Developing a Massively Parallel Reporter Assay for Studying Gene Regulatory Elements

Ban Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Georg Seelig, Chair

Sreeram Kannan

Andrew C. Hsieh

Program Authorized to Offer Degree:

Department of Electrical and Computer Engineering

University of Washington

Abstract

Developing a Massively Parallel Reporter Assay for Studying Gene Regulatory Elements

Ban Wang

Chair of the Supervisory Committee:
Associate Professor Georg Seelig
Electrical and Computer Engineering
Computer Science & Engineering

Although it took roughly 13 years for the Human Genome Project to be finished and cost roughly $2.7 billion, it was a big step that helped us get the genetic information to understand how genomes have impact on individuals and populations. The Human Genome Project provided a view of human genome sequences that is not representative of any one individual. Finding genetic variants any individual may carry and then to associate these variants with phenotypes or even diseases, was unrealistic back in the beginning of the 21th century given the time and cost of genome sequencing. However, with the rapid development of DNA sequencing technology, the cost of sequencing one genome has dropped to around $1000 today. Now, the main challenge is not sequencing the genome but to link the genotypes and phenotypes. In particular, the regulatory rules

that govern how non-coding regions such as untranslated regions (UTRs) and introns control gene expression are not fully understood and making it challenging to interpret genetic variants that occur in such regions.

In this dissertation, we reported how data from a massively parallel reporter assay (MPRA) can be combined with deep learning to build a predictive model that can be used to score any variant in an important class of non-coding regions. The MPRA we presented in this dissertation was specifically used on the 5' untranslated region (5' UTR), which is the region on an mRNA that is directly upstream from the coding sequence. We were specifically interested in 5' UTRs because of their significant role in translation regulation. Although this predictive model was trained on data from a fully synthetic reporter library containing random sequence, it performed well on the task of predicting the impact of variants in the human genome. We also showed that the model could be used to design new sequences with targeted level of protein production, which provides a valuable tool for applications in mRNA therapeutics. This assay could be applied to any regions of interest with the idea of building predictive models through machine learning on big datasets collected from synthetic random library, showing the power of how machine learning coupled with synthetic biology in helping us understand the fundamentals of nature and life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

When I tried to look back, six years definitely would not be considered as a short time period in my life, and I am so grateful that I can spend the six years to obtain my doctorate degree here at the University of Washington, majoring in Electrical and Computer Engineering and working in Seelig Lab. This was truly a joyful and fruitful experience beyond all my imagination and expectation when I started.

First, I want to express my deepest appreciation to Georg Seelig for everything. Georg is such an amazing scientist, principle investigator, advisor and mentor. The support I received from him covered all aspects, including but not limited to how to dive into scientific research, communicate with colleagues, solve problems, overcome obstacles, fight with frustration. These could all be long-term classes that I would learn from that helped me go deep into science. I also want to thank my committee members: Prof. Sreeram Kannan, Prof. Andrew C. Hsieh and Prof. James M. Carothers for their time and helpful advices and feedback on my dissertation.

Secondly, I want to thank all the Seelig lab members (Alberto Carignano, Alex Baryshev, Alexander Rosenberg, Alyssa LaFleur, Anna Kuchina, Arjun Khakar, Ben Groves, Charlie Roco, Erin Wilson, Gourab Chatterjee, Johannes Linder, Matthew Hirano, Max Darnell, Nick Bogard, Paul Sample, Randolph Lopez, Sergii Pochekailov, Sifang Chen, Sumit Mukherjee, Sunny Rao, Yuan-Jyue Chen, Yue Zhang) deeply, for all helpful discussions on research and all fun conversations on life. Specially, I want to thank Paul who was the major collaborator I had in the lab. We finished and published the project together and I learned so much from him along the way.

I also want to thank Nick who was actually my first mentor when I just joined the lab and offered me so much patience to teach me how to start even just from pipetting, and we had so many fun conversations talking about any topics we would like to talk about. Additionally, I want to thank Johannes who generously spent a lot of time on helping me set up the Optimus 5-Prime webtool and discussing all sorts of computational questions I may have with me.

Beyond research life, I want to thank all my friends here at UW: Xindi Liu, Bin Yu, Bowen Xue, Yedi Luo, Boling Yang. I would cherish all the memories that we had together, exploring fun places and having good food. I also want to thank all my friends not here locally in Seattle and even back in China for all kinds of mental support.

Most importantly, I want to thank my boyfriend Liangqi Gong, my parents and my family for their love and support. Without them, I would never be able to chase my dream so bravely and even finish my degree. I am who I am today only because all the love and support I got from them.

DEDICATION

This is dedicated to my parents, Xuequn Wang and Chunlei Wang.

Chapter 1. INTRODUCTION

The human genome contains 3 billion base pairs, and for any individual about 0.5% of these bases are distinct from a reference genome [1]. Nowadays, we can get human reference genome sequences very easily through all sorts of public datasets and we can sequence individual genomes for about $1000 [2].

Studies have shown that variation at the DNA level can explain traits such as skin pigmentation [3] or genetic disease [4]. To understand how DNA sequence determines these traits we first need to learn how a change in DNA sequence affects protein production. The genetic code[5] explains the relationship between coding sequences and protein; each three nucleotides correspond to a specific amino acid or a stop signal, the building blocks of proteins. Changes in coding regions result in different amino acid selection and can result in misfolded, truncated or otherwise non-functional proteins. However, the vast majority of the genome does not get translated but can still significantly regulate protein production [6]. How sequences in untranslated regions exert their effects is still poorly understood.

In the process of gene expression, genes are transcribed to mRNA first and mRNA is then translated into protein. Many non-coding regions mediate this complex process. For example, alternative splicing results in the inclusion or exclusion of particular exons to generate multiple mRNA isoforms co-transcriptionally [7]. In addition to that, alternative polyadenylation, which makes one gene code for several mRNAs that have different 3’ end, is also involved in mRNA maturation [8]. Protein levels are also influenced by mRNA stability which in turn is regulated by regulatory features including the 5’ cap structure, the 5’ UTR, the protein coding region, the 3’ UTR and the 3’ tail [9]. After mRNA is produced, translation starts where translational regulating

mechanisms start to work as well. The scanning process of the ribosome from the 5' to 3' direction enables the ribosome to initiate at an appropriate start codon. However, this can also be mediated by regulatory elements across the transcript through leaking scanning and initiation at internal ribosomal entry sites (IRES) [10].

In this dissertation, we were specifically interested in 5' UTR mediated translation. The 5' UTR is the region on an mRNA that is directly upstream of the coding sequence (CDS). Human 5' UTRs vary in length from tens to thousands of nucleotides with an average length of approximately 200 nucleotides [[11], [12]], and the 5' UTR plays a significant role in translation regulation as ribosomes scan this region to initiate at a start codon (Figure 1.1).

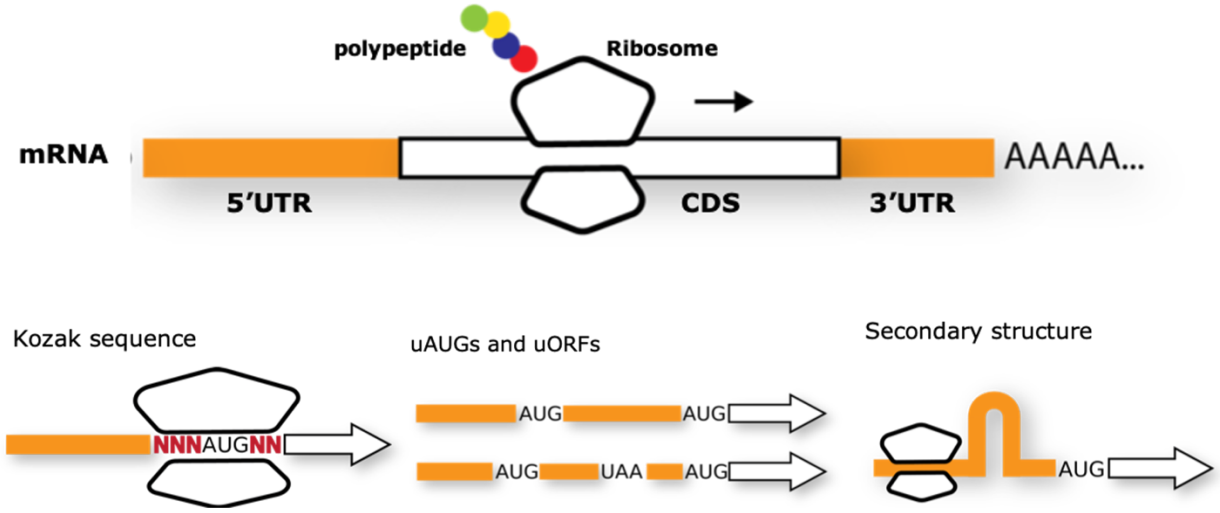


Figure 1.1. The 5' UTR sits immediately upstream of coding sequence and there are several regulatory elements have been individually characterized in 5' UTR.

The sequence of a 5' UTR is a primary determinant of translation efficiency [[13], [14]]. Previous studies have identified many *cis*-regulatory elements, such as upstream AUGs (uAUGs), upstream open reading frames (uORFs) [[14]–[16]], the Kozak sequence [[17], [18]], secondary

structure [[19], [20]] and recruitment sites of trans-acting proteins [[21], [22]]. In Figure 1.1, three regulatory elements identified in 5' UTR previously are shown as examples. One of the most famous regulatory elements is a region called the Kozak sequence which was studied in the mid-1980s and it is the region that is immediately surrounds the start codon. Specifically, the Kozak sequence includes three nucleotides upstream and two nucleotides downstream of the start codon. The surrounding sequences determine the strength of the start site - a purine (A or G) at -3 position and a G at +1 position are considered to be the strongest Kozak motif. Upstream AUGs and upstream open reading frames (uORFs) are well studied as well. The translation of main ORF is repressed when the translation starts at upstream AUGs or uORFs. Since the ribosome scans through the transcript from the 5' UTR to initiate translation, strong RNA secondary structures can exert repressive effects on translation as well, because a hairpin structure in the 5' UTR shown in Figure 1.1 could block ribosomes from scanning through to reach CDS. All these factors have been well characterized individually but there lacks a way to build up a comprehensive model to characterize 5' UTR by its sequence alone.

One comprehensive 5' UTR model will have a wide range of applications as shown in Figure 1.2. such as predicting expression levels for a given 5' UTR sequence or predicting how variants modulate protein production. Existing approaches such as quantitative trait locus analysis and genome wide association studies are limited to common variants and cannot scale to the enormous number of rare 5' UTR variants occurring in the human population [[23], [24]]. The 5' UTR predictive model we want to build should be able to score the impact of any 5' UTR variant on translation and provide a molecular basis reference for diagnosis potentially. Moreover, we can also use the 5' UTR predictive model to design new sequences to hit some targeted protein

expression level, and this could be a valuable asset for mRNA therapeutics, metabolic engineering, and protein manufacturing.

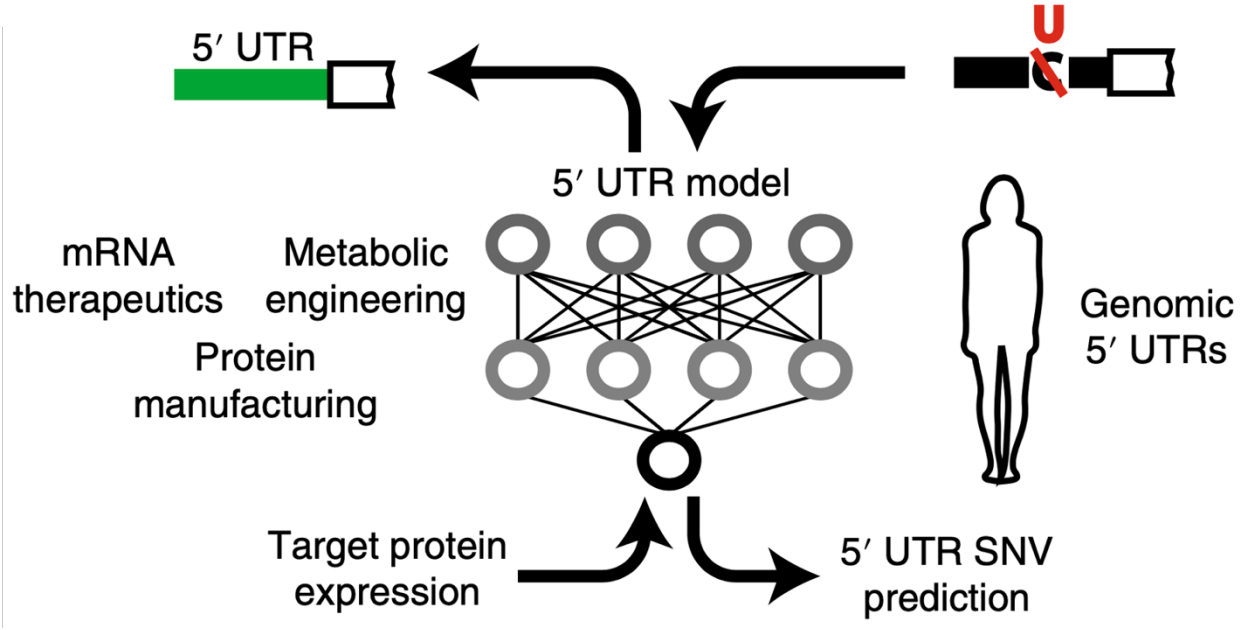


Figure 1.2. A wide range of potential applications indicates a strong motivation to build a 5' UTR model.

