.; McDowall, K. J., Direct entry by RNase E is a major pathway for the degradation and processing of RNA in Escherichia coli. *Nucleic acids research* **2014,** *42* (18), 11733-11751.
105. Court, D. L.; Gan, J.; Liang, Y.-H.; Shaw, G. X.; Tropea, J. E.; Costantino, N.; Waugh, D. S.; Ji, X., RNase III: genetics and function; structure and mechanism. *Annual review of genetics* **2013,** *47*, 405-431.
106. Vogel, J.; Luisi, B. F., Hfq and its constellation of RNA. *Nature Reviews Microbiology* **2011,** *9* (8), 578-589.
107. Clifton, K. P.; Jones, E. M.; Paudel, S.; Marken, J. P.; Monette, C. E.; Halleran, A. D.; Epp, L.; Saha, M. S., The genetic insulator RiboJ increases expression of insulated genes. *J Biol Eng* **2018,** *12*, 23-23.
108. Vlková, M.; Morampalli, B. R.; Silander, O. K., Efficiency of the synthetic self-splicing RiboJ ribozyme is robust to cis- and trans-changes in genetic background. *bioRxiv* **2021**, 2021.03.17.435894.
109. Lou, C.; Stanton, B.; Chen, Y.-J.; Munsky, B.; Voigt, C. A., Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature biotechnology* **2012,** *30* (11), 1137-1142.
110. Esquerré, T.; Laguerre, S.; Turlan, C.; Carpousis, A. J.; Girbal, L.; Cocaign-Bousquet, M., Dual role of transcription and transcript stability in the regulation of gene expression in Escherichia coli cells cultured on glucose at different growth rates. *Nucleic Acids Research* **2013,** *42* (4), 2460-2472.
111. Laguerre, S.; González, I.; Nouaille, S.; Moisan, A.; Villa-Vialaneix, N.; Gaspin, C.; Bouvier, M.; Carpousis, A. J.; Cocaign-Bousquet, M.; Girbal, L., Large-scale measurement of mRNA degradation in Escherichia coli: to delay or not to delay. *Methods in enzymology* **2018,** *612*, 47-66.
112. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y., Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **2017,** *30*, 3146-3154.
113. Chen, H.; Shiroguchi, K.; Ge, H.; Xie, X. S., Genome-wide study of mRNA degradation and transcript elongation in E scherichia coli. *Molecular systems biology* **2015,** *11* (1), 781.
114. Lorenz, R.; Bernhart, S. H.; Zu Siederdissen, C. H.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L., ViennaRNA Package 2.0. *Algorithms for molecular biology* **2011,** *6* (1), 1-14.
115. Shen, P.; Huang, H. V., Homologous recombination in Escherichia coli: dependence on substrate length and homology. *Genetics* **1986,** *112* (3), 441-457.
116. Hua, S.-b.; Qiu, M.; Chan, E.; Zhu, L.; Luo, Y., Minimum Length of Sequence Homology Required forin VivoCloning by Homologous Recombination in Yeast. *Plasmid* **1997,** *38* (2), 91-96.