Building such a comprehensive model requires a huge dataset for analysis and modeling. MPRAs allow for simultaneous analysis of thousands to millions of sequence variants in a single experiment. By collecting such big datasets, researchers have been able to study biological processes including transcription [[20], [25]–[27]], mRNA stability [[28], [29]], alternative splicing [[30], [31]], alternative polyadenylation [[32]]. Machine learning algorithms have been shown to work incredibly well on these biological data to capture hidden information that may lie underneath obscure biological phenomena going beyond statistical results [[18], [31]–[33]].

In this dissertation, we reported the development of a polysome profiling-based MPRA that was able to assess the translation efficiency of each transcript in a library. We then report on the

design of a convolution neural network model that could explain up to 93% of the variation observed in the dataset [34]. With this model, we showed that we could engineer new sequences to achieve targeted translation levels and could predict the effect of single nucleotide variants (SNVs) on gene expression in human gene contexts. The idea of coupling MPRA and machine learning algorithm together would be very useful to study a variety of interested regions and the assay described in this dissertation was specifically used in studying 5' UTR, but the generalization of this assay is very straightforward and it by no means is only specific to 5' UTR. We expect more studies to take advantage of this assay to investigate regulatory elements in different regions of the genome.

Chapter 2. BACKGROUND AND RELATED WORKS

In this chapter, the background and previous works related to the idea in this dissertation are introduced. The massively parallel reporter assay (MPRA) was first introduced in 2009 by Patwardhan et al [26], where they managed to measure the effects of all possible single-nucleotide mutations for three bacteriophage promoters and three mammalian core promoters in one single experiment per promoter, and this hugely improved the efficiency of experiments as low-throughput assay would only be able to measure one mutation each time. After that, there were many MPRAs developed to achieve high-throughput studies with different goals, and a large portion of them relied on direct DNA/RNA sequencing. Two MPRAs were reviewed in this chapter to show how previous studies developed assay measurements other than sequencing only. Polysome profiling is introduced in this chapter which is one of the key technologies of our MPRA to show how this technique overcomes limitations of other MPRAs.

Even though MPRAs make it possible to provide huge amount of data in labor effective experiments, there is no way to test all variants that occur, and we might not take advantage of the collected datasets fully by only using standard statistical analysis. Applications of machine learning algorithms have been shown in various fields including image recognition [35], speech recognition [36], self-driving [37] and gaming [38] etc. Machine learning was also proven to be very useful in biological perspectives, two papers were reviewed here to demonstrate the power of coupling machine learning with biological studies.

2.1 MASSIVELY PARALLEL REPORTER ASSAY

The genetic reporter assay is a very well established and powerful tool that can be used to study the relationship between DNA sequences and gene regulatory activities. For example, how a

mutation in the 5' UTR affects the translation efficiency of that CDS can be studied by making a mutation, cloning it into a reporter construct, and testing it using flow cytometry to determine if there is a fluorescence signal level change that can be detected. However, the throughput of this assay is limited by the need to individually clone and assay the activity of each sequence of interest. MPRAs are made possible based on the developments in high-throughput DNA synthesis and sequencing technologies. MPRAs offer a high-throughput method to do directly comparison among large numbers of variants.

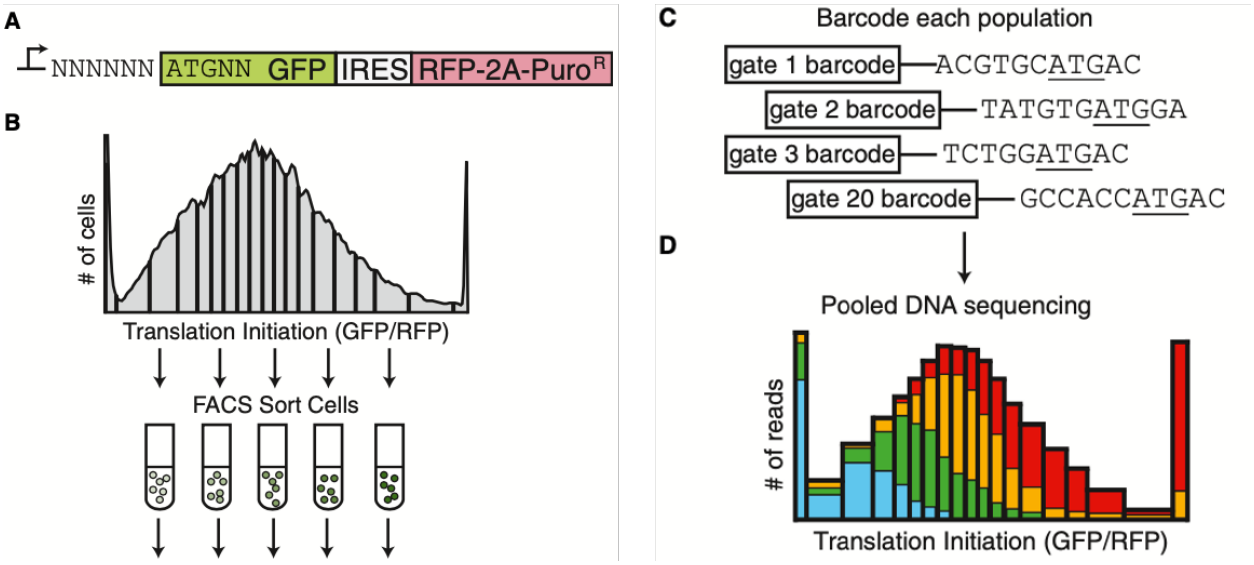


Figure 2.1. Noderer et al. [18] used FACS-seq to study mammalian translation initiation sites.

One massively parallel method to study translation is FACS-seq. Noderer et al. used FACS-seq to do quantitative analysis of mammalian translation initiation sites [18] (Figure 2.1). FACS-seq is composed of fluorescence-activated cell sorting (FACS) and high-throughput DNA sequencing (seq). Cells which are integrated with a single copy of a library member into the

genome are analyzed one by one and then binned based on their fluorescence levels. The PCR products of each bin will be tagged with an identifying DNA barcode and subjected to high throughput sequencing to connect the sequence to its fluorescence level. This study built up a library that covered all possible translation initiation sites (TIS) spanning from position -6 to +5 with start codon ATG fixed (NNNNNNATGNN), which formed a length of 8 nucleotides (nt) randomized region (Figure 2.1A) and stable transduced cells with library were sorted based on fluorescence signal (Figure 2.1B). The library sequence from each sorted bin were amplified and barcoded (Figure 2.1C), and finally the barcoded library was pooled and sequenced (Figure 2.1D). The experimental result agreed with the Kozak sequence, that was, the -3R (purine, R = A or G) and +4G to be the first and second most important bases for efficient initiation, respectively. Since the single values from the assay could not stand on their own but reveal trends, a dinucleotide position weight matrix (PWM) model was also developed to explain the relationship between TIS sequence and initiation efficiency, and a general high-efficiency TIS motif was defined through this model: RYMRMVAUGGC, where Y = U or C, M = A or C, R = A or G, and V = A, C or G.

Another growth-selection based MPRA was reported focusing on the 5' UTR region in yeast [33] (Figure 2.2). Cuperus et al. constructed a library structured as 50 nt randomized region in 5' UTR upstream of a His3 protein and made a library with nearly half a million distinct 5' UTRs. Yeast were grown in selective media devoid of histidine so only yeast carrying a library construct with the His3 reporter gene could grow. Since the growth rate was proportional to His3 protein expression, the growth rate for each variant could be determined by DNA-sequencing. The sequencing experiment effectively measured the frequency of each construct in the population in before and after growth. The resulting data agreed with previous studies in the Kozak motif and effects from uORFs and secondary structures.

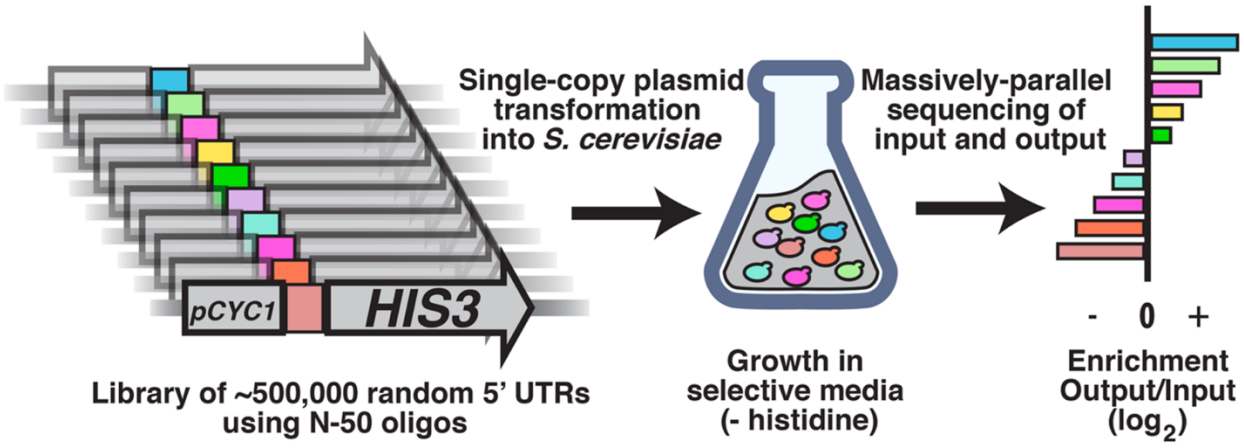


Figure 2.2. Cuperus et al. [33] used growth-selection based MPRA to study 5' UTR in yeast.

From these two reported MPRAs, we have seen the power of MPRAs for collecting huge datasets, for FACS-seq, 65,000 variants could be analyzed in one single experiment which spanned over all possible sequence combinations in that space, and the growth assay could go even higher to nearly half a million distinct variants. However, there are still several limitations for these MPRAs. Firstly, FACS-seq requires fluorescent signal and the growth assay requires His3 reporter gene as the coding sequence contexts, which means that these two MPRAs cannot be generalized to other CDS when arbitrary CDS regions are required for study. Secondly, both assays require each cell carry exactly one variant which requires genomic integration. Thirdly, growth rate or fluorescent level measures a combination of transcription and translation, therefore, it is not amenable to separate the regulatory impact of these two processes. Finally, these assays are not compatible with the use of synthetic or modified RNA which are of interest for mRNA therapeutics. To overcome all these limitations, we used polysome profiling as our MPRA approach to analyze our library which consisted of *in vitro* transcribed (IVT) mRNA.

2.2 POLYSOME PROFILING

Polysome profiling is a technique that can measure translation by quantifying the number of ribosomes on individual transcripts through a sucrose gradient and ultracentrifugation. The protocols for polysome profiling vary in details between different types of cells and treatment conditions, but in general cells are lysed first and loaded on top of prepared sucrose gradients that have light sucrose density on top and heavy sucrose density at the bottom so that it will harvest poor-translated transcripts (few ribosomes on transcripts) on the top and strong-translated transcripts (many ribosomes on transcripts) at the bottom after ultracentrifugation. The distribution of mRNAs is determined by the number of ribosomes bound but not the mRNA itself since the ribosome weighs much more than RNA. The sucrose gradient mixed with cell lysates after ultracentrifugation are processed on a fraction collector. The fractions are collected from the top to bottom which corresponds to the left to right direction of a polysome profile. Peaks are identified by the optical density determined by a UV detector. 40S, 60S, 80S (monosome) to polysomes. Peaks can be visualized and fractionated from the sucrose gradient and are collected for RNA extraction and future analysis (Figure 2.3).

Polysome profiling can be used to study translational regulation and it has been used for analysis on targeted mRNAs [[39]–[41]] and transcriptome-wide analysis [[42], [43]]. In 2016, nearly all human transcript isoforms were analyzed by polysome profiling combined with RNA-seq and the significance of features such as GC content, UTR length, and codon frequency were statistically determined [44]. Polysome profiling was applied to native transcripts which varied in all aspects including 5’ UTR, 3’ UTR, protein coding region sequence and length; as a result, the contribution from each component individually was difficult to determine. Our MPRA used a library with a fixed CDS and 3’ UTR, and 5’ UTR length so that only the sequence of the 5’ UTR

was different among variants. Then by using polysome profiling to collect translation profiles for all variants in our library, we could identify the isolated influences brought by the sequence in the 5' UTR.

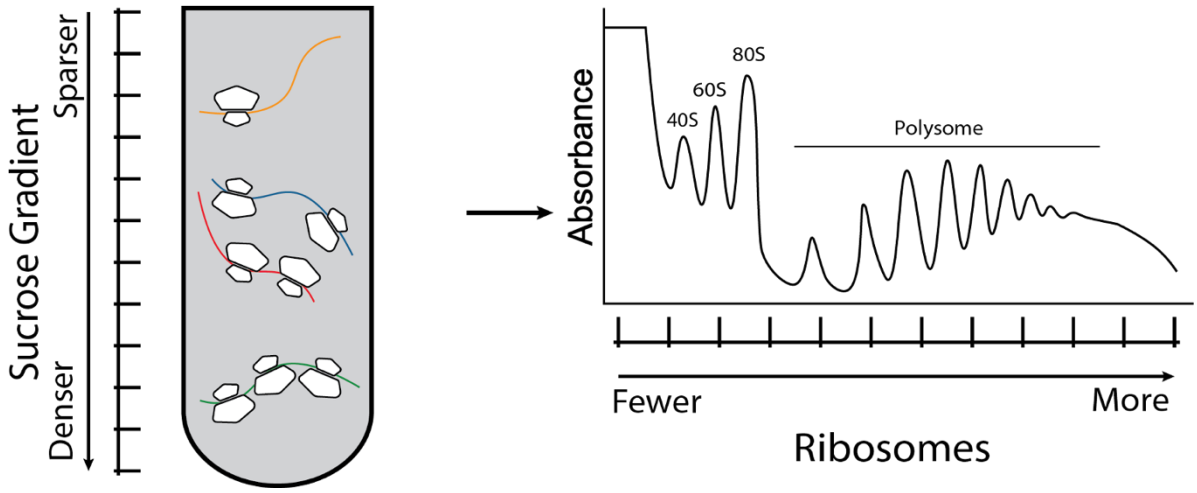


Figure 2.3. Overview of polysome profiling where sucrose gradient was used to separate transcripts bound by different number of ribosomes and a polysome profile was generated.

2.3 Machine Learning in Biology

Besides the board range of applications for machine learning algorithms such as image recognition, speech recognition and so on, machine learning has also proven useful in addressing challenges by producing quantitative biological insights [45].

The work used growth-selection based MPRA done by Cuperus et al. [33] introduced in previous section showed decent results after applying a convolutional neural network (CNN) to their data. The model they built using CNN was able to explain 62% of the variation in a test set. By taking advantage of CNN, they also visualized motifs generated by the trained CNN which contain upstream AUGs (uAUGs), stop codons and some guanine-rich structures.

Bogard et al. built a model called APARANT based on a deep neural network to predict alternative polyadenylation (APA) which is a major driver of transcriptome diversity in human cells [32]. Their model could predict synthetic and human 3' UTRs with high accuracy and recognize sequence motifs both known and novel from filters' visualization.

Jaganathan et al. worked on predicting splicing from primary sequence by a deep residual neural network and built SpliceAI to predict whether each position in a pre-mRNA transcript is a splice donor, splice acceptor, or neither [46]. SpliceAI could take up to 10,000 nucleotides of flanking sequence and accurately predict splice junctions. It could be used to study mutations that may associate with autism and intellectual disability that associated with altering splicing.

Alipanahi et al. used deep learning to build a model called DeepBind to predict the sequence specificities of DNA- and RNA- binding proteins [47]. DeepBind was able to be integrated into downstream applications including studying RNA binding proteins (RBP) in alternative splicing and analyzing variants that can affect transcription factor binding.

There are many more excellent studies other than the few introduced above to apply machine learning algorithms in biology. Computer scientist, molecular biologists, computational biologist and many other interdisciplinary scientists are working together closely to decode the complicated biological mysteries in a much faster manner nowadays than 10 years ago. In this dissertation, we will show our way of how to apply machine learning to MPRA data we have collected in the 5' UTR.

Chapter 3. EXPERIMENTAL WORKFLOW

In Chapter 1, we have described that the ultimate goal is to build a 5' UTR model which would have a variety of applications, and we would couple MPRA with machine learning to achieve this goal. Machine learning relies on big training dataset and a bigger dataset yields a better model in practice, and we can use MPRA to generate big datasets in a very high-throughput way. A majority of contents introduced in Chapter 3, 4 and 5 have been published by Dr. Paul Sample and I as co-first authors, and with many other collaborators in 2019 [34].

The first question was what kind of data should be collected to build such a model. Intuitively, we could use the human genome, since we can access human genome sequences very easily. We could track down all the 5' UTR sequences from human and map each sequence to its protein expression level. However, the model built on top of this dataset would not be able to decouple the regulatory effects from 5' UTR only, because for each native transcript, it has its own coding sequence (CDS) and 3' UTR which vary in both length and sequence content. CDSs can control protein expression since they encode for protein sequences and 3' UTRs can play a role in regulating translation. It would not be feasible for us to study 5' UTR-mediated translational regulation from this dataset and resulting model. What's more, the human genome encodes roughly 45,000 5' UTRs, which is not big enough to build accurate models. For the purpose of studying 5' UTR-mediated translational regulation specifically, we built an *in vitro* transcribed mRNA library instead of doing genomic integration.

The MPRA we developed here was based on polysome profiling with a lot of advantages over the other types of MPRAs introduced in Chapter 2. We designed the polysome profiling-based MPRA to specifically enrich the library signal of our interest and used a new metrics - mean ribosome load (MRL) as the proxy for protein expression level. The whole experimental workflow

is shown in Figure 3.1 – 1. random DNA plasmid library synthesis from random 50 nt oligos; 2. IVT mRNA library synthesis; 3. transfection in human cells; 4. polysome profiling; 5. RNA extraction and next-generation sequencing.

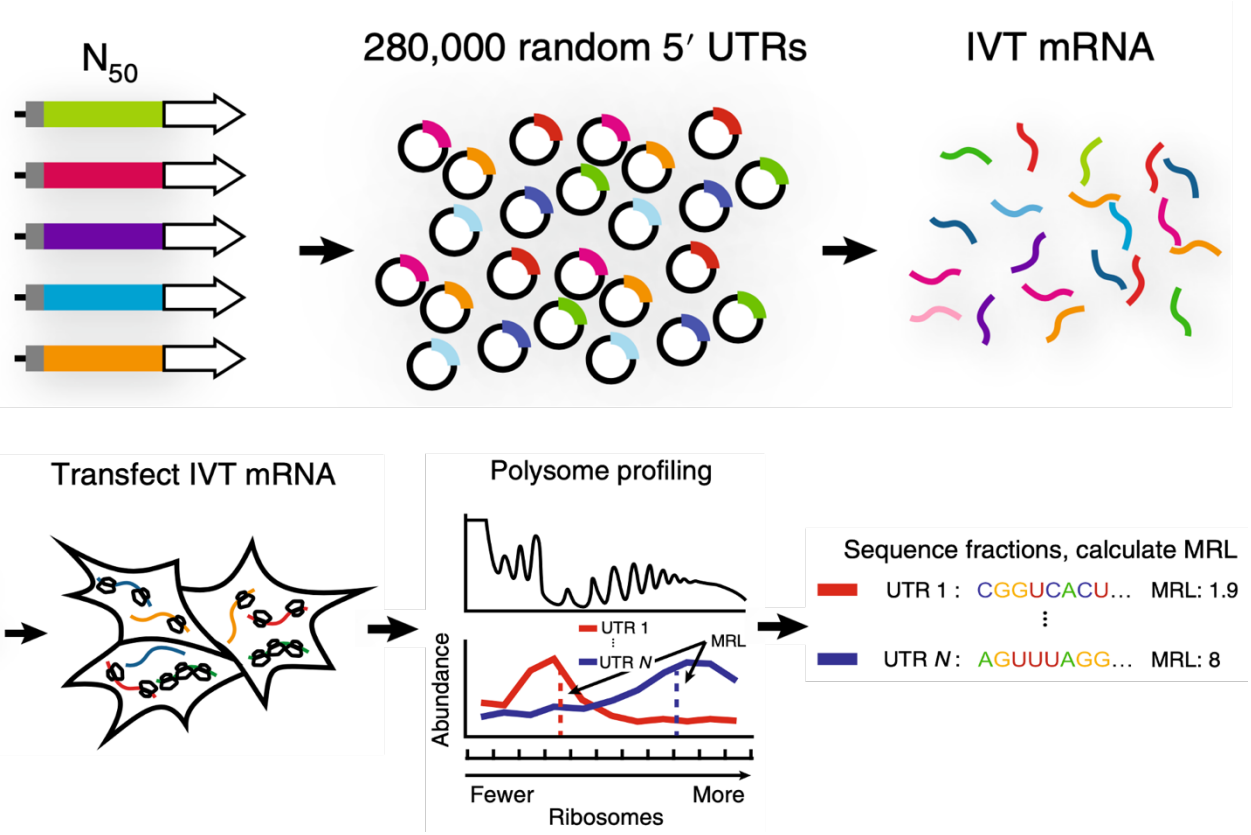


Figure 3.1. Overview of experimental workflow: random DNA plasmid library – IVT mRNA library – transfection in human cells – polysome profiling – next generation high-throughput sequencing.

3.1 LIBRARY CONSTRUCTION

The library construction introduced in this section was for the major random 50 nt library with defined 25 nt 5' end using eGFP as the CDS. Other versions of libraries were constructed in a similar way but slightly different which will be introduced individually below.

A DNA plasmid library was first built to serve as the template in *in vitro* transcription. A vector (pET 28) encoding a T7 promoter was followed by 25 nt of a defined 5' UTR (GGGACATCGTAGAGAGTCGTACTTA), then followed by eGFP coding sequence and a BGH poly-A signal. Between the 25 nt 5' UTR and eGFP CDS, there was an AgeI restriction site that allowed AgeI restriction enzyme to cut to allow for insertion of the 5' UTR library between the defined sequence and the CDS. This plasmid contained kanamycin marker in the backbone for growth selection. The oligonucleotide (Table 3.1 - primer 282) ordered from IDT that was used for library insertion contained the defined 5' UTR sequence, followed by 50 nucleotides of randomized bases and 21 nucleotides that overlapped the eGFP CDS (including the ATG start site). A reverse primer (Table 3.1 - primer 283) complementary to the 21-nucleotide eGFP overlap was used to produce a double-stranded product via Klenow extension with Klenow polymerase I (NEB). The vector and insert were assembled by Gibson reaction. Therefore, an eGFP DNA plasmid library was constructed containing a T7 promoter, a 25 nt defined region, a 50 nt random region, eGFP CDS, BGH poly-A signal in consequence (Figure 3.2). Design of the 25 nt defined region would allow for amplification after reverse transcription in later stage of the experiment. There were some modifications on eGFP CDS: two nucleotides at positions +11 (C to A) and +14 (C to T) in the EGFP CDS were changed in order to introduce stop codons (TAA) in frame -1 and -2 relative to ATG. The Gibson product was electroporated into 5-alpha electrocompetent *Escherichia coli*. A small portion of the electroporation was plated and resulted in ~750,000

colony-forming units and the rest was grown in liquid culture overnight (all bacteria were grown under kanamycin selection). The isolated plasmid was the eGFP library.

An mCherry library used mCherry instead of eGFP in the CDS was constructed in the same process and structured the same as described above. The same defined 5’ UTR that lay upstream of the randomized 50-nucleotide UTR in the eGFP library was used (Table 3.1 - primer 252). Klenow extension with primer 253 (Table 3.1) created the double-stranded insert that was assembled with the AgeI-linearized backbone by Gibson reaction. However, we didn’t intentionally place stop codons in the mCherry CDS.

Primer #	**Sequence (5′ to 3′)**
220	GACGTGTGCTCTTCCGATCTNNNNNNNNNNGTCTGGGTGCCCTCGTA
252	ATAGGGACATCGTAGAGAGTCGTACTTANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATGCCTCCCGAGAAGAAGATC
253	CACGCTCTTGATCTTCTTCTCGGGAGGCAT
254	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCAAACAACAGATGGCTGGCA
255	GCGAAATTAATACGACTCACTATAGGG
282	ATAGGGACATCGTAGAGAGTCGTACTTANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATGGGCGAATTAAGTAAGGGC

283	ACAGCTCCTCGCCCTTACTTAATTCGCCCA
289	GACGTGTGCTCTTCCGATCTNNNNNNNNNNAGATGAACTTCAGGGTCAGC
300	AGCGTGACAGGGACATCGTAGAGAGTCGTACTTA

Table 3.1. Primers used for library construction, IVT mRNA template, reverse transcription and high-throughput sequencing.

Library Name	**Library Sequence (5′ to 3′)**
eGFP	**GGGACATCGTAGAGAGTCGTACTTA**(N50)atgggcgaattaagtaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgccacctacggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccaccctcgtgaccaccctgacctacggcgtgcagtgcttcagccgctaccccgaccacatgaagcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaactacaagacccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcacaagctggagtacaactacaacagccacaacgtctatatcatggccgacaagcagaagaacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccgaccactaccagcagaacaccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccaagctgagcaaagaccccaacgagaagcgcgatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaagttcgaataaagctagcgcctcgactgtgccttctagttgccagccatctgttgtttg

mCherry	**GGGACATCGTAGAGAGTCGTACTTA**(N50)atgcctcccgagaagaagatcaagagc gtgagcaagggcgaggaggataacatggccatcatcaaggagttcatgcgcttcaaggtgcacatggagggctc cgtgaacggccacgagttcgagatcgagggcgagggcgagggccgcccctacgagggcacccagaccgcca agctgaaggtgaccaagggtggcccccctgcccttcgcctgggacatcctgtcccctcagttcatgtacggctccaa ggcctacgtgaagcaccccgccgacatccccgactacttgaagctgtccttccccgagggcttcaagtgggagcg cgtgatgaacttcgaggacggcggcgtggtgaccgtgacccaggactcctccctgcaggacggcgagttcatcta caaggtgaagctgcgcggcaccaacttcccctccgacggccccgtaatgcagaagaagaccatgggctgggag gcctcctccgagcggatgtaccccgaggacggcgccctgaagggcgagatcaagcagaggctgaagctgaag gacggcggccactacgacgctgaggtcaagaccacctacaaggccaagaagcccgtgcagctgcccggcgcct acaacgtcaacatcaagttggacatcacctcccacaacgaggactacaccatcgtggaacagtacgaacgcgccg agggccgccactccaccggcggcatggacgagctgtacaagtcttaacgcctcgactgtgccttctagttgccagc catctgttgtttg

Table 3.2. eGFP and mCherry library sequences.

A linear template for *in vitro* transcription was produced via PCR using primer 254 and primer 255 (Table 3.1) of the library DNA plasmid. The reverse primer (primer 254) for the PCR was designed in a way such that this double-stranded PCR product had a truncated BGH poly-A signal sequence which was adopted from the plasmid and 70 nt poly-A signal from the reverse primer overhang. Therefore, the double-stranded DNA product had a T7 promoter at the 5' end and a truncated BGH poly(A) signal sequence followed by a 70-nucleotide poly(A) sequence (Figure 3.2). Agarose gel was used to exam the correct length of DNA linear template and purified the linear fragment out of circular plasmid. Then *in vitro* transcription was carried out using the linear

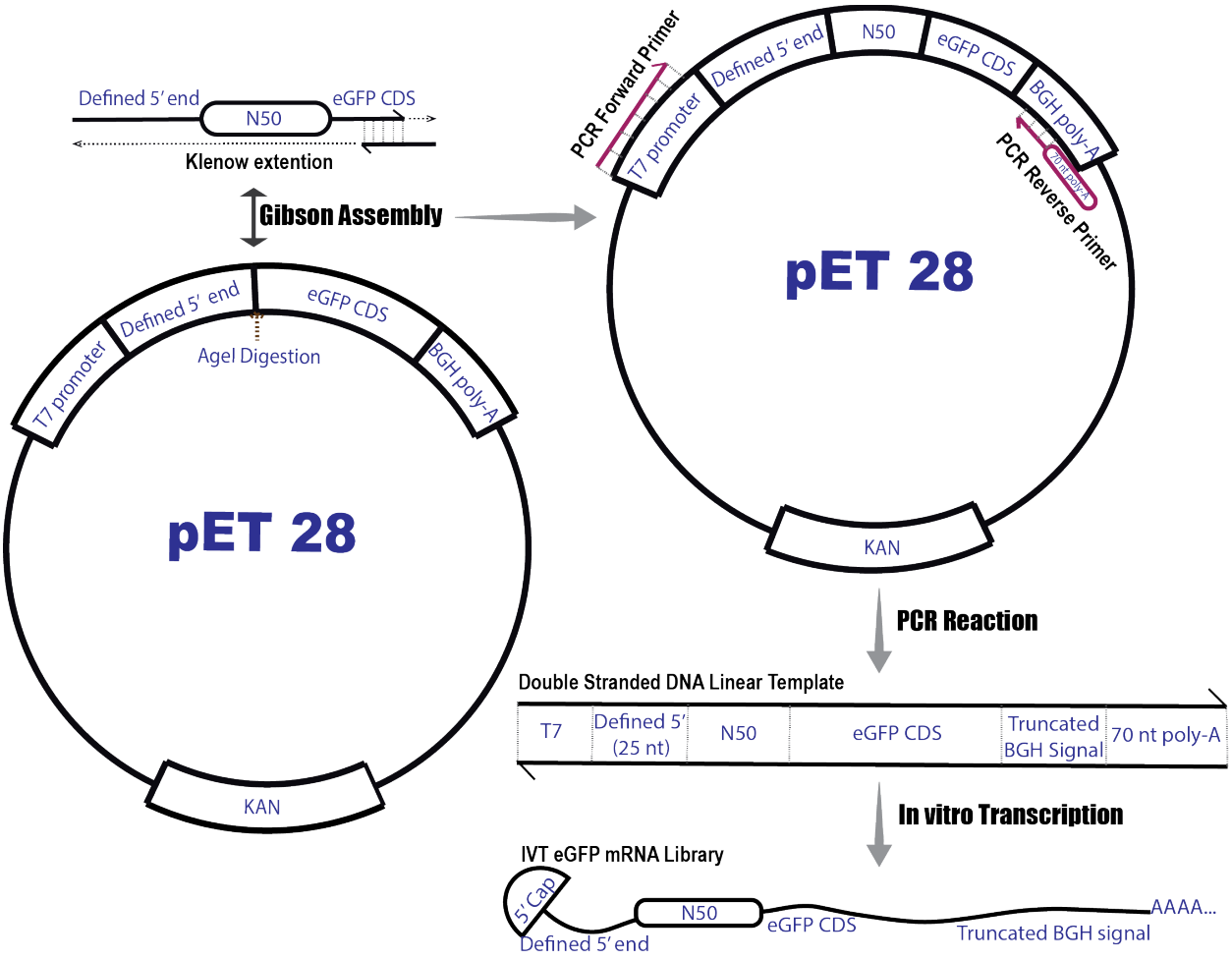


Figure 3.2. Overview of synthetic random library construction.

template to produce mRNA library while a 5' cap analog (m^7G(5')ppp(5')G RNA Cap Structure Analog from NEB) was added into the *in vitro* transcription in the same time to produce capped mRNA. Denaturing urea polyacrylamide gel electrophoresis (PAGE) was used as RNA diagnostic gel to check the correct resulting length of IVT mRNA. We made the unmodified eGFP and mCherry IVT mRNA libraries using this protocol. The 3 Gs at the very 5' end of both eGFP and mCherry library sequence were adapted in the process of *in vitro* transcription from T7 promoter. The detailed eGFP and mCherry sequences were shown in Table 3.2, and bold indicated the defined 5' end of the 5' UTR. The 50-nucleotide oligomer random UTR immediately followed.

The underlined sequence corresponded to a truncated BGH poly(A) signal and followed by a 70-nucleotide-long poly(A) tail which was not shown in Table 3.2. After purification, the mRNA was diluted in citrate buffer to the desired concentration. For mRNA libraries containing alternative uridine, UTP was replaced with pseudouridine-5'-triphosphate or utilized Cap1 to increase mRNA translation efficiency, and the pseudouridine and m1pseudoridine eGFP IVT mRNA libraries were synthesized courtesy of Moderna.

3.2 POLYSOME PROFILING BASED MPRA

We performed our studies in HEK293T cells to study 5' UTR dependent translation in the context of mammalian cells. HEK293T cells were placed on 10-cm cell culture dishes 24 hours before transfection (between 1 and 2 million cells per plate). Two plates were seeded each time to generate biological replicates. Then, we transfected 14.5 μg of library mRNA using Lipofectamine MessengerMAX (Thermo Fisher Scientific) following the manufacturer's protocol to the pre-seeded plate, and the cell density was around 60% to 80% confluency. After 1 hour of incubation with Lipofectamine MessengerMAX, the plates were washed with 10ml 1x Dullbecco's PBS once and replaced with 10 ml new warm media (DMEM with 10% FBS and 1% penicillin-streptomycin). Then after 12 hours of incubation after transfection, cells were lysed. The lysis protocol was adopted from many polysome profiling protocols and optimized by our several rounds of polysome profiling experiments [[44], [48]]. After aspirating cell growth media, 5 ml wash buffer containing 100 μg/ml cycloheximide (Table 3.3) was first added to halt ribosomes, and the plate was then incubated at 37 °C for 5 minutes followed by aspiration on ice. Another 5 ml wash buffer was added to the plate and aspirated thoroughly. Then cells were lysed with 300 μl ice-cold lysis buffer (Table 3.3) added, scraped using cell scraper and broke by pipetting for around 5 times. Wash

buffer and lysis buffer were chilled throughout the protocol. We collected the lysate and placed the tube on ice for 10 minutes. Then the lysate was triturated by passing through a 25-gauge needle 10 times [44] and spun at 16,000 x g for 5 minutes to pellet cell debris and nuclei. The transparent supernatant was collected and added DNase I to a final concentration of 0.005 U/µl to digest all DNA. The lysate was kept on ice for 30 minutes to allow DNase I digestion complete, and then lysate could be stored at -80C or directly proceeded to polysome profiling. All reagents used in this porotocol should be either dissolved in RNAse-free H2O or RNAse-free solvent to avoid RNA degradation.

10x Salt Solution

NaCl	100 mM
$MgCl_2$	100 mM
Tris-HCl pH 7.5	100 mM

Wash Buffer

Cycloheximide	100 µg/ml
DPBS	10 ml

Lysis Buffer

Salt Solution	1x
Cycloheximide	100 µg/ml
20% Triton X-100	1%
DTT	1 mM

SUPERase-In	0.2 U/µl

Sucrose Buffer

KCl	100 mM
$MgCl_2$	10 mM
HEPES pH 7.2	20 mM

Table 3.3. Components of salt solution, wash buffer, lysis buffer and sucrose buffer.

Sucrose gradients were prepared a day in advance to allow fully dissolve of sucrose in sucrose buffer (Table 3.3). Two kinds with different sucrose percentages of sucrose solution (20% and 55%) were well prepared in advance. We dissolved 10 g and 27.5 g RNAse-free sucrose about three days in advance into 50 ml sucrose buffer respectively to get 20% and 55% sucrose solution. In order to make the sucrose gradient, 5.4 ml of 55% sucrose solution was first added to an ultracentrifuge tube, and then 5.4 ml of 20% sucrose was added on top of that drop by drop very slowly and carefully to keep the interface of these two concentrations of sucrose solutions undisrupted. We sealed the tube with parafilm and slowly put the tube down horizontally and left it overnight. The make of sucrose gradient should be performed all in 4 °C cold room. Approximately 2 hours before use, the tube should be returned to up straight position, and 300 µl cell lysate could be loaded carefully on top of the sucrose gradient. Then the tube was ultracentrifuged for 3 hours at 151,000 x g using a Beckman SW-41 Ti rotor. Polysome profiles were generated for each tube and fractions of 500 µl corresponding to ribosome peaks were individually collected and processed using pump, UV detector, and fraction collector [49].

Fractions of 500 µl corresponding to ribosome peaks including the 40S and 60S peaks were individually collected and processed. Five-hundred microliters of TRIzol (Thermo Fisher Scientific) was added to each fraction and the fractions were vortexed. After incubating at room temperature for 5 min, 100 µl of chloroform was added and the mixture was vortexed and then incubated for another 5 min at room temperature. Fractions were spun at 13,000 r.p.m. for 10 min and the RNA from the supernatant was purified following the protocol for RNA Clean & Concentrator (Zymo Research). Elution was performed with 15 µl of RNase-free water. The purified RNA was reverse transcribed (RT) using SuperScript IV (Thermo Fisher Scientific) and gene-specific primers (Table 3.1 - primer 289 for eGFP libraries and primer 220 for mCherry libraries). Both reverse transcription primers had 10-nucleotide unique molecular indices (UMIs). The RT products were then amplified through qPCR with overhangs for Illumina-based sequencing using primers containing barcodes to identify the fraction it belonged to. Then agarose gel extraction was performed to examine the size of PCR products and Sanger sequencing was performed to make sure the purified DNA contained the Illumina sequencing required bases. The products were ready for high-throughput sequencing. A custom forward primer (Table 3.1 - primer 300) for read 1 annealed to the defined 5' end of the 5' UTR. The mCherry library and the eGFP library had the same 5' end sequence. Products were sequenced with the Illumina NextSeq platform using NextSeq 500/550 v2 High Output 75 cycle kits.

We have performed polysome profiling runs on the libraries we have built introduced in Chapter 3.1, including unmodified (U) RNA eGFP, pseudouridine (Ψ) eGFP, 1-methyl-pseudouridine (m1Ψ) eGFP, and unmodified (U) mCherry libraries. For each library, we did two biological replicates, and for each two biological replicates, a correlation figure was generated on the right and the r-squared values were remarkably high suggesting strong reproducibility of our

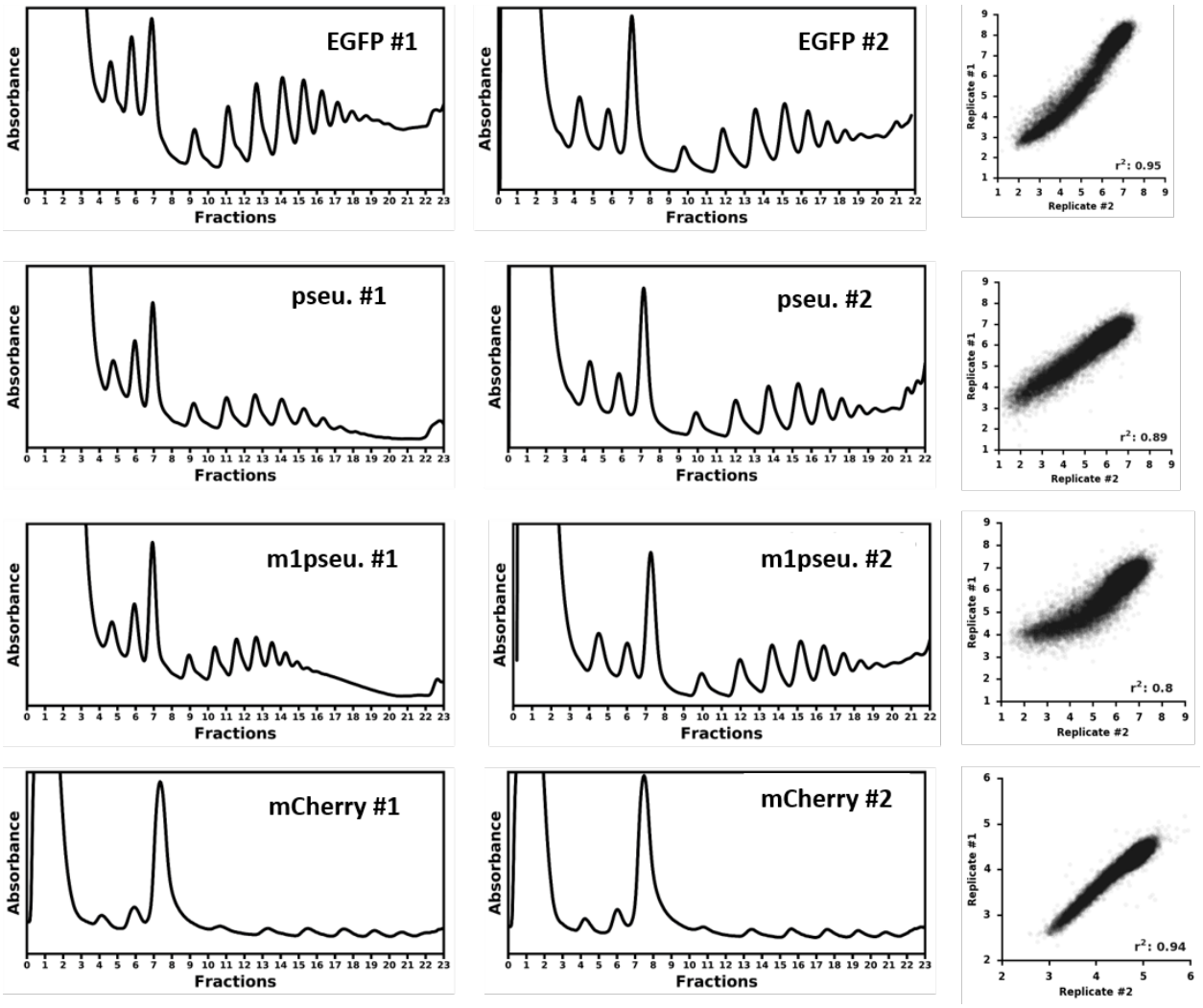


Figure 3.3. Polysome profiles for 4 different libraries with 2 biological replicates for each showing high reproducibility.

experiments (Figure. 3.3). The r-squared values for unmodified (U) RNA eGFP, pseudouridine (Ψ) eGFP, 1-methyl-pseudouridine (m1Ψ) eGFP, and unmodified (U) mCherry libraries were 0.95, 0.89, 0.90, 0.80 respectively. It was easy to observe that polysome profiles for mCherry libraries were not optimal and quite different as the other 6 profiles with much smaller peaks and the correlation between the two biological replicates for mCherry libraries was slightly worse than the other three libraries. The reason for that was this was done using an earlier version protocol of

polysome profiling and we updated our protocol after that with several rounds of optimization experiments. The protocol for mCherry library was carried out slight differently. Specifically, sucrose gradient buffers contained either 7% or 47% (wt/vol) sucrose as well as 150 mM NaCl, 20 mM Tris-HCl pH 7.2, 5 mM MgCl2 and 1 mM dithiothreitol. Sucrose (7%, 5.4 ml) was gently layered over 5.4 ml of 47% sucrose in an ultracentrifuge tube. The tube was ultracentrifuged for 1 h and 45 min at 39,000 r.p.m.

3.3 Mean Ribosome Load (MRL)

Raw sequencing reads, separated by their fraction-associated barcodes, were processed using Cutadapt [50] and Bartender [51] to be grouped and PCR effects were removed through clustering on UMIs. The eGFP library contained approximately 750,000 unique sequences and the mCherry library contained approximately 500,000 sequences. UTRs were removed if the CDS sequence did not match the intended start region of that CDS. Because many of the remaining sequences had very few reads, the top 280k eGFP UTRs and 200k mCherry UTRs were selected based on total number of reads for each UTR. We found that there were no UTR sequences between the eGFP and mCherry libraries were shared.

We transfected our libraries into cells and collected the polysome profile from the cell lysate which showed the cells' translatome profile (Figure 3.4A), while each UTR had its own profile by reading out the counts in all bins (Figure. 3.4B). Three polysome profiles for selective distinct 5' UTR library members were shown in Figure 3.4B representing high, moderate, and low ribosome loads. For high ribosome load profile in red curve, the sequencing reads came more from 'more ribosomes' fractions than 'less ribosomes' fractions, and for low ribosome load profile in blue curve, the sequencing reads came more from 'less ribosomes' fractions than 'more ribosomes'

fractions. Therefore, for each UTR variant, mean ribosome load (MRL) needed to be computed as a measurement for the translation efficiency.

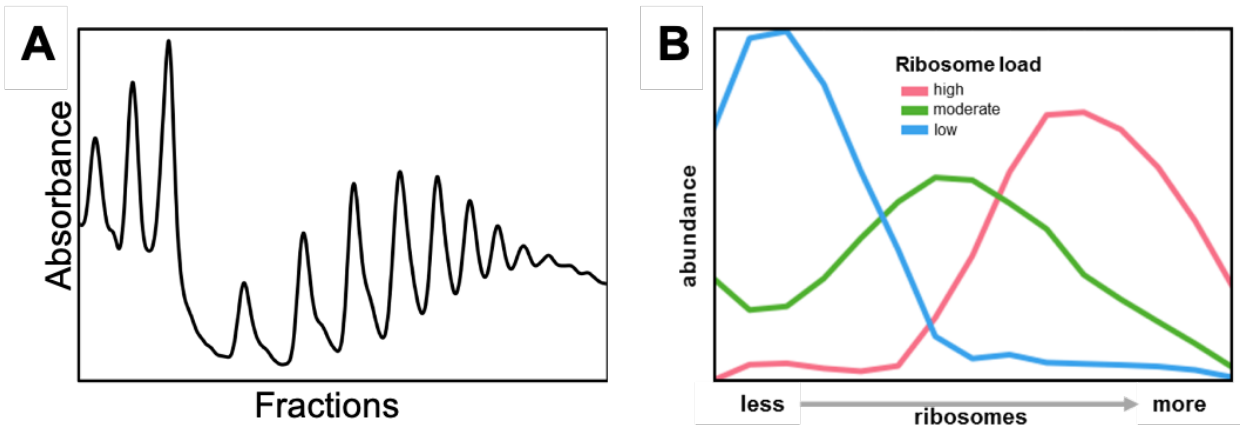


Figure 3.4. Polysome profiles for cells' translatome and selective distinct 5' UTR library members.

Variant	**Fraction-0**	**Fraction-1**	**…**	**Fraction-m**	**…**	**MRL**
UTR-1	R_{10}	R_{11}	…	R_{1m}	…	MRL_1
UTR-2	R_{20}	R_{21}	…	R_{2m}	…	MRL_2
…	…	…	…	…	…	…
UTR-n	R_{n0}	R_{n1}	…	R_{nm}	…	MRL_n
…	…	…	…	…	…	…

Table 3.4. Data collected representing total reads in each fraction for each UTR after high-throughput sequencing and MRL was computed for each distinct library member.

$$\text{Relative-R}_{nm} = \frac{R_{nm}}{\sum_n R_{nm}} \quad \text{(Equation 3.1)}$$

$$\text{Normalized-R}_{nm} = \frac{\text{Relative-R}_{nm}}{\sum_m \text{Relative-R}_{nm}} = \frac{\frac{R_{nm}}{\sum_n R_{nm}}}{\sum_m \frac{R_{nm}}{\sum_n R_{nm}}} \quad \text{(Equation 3.2)}$$

$$\text{MRL}_n = \sum_m (\text{Normalized-R}_{nm} \times m) = \sum_m \left(\frac{\frac{R_{nm}}{\sum_n R_{nm}}}{\sum_m \frac{R_{nm}}{\sum_n R_{nm}}} \times m \right) \quad \text{(Equation 3.3)}$$

In order to deal with the artifacts that different total reads in each fraction for each UTR may bring, we needed to normalize the reads to compute out MRL for each UTR variant. The collected data would be listed as a table where each UTR variant had multiple reads depending on how many fractions were collected (Table 3.4). Note that Fraction-0 included the reads from fractions corresponding to ribosome subunits 40s and 60s.

From the table, each row represented a unique UTR variant, and each column represented read counts collected in that fraction, which meant for each UTR-n, it had reads from R_{n0}, R_{n1}, ..., to R_{nm} etc., where m represented the number of ribosomes presented in that peak in the polysome profile. Firstly, relative reads for each UTR in the fraction (Relative-R_{nm}) would be computed in order to normalize differences in total read counts between fractions (Equation 3.1). Then, the normalized reads (Normalized-R_{nm}) for each UTR would be computed based on Relative–R_{nm} in order to normalize differences in total read counts between UTRs (Equation 3.2). Finally, MRL

would be computed as the sum of product of Normalized-R_{nm} with its corresponding number of ribosomes (Equation 3.3), which would be used as the proxy to protein expression level.

Chapter 4. DATA ANALYSIS AND MODELING

Having collected a dataset with 280,000 datapoints as described in Chapter 3, we needed to perform data analysis first to see if our data was valid or biological meaningful before getting into modeling. The upstream AUGs (uAUGs), upstream open reading frames (uORFs) and Kozak motif were well known regulatory elements in 5' UTR region which could be used as the references in our dataset. We also looked at alternative initiation at non-AUG start codons [[52]–[54]] which other studies reported a widespread usage in addition to AUGs.

A predictive model, Optimus 5-Prime, based on convolutional neural network (CNN) was built to predict regulatory effects from 5' UTR sequence to translation and showed excellent prediction accuracy. Many motifs were visualized from the CNN to provide insight for possible important motifs that worth future studies. The model was also explored to investigate different library sequences contexts including CDS and modified RNA, showing good capability of generalization.

4.1 UPSTREAM AUG AND KOZAK MOTIF

We first did several statistical analyses using UTR sequences and corresponding MRLs to see if they agreed with previous studies. The efficiency of translation can be largely dependent on the existence or absence of in-frame or out-of-frame (OOF) uAUGs and uORFs (Figure. 4.1A). In the library, UTRs without uAUGs and those with uAUGs but are in-frame relative to the designated start codon for the CDS, had higher MRL in general than those with uAUGs but were out-of-frame and also for those containing uORFs (p-values from two-sided t-test are shown). The known Kozak motif [[18], [55]] was noticeable as well, where sequences that encoded a purine at the -3 position

of a start codon had stronger initiation potential than those with pyrimidines. For UTRs without uAUGs and in-frame uAUGs, purine at the -3 position increased the MRL, while for UTRs with out-of-frame uAUGs and uORFs, had purine at -3 position decreased the MRL. P-values from two-sided t-test of impact of purines vs. pyrimidines at -3 position in four groups were computed; No uAUG: $5e^{-103}$, IF uAUG: $8e^{-29}$, OOF uAUG: 0, OOF uORF: 0.

Previous studies have reported that the secondary structures formed in 5' UTR affect translation efficiency negatively [[20], [56]], and our data showed that 5' UTR sequences with higher minimal secondary energy (MFE), calculated by finding the MFE of 20,000 UTRs with top total reads using Nupack [57], outperformed those with lower MFE, which agreed on those findings (shown in Figure. 4.1B and p-values from two-sided t-test were shown).

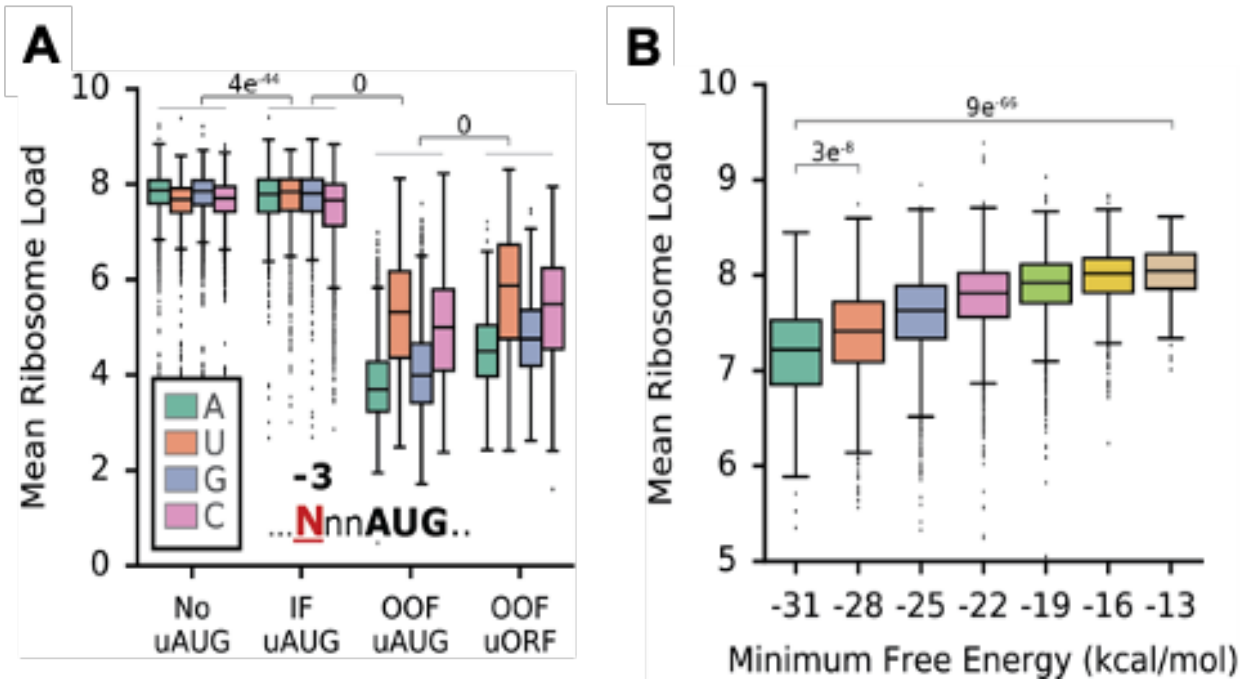


Figure 4.1. Presence or absence of upstream AUGs and minimum free energy showed expected influences on MRL.

Next, we asked if we could observe non-canonical start codon usage based on what has been reported in other studies [[52]–[54]]. On average, we did not see very strong indications that non-canonical start codons were heavily used in the context of our library. This difference was possibly due to these alternative start codons being used more often under stress conditions [58]. However, we found that CUG and GUG start codons could impact ribosome loading, especially when surrounded by strong sequence context as shown in Figure 4.2 and Figure 4.3, where strong TIS (translation initiation sites) context meant a purine at -3 position and a G at +4 position relative to the start codon which was assigned as +1, +2, +3 positions, moderate TIS context meant a purine at -3 position and a non-G (C, U or A) at +4 position, and weak TIS context meant a pyrimidine at -3 position relative to the start codon.

In Figure 4.2A, UTRs were grouped based on the presence of AUG, CUG, GUG, or UUG as the potential start codons between positions -21 through -8 and by the TIS context in which they were found. All AUG TISs dramatically reduced ribosome loading when in an out-of-frame position and the extent of repression was dependent on the strength of the TIS context. For CUG and GUG, there was a minor reduction in ribosome loading when in the in-frame position and within a strong TIS context (p-values from two-sided t-test: AUG weak: 6e-245, AUG Moderate: 0, AUG Strong: 0, CUG Strong: 2e-7, GUG Strong: 5e-3). UUG showed no effect. When analyzing CUG, GUG, and UUG, all sequences with AUG were removed. In Figure 4.2B, the three-nucleotide periodicity explained the differences between in-frame and out-of-frame, which could be easily seen in uAUG data, where out-of-frame uAUGs dramatically reduced the mean of MRL compared to in-frame uAUGs. There were minor three-nucleotide periodicity effects which could be observed for data of CUG and GUG. However, presence of out-of-frame CAG and GAG, single base mismatches of CUG and GUG did not cause reduced mean MRLs.

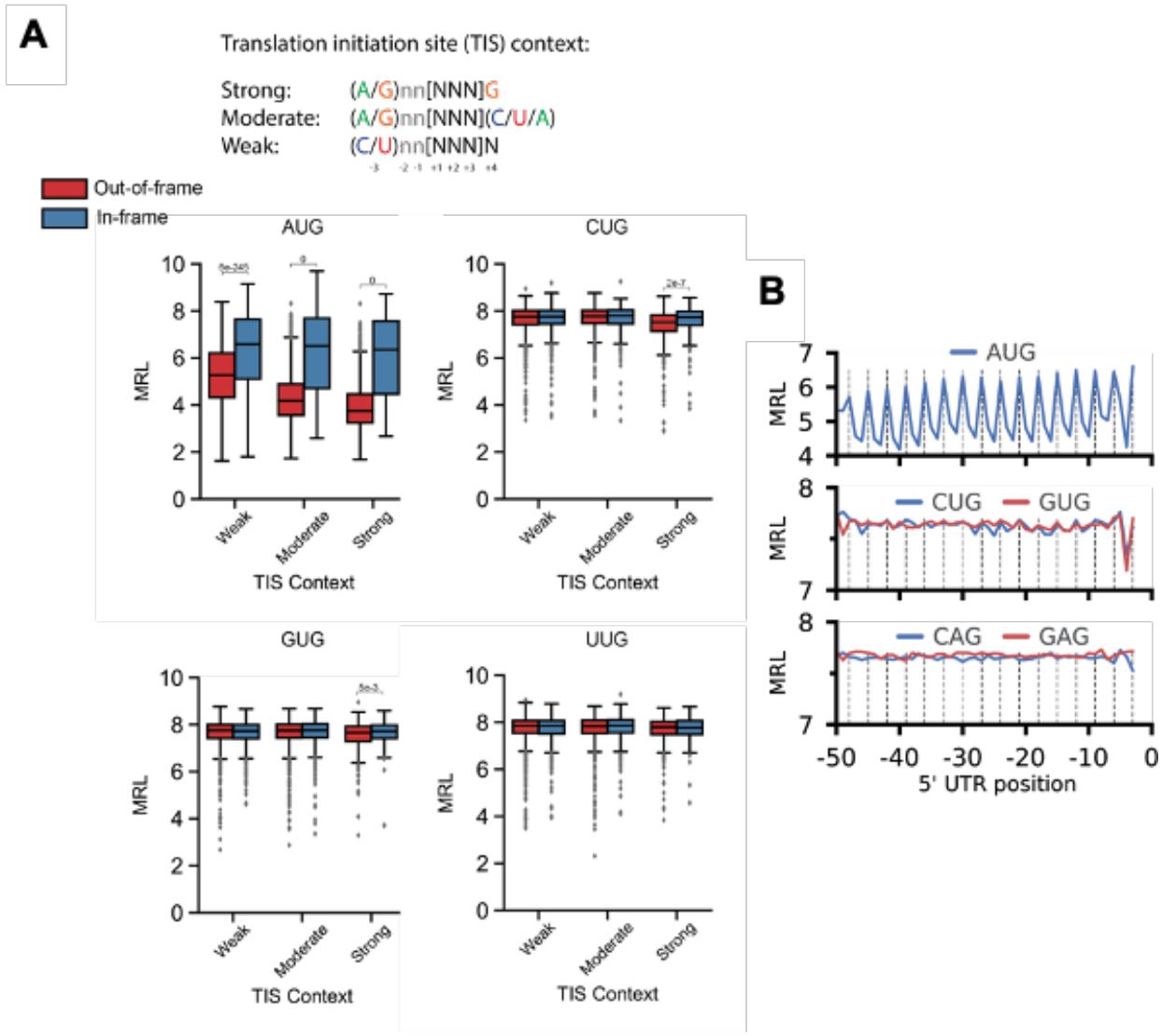


Figure 4.2. Effect of non-AUG translation initiation sites (TIS) on ribosome loading.

Next, we scored the repressive strength of all out-of-frame TISs by finding the mean MRL of sequences with all permutations of NNNAUGNN (except where NNN was AUG) in Figure 4.3A. Using the 20 most repressive and 20 least repressive sequences, we calculated nucleotide frequencies for the strongest and weakest TIS. This analysis recapitulated the importance of a purine (A or G) at position -3 relative to the AUG and a G at position +4 [[18], [55], [56]]. These data together suggested that each TIS sequence could uniquely tune translation initiation to a fine

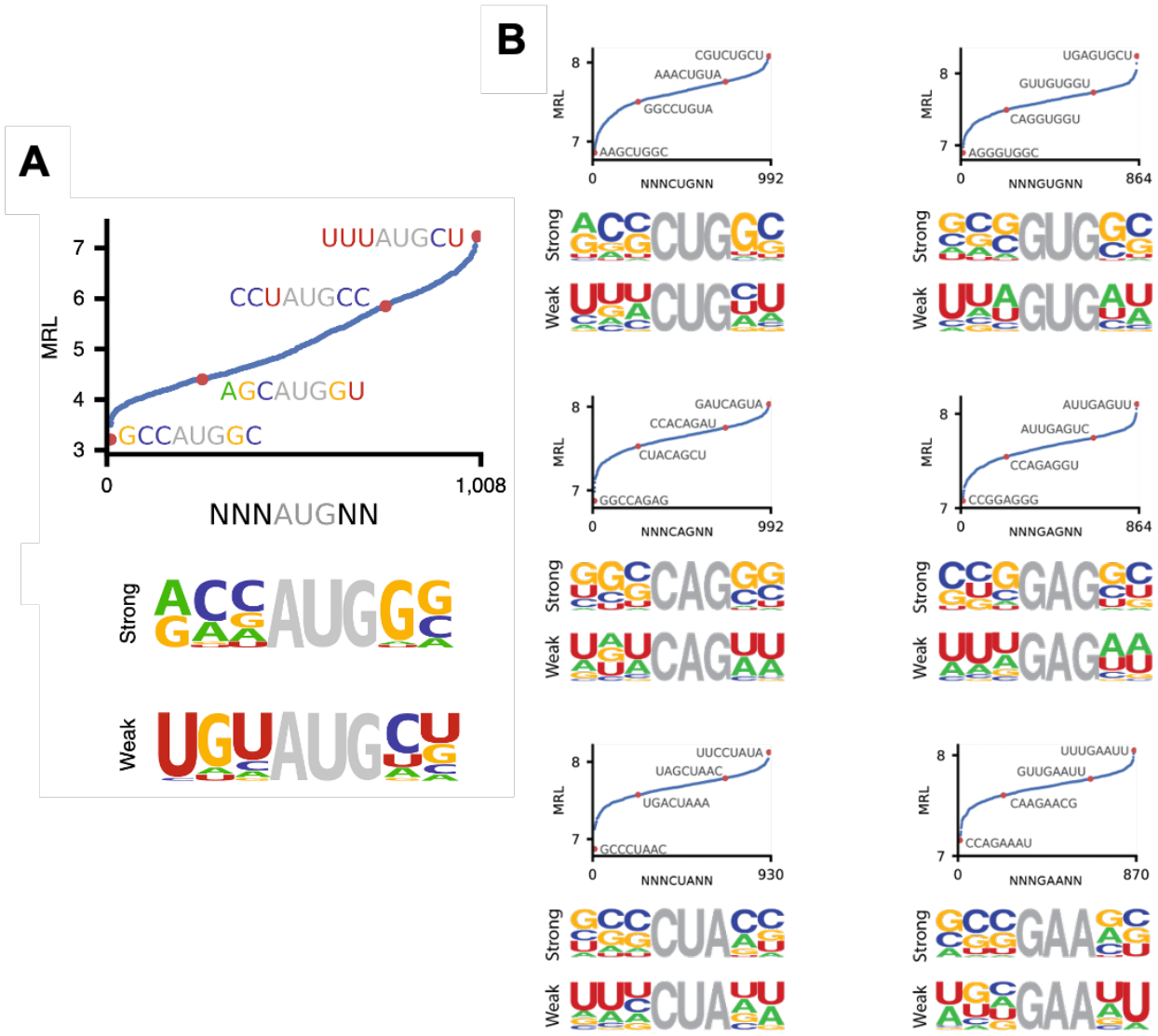


Figure 4.3. The repressive strength of all out-of-frame variations of NNNAUGNN and other non-AUG start codons.

degree. Figure 4.3B shows the repressive strength of surrounding sequences for non-AUG starts, when analyzing non-AUG TIS, all sequences with AUG were removed. The first column in Figure 4.3B shows the repressive strength of all out-of-frame variations of NNNCUGNN, similar as what has been done for NNNAUGNN while here only no uAUG data was used. The single-mismatch 3-mers CAG and CUA were included as controls. The ‘strong’ (most repressive) TIS consensus

sequence matched that of AUG – A/G at -3, CC at -2 and -1, and GC at +4 and +5. The ‘strong’ and ‘weak’ TIS sequences of CAG and CUA were GC-rich and AU-rich respectively, reflecting the repressive nature of GC-rich sequences rather than likelihood of translation initiation. In the second column of Figure 4.3B, the repressive strength of all out-of-frame variations of NNNGUGNN was shown. The 3-mers GAG and GAA were included as controls. The ‘strong’ TIS consensus sequence did not match the pattern of the consensus sequence of AUG and CUG. Like the control 3-mers, the ‘strong’ and ‘weak’ sequences were simply GC-rich and AU-rich, respectively.

4.2 Convolutional Neural Network (CNN) Model

Our libraries consisted of 50 nt randomers and we built them to contain hundreds of thousands of members. For example, the unmodified eGFP library that we used had 280,000 members after initial filtering based on total reads. We could have designed a library with millions of members. Obviously, it was infeasible to build library that could cover all variances that might happen in a 50 nt region since each base has four choices of adenine(A), thymine(T), cytosine(C) and guanine(G), which would yield 4^50 possible sequences. However, we did not necessarily need all 4^50 combinations, and we could apply machine learning algorithms to learn the features from a sub-pool but still randomly distributed sequences to capture important features hidden in the sequence information. Machine learning algorithms, such as linear regression, convolution neural networks (CNN), have been shown to powerfully learn mixed features from numerous datasets. In 2016, the artificially intelligence Alpha-Go, a go player trained using neural network defeated the world champion Go player Lee Sedol in the world by learning from all human go plays, and in 2017, the developing group claimed that they made an even stronger version Alpha-Go to make it

play with itself [38], showing the stunning power of machine learning. Machine learning algorithms are also proven to be very useful in biological perspectives, the two papers introduced using alternative MPRAs also used linear regression [18] and convolutional neural network [33] to build their models.

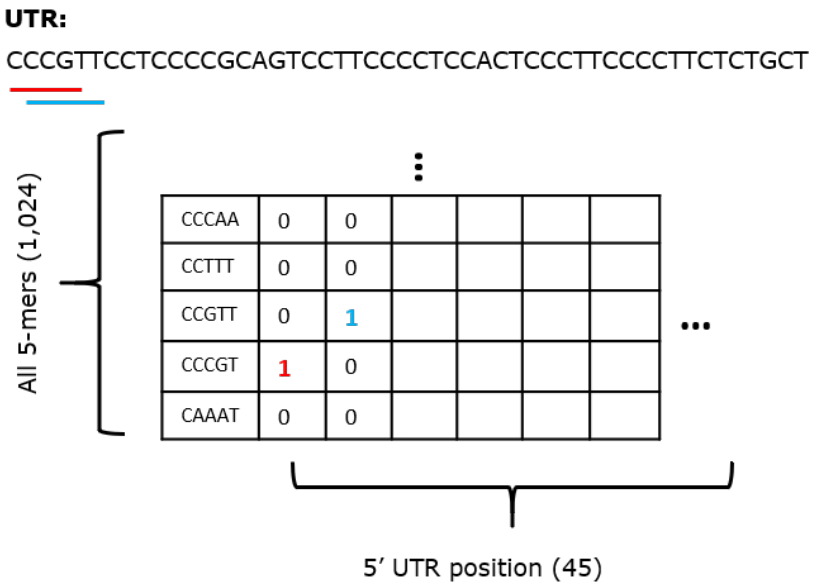


Figure 4.4. A position-specific 5-mer linear regression model.

With such diverse and great amount of data, we wanted to build a comprehensive model explaining and predicting how sequences in 5' UTR regions affect translation using machine learning algorithms. We first started with the *k*-mer linear regression with range from 1 to 6, where UTR sequences were represented as *k*-mers at each position of the UTR. These position-specific *k*-mers were used as features for training a model via linear regression. For example, Figure 4.4 shows the representation of a 5-mer linear regression model. A full list of all possible 5-mers was created first with 1,024 distinct 5-mers, and the UTR sequence was represented as 5-mer at each position from 1 to 45, the length was reduced to 45 since that was the last position a 5-mer could be fully represented. 260,000 sequences with corresponding MRL out of the 280,000-size eGFP

dataset we collected through polysome profiling were used for training and the model was tested on the held-out 20,000 sequences data, where the best performance came from a position-specific 5-mer linear model. UTRs were encoded such that 5-mers and the positions in which they occurred in a sequence served as features for linear regression. Position information was especially important for uAUGs where placement relative to the CDS determined whether they were in-frame or out-of-frame. This position-specific 5-mer linear regression model could explain 66% on the held-out test set (Figure 4.5A). Training involved regularization to limit overfitting and fivefold cross-validation.

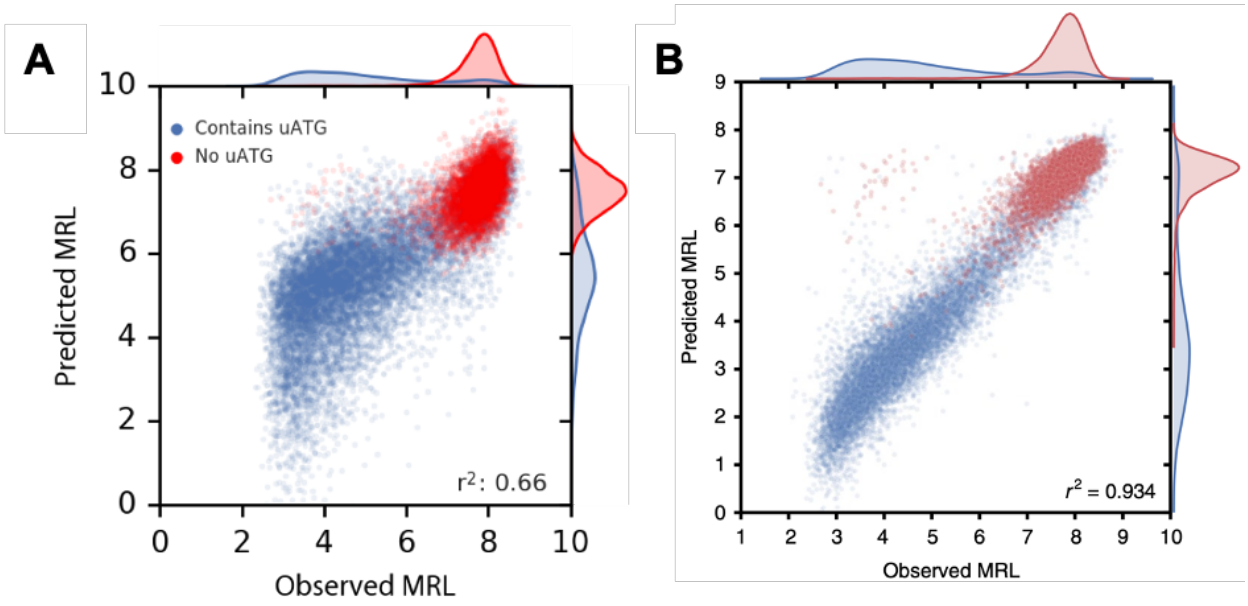


Figure 4.5. Models' performances on held-out 20,000-size test set for linear regression and CNN.

Sequences containing uAUGs in the figure were labeled with blue dots including in-frame and out-of-frame, so they had spread-out distribution with strong and poor translators, and sequences without uAUGs were labeled with red dots and had generally higher MRLs. The r-squared values from testing on the test set for 1 to 6-mer were 0.127, 0.352, 0.628, 0.640, 0.655,

0.561, and 5-fold cross validation found the best regularization (ridge) parameters to limit overfitting. Then we considered CNNs, which could capture important features including linear and non-linear interaction while linear regression would not be able to do. All code was written in Python 2.7 and all neural network development was done using the Keras (https://keras.io) and TensorFlow backends [59].

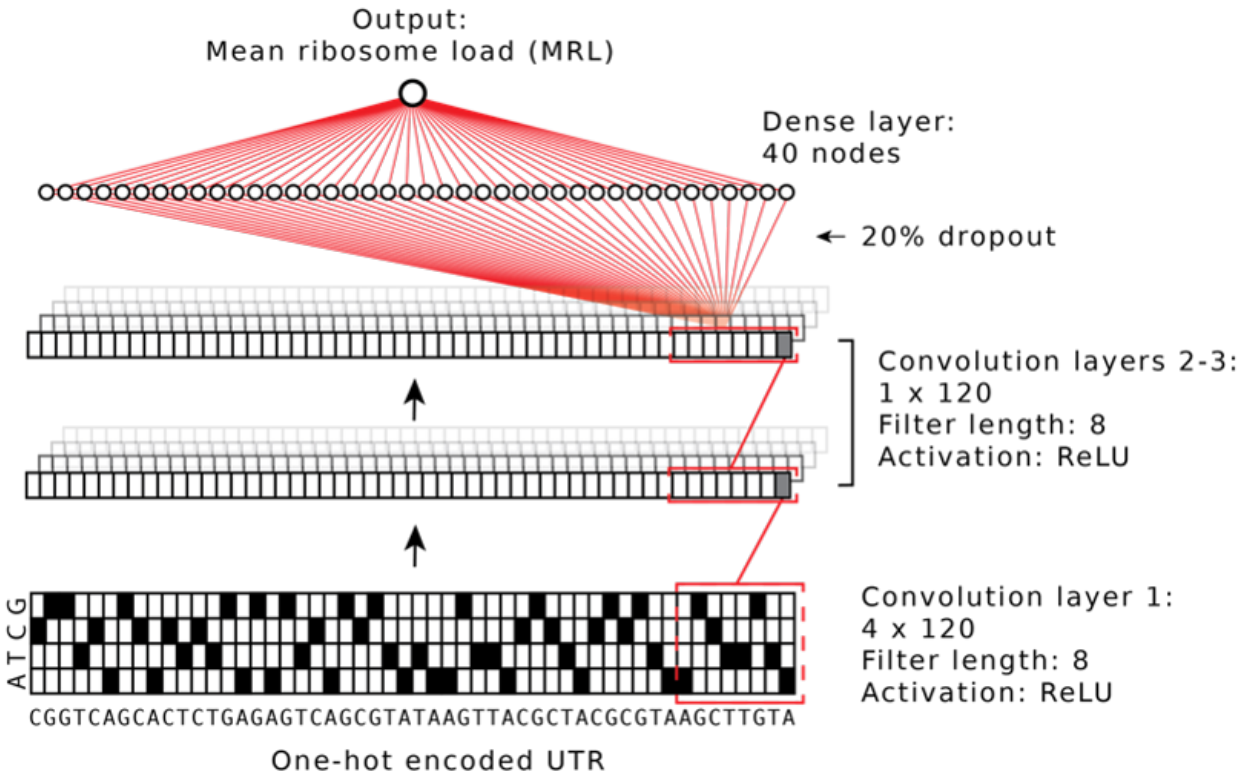


Figure 4.6. CNN architecture of Optimus 5-Prime to predict mean ribosome load (MRL) from 50-mer 5' UTR sequences.

Remarkably, the model we set out called, Optimus 5-Prime could hit the accuracy of r-squared 0.93 (Figure 4.5B) and the two models shared the exact same training and test datasets for better comparison purposes. We used grid search to exhaustively test parameter combinations of convolution layers (2,3), convolution filter lengths (8, 10, 12), number of convolution filters (40, 80, 120), number of nodes in the dense layer (40, 80, 120), dropout probability between all layers

(0, 0.2, 0.4). The best hyperparameter combination we used for the CNN was with 3 convolution layers, and in each layer, there were 120 filters and filter length was 8, ReLU activation and 0% dropout, in the dense layer, there were 40 nodes with 20% dropout and in the output layer, there was only 1 linear output (Figure 4.6).

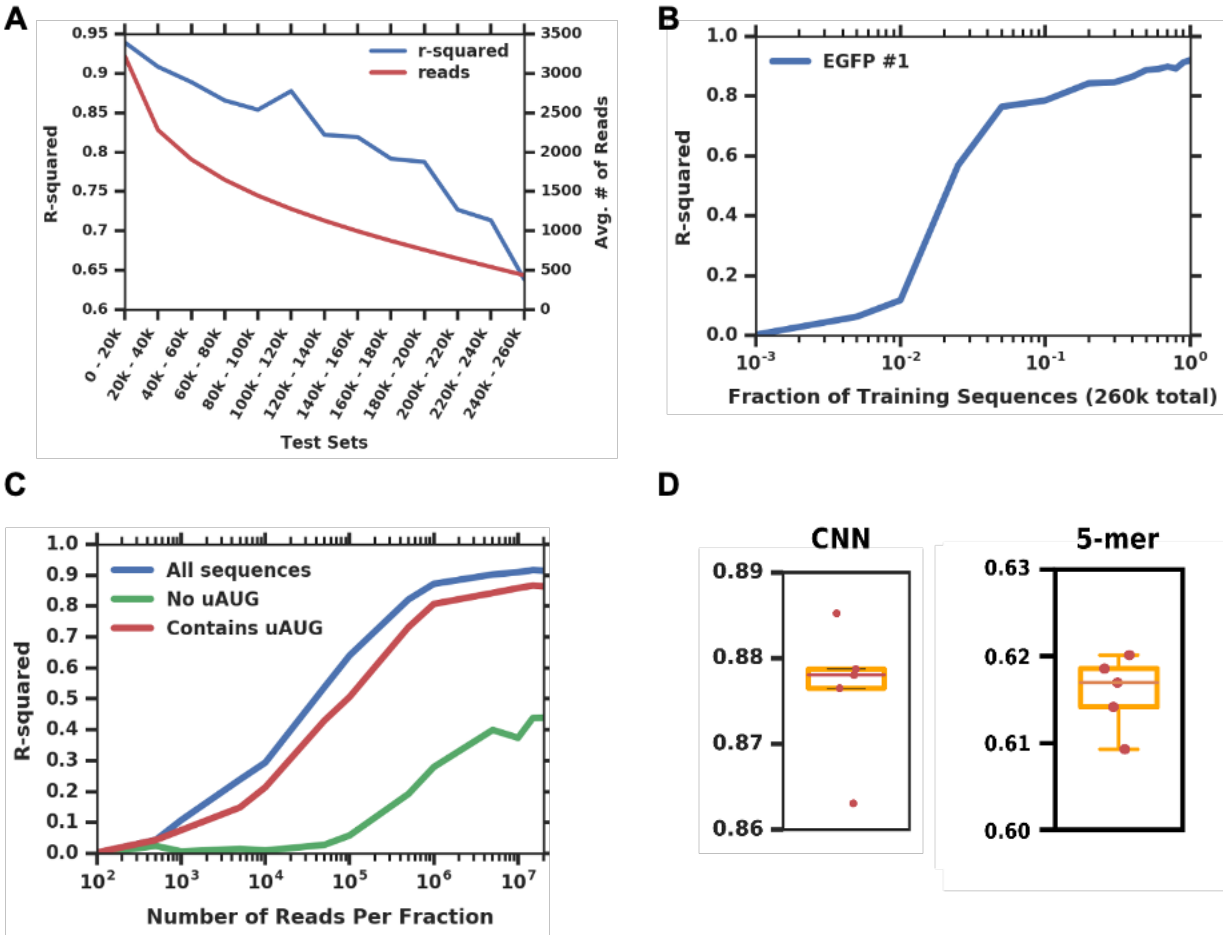


Figure 4.7. The effects for splitting training and test set on the model performance.

Optimus 5-Prime was trained over three epochs before overfitting occurred. Before training, we first sorted the UTRs on the basis of the number of total reads; those with the highest read count were used for the test set. UTRs with more reads had higher resolution and so more accurately

reflected their MRL as compared to UTRs with fewer reads that were noisier. However, the model performed nearly as well after randomly splitting the training and test sets.

In Figure 4.6, how the training and testing set split and data composition in the two sets would affect the performance of the model was shown. Figure 4.6A shows that how the number of sequencing reads per UTR in the test set affects the observed model performance. Sequences were sorted high to low by the total number of sequencing reads across all polysome fractions. A sliding window of 20,000 sequences were used as test sets and the remaining sequences were used for training. The blue curve showed that r-squared dropped when the test set got a slice of lower quality data from over 0.9 to roughly 0.65 and the red curve showed the average number of reads in that testing set with the sliding window. Figure 6.5B shows that model quality decreased as fewer UTRs were used for training. The full set consisted of 260,000 unique 5' UTRs. A dramatic rise in model performance occurred from training on 6,500 (2.5%) to 26,000 (10%) sequences, which was likely due to the model learning the rules of uAUGs and uORFs. Figure 6.5C shows the evaluation of model performance as a function of the number of reads per fractions. Fraction reads were down sampled and used for training and testing. Initial read counts per fraction ranged from 21 million to 33 million except for fraction 13 which had 10 million. The plot suggested that sequences without uAUGs required more training examples than those with uAUGs. Though from the three subfigures above, we could see that the choice of testing and training sets affected r-squared which was treated as the metrics for our model's accuracy. However, Figure 6.5D shows five models were trained and tested independent through CNN (left) and 5-mer linear model (right) using randomly selected subsets of sequences, and the median r-squared for CNN was 0.878 and the median r-squared for 5-mer linear model was 0.617 suggesting that we were not strongly biased the r-squared value by picking testing set and training set manually.

4.3 Polysome Profile Model and Fluorescence Data

Up to this point, we used MRL as a simple measure for translation, but the raw data also captured how often a given sequence occurred in each polysome fraction. We thus set out to build a model capable of predicting the full polysome distribution for a given sequence.

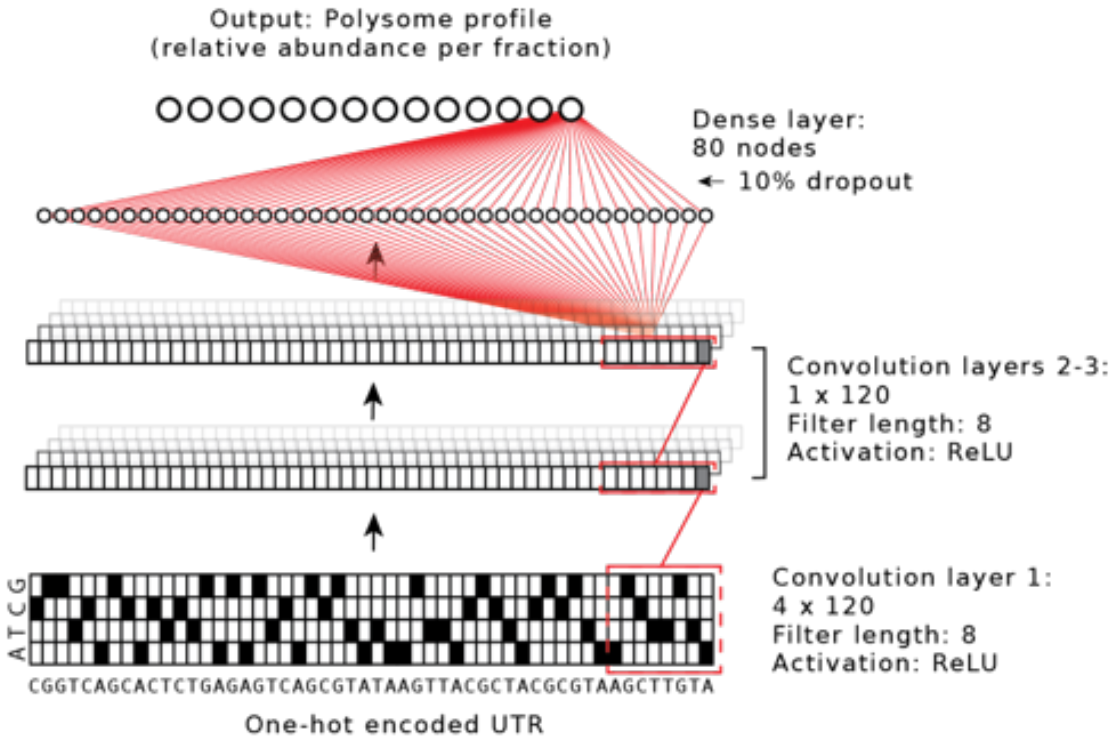


Figure 4.8. CNN architecture for predicting the polysome profile of a given 5' UTR.

This polysome profile model used a similar network architecture but with 14 linear outputs representing the polysome fractions (Figure 4.8). After performing the same grid search as for the model trained to predict the MRL of a sequence, the best hyperparameters for the polysome profile CNN were as follows. First convolution layer: 120 filters (8x4), ReLU activation and 0% dropout. Second convolution layer: 120 filters (8x1), ReLu activation and 0% dropout. Third convolution layer: 120 filters (8x1), ReLU activation and 0% dropout. Dense layer: 80 nodes and 10% dropout.

Output layer: 14 linear outputs. The splitting of the training and test sets was the same as performed in Optimus 5-Prime. As a result, the model differed from the MRL model in the number of nodes in the dense layer (80 rather than 40), the dropout from the dense layer (10% rather than 20%), and the single final linear node was replaced by 14 linear nodes corresponding to each fraction collected during polysome profiling. The output values of this polysome profile would sum to 1 and represented the relative abundance per fraction.

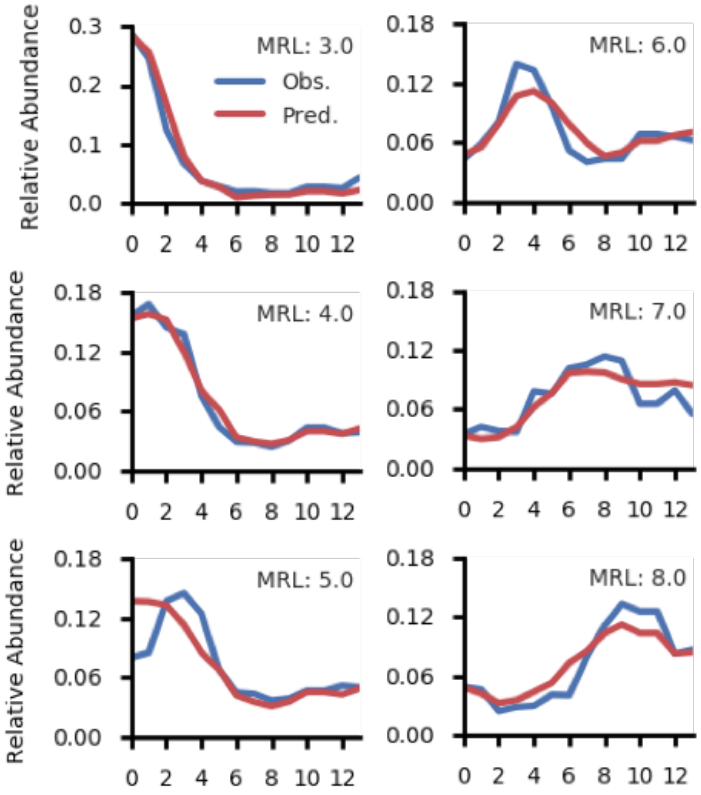


Figure 4.9. Model performance of the polysome profile model on selective examples covering a wide range of MRL values.

The polysome profile model captured the relationship between 5' UTR sequence and the distribution of ribosome occupancy on test data remarkably well as shown in Figure 4.9, where

the predicted polysome profiles in red matched observed polysome profiles in blue very closely. Figure 4.10 shows the model performance of the polysome profile model per fraction with r-squared values ranged from 0.621to 0.915 with an average of 0.83 across all fractions (n = 20,000).

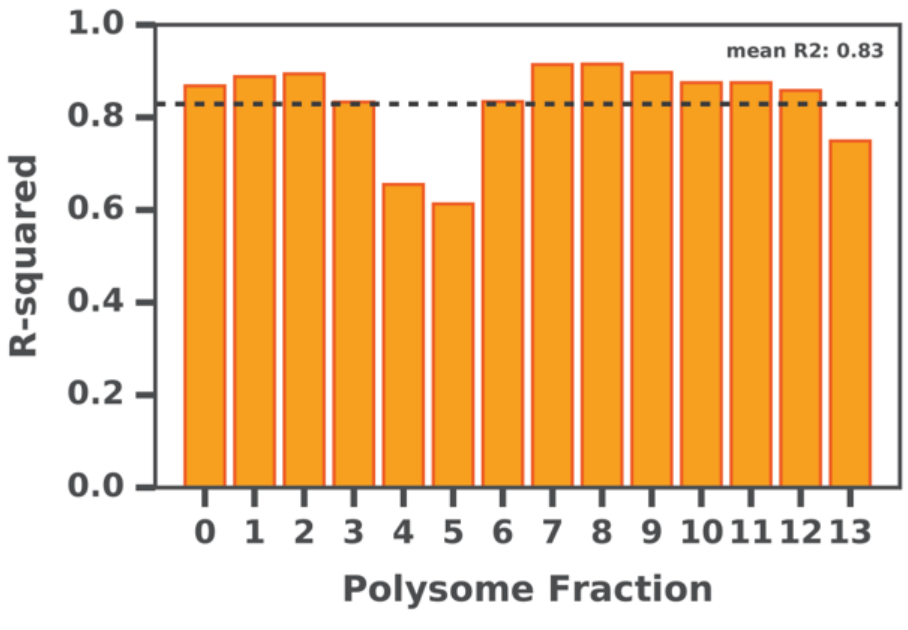


Figure 4.10. Model performance of the polysome profile model per fraction.

More empirically, gene's expression level is determined by the fluorescent signal level when embedded with a fluorescent CDS. Since we still built our library with eGFP as the CDS, although we have demonstrated that one of the several advantages that our assay had was that a fluorescent CDS is not necessary, we picked out ten 5' UTRs from our library and individually cloned them into the same vector as the randomized library to test out their fluorescent protein expression level. IVT