117. Peeters, B. P.; de Boer, J. H.; Bron, S.; Venema, G., Structural plasmid instability in Bacillus subtilis: effect of direct and inverted repeats. *Molecular and General Genetics MGG* **1988,** *212* (3), 450-458.
118. Handler, R. M.; Shonnard, D. R.; Griffing, E. M.; Lai, A.; Palou-Rivera, I., Life Cycle Assessments of Ethanol Production via Gas Fermentation: Anticipated Greenhouse Gas Emissions for Cellulosic and Waste Gas Feedstocks. *Industrial & Engineering Chemistry Research* **2016,** *55* (12), 3253-3261.
119. Lehnik-Habrink, M.; Schaffer, M.; Mäder, U.; Diethmaier, C.; Herzberg, C.; Stülke, J., RNA processing in Bacillus subtilis: identification of targets of the essential RNase Y. *Molecular Microbiology* **2011,** *81* (6), 1459-1473.
120. Even, S.; Pellegrini, O.; Zig, L.; Labas, V.; Vinh, J.; Bréchemmier-Baey, D.; Putzer, H., Ribonucleases J1 and J2: two novel endoribonucleases in B.subtilis with functional homology to E.coli RNase E. *Nucleic Acids Research* **2005,** *33* (7), 2141-2152.
121. Kaberdin, V. R.; Bizebard, T., Characterization of Aquifex aeolicus RNase E/G. *Biochemical and Biophysical Research Communications* **2005,** *327* (2), 382-392.
122. Wade, J. T.; Roa, D. C.; Grainger, D. C.; Hurd, D.; Busby, S. J. W.; Struhl, K.; Nudler, E., Extensive functional overlap between σ factors in Escherichia coli. *Nature Structural & Molecular Biology* **2006,** *13* (9), 806-814.

VITA

Daniel P. Cetnar

Education:

Ph.D. in Chemical Engineering, The Pennsylvania State University, 2015-2021
B.S in Biochemistry, The Pennsylvania State University, 2011-2015

Publications:

- **Cetnar, D.** and Salis, H.M. (2021). Systematic Quantification of Sequence and Structural Determinants Controlling mRNA stability in Bacterial Operons. ACS Synthetic Biology.
- Borujeni, A. E., **Cetnar, D.**, Farasat, I., Smith, A., Lundgren, N., and Salis, H.M. (2017). Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. NAR.
- Halper, S., **Cetnar, D.**, and Salis, H.M. (2017). An Automated Pipeline for Engineering Many-Enzyme Pathways: Computational Sequence Design, Pathway Expression-Flux Mapping, and Scalable Optimization. Methods in Metabolic Engr.
- Reis, A., Halper, S., Vezeau, G., **Cetnar, D.**, Hossain, A., Clauer, P., Salis, H.M. (2019). Simultaneous repression of multiple bacterial genes using nonrepetitive extra-long sgRNA arrays. Nature Biotechnology.
- Hossain, A., Lopez, E., Halper, S., **Cetnar, D.**, Reis, A., Strickland, D., Klavins, E., Salis, H.M. (2020). Automated design of thousands of nonrepetitive parts for engineering stable genetic systems. Nature Biotechnology.
- Ferlez, B., Cowgill, J., Dong, W., Gisriel, C., Lin, S., Flores, M., Walters, K., **Cetnar, D.**, Redding, K., and Golbeck, J. (2016). Thermodynamics of the electron acceptors in Heliobacterium modesticaldum: An exemplar of an early homodimeric type I photosynthetic reaction center. Biochemistry.
- **Cetnar, D.**, Hossain, A., Vezeau, G., Salis, H.M. (2021). Comprehensive modeling and design of 5' UTRs for RNA stability using a 62,000 unique 5' UTR member library. Manuscript in preparation.
- Feroz, H., **Cetnar, D.**, Hewlett, R., Sharma, S., Holstein, M., Ghose, S., Li, Z. (2021). Establishing a Surrogate Model for Inactivation of Enveloped Viruses to Screen Viral Clearance Conditions During Biotherapeutics Process Development. Biotechnology Journal. Manuscript submitted.
- Hossain, A., **Cetnar, D.**, La Fleur, T., Salis, H.M. (2021). Automated Design and Analysis of Oligopool Libraries. Manuscript in preparation.

Conference Presentations:

- EBRC, Virtual. April 2021. Poster.
- Annual AIChE Meeting, Pittsburgh, PA. November 2018. Presentation.
- EBRC, Seattle, WA. March 2018. Poster.
- EBRC, Atlanta, GA. Sept 2017. Poster.
- IWBDA, Pittsburgh, PA. Aug 2017. Presentation.
- Synthetic Biology: Engineering, Evolution & Design (SEED), Chicago, IL. July 2016. Poster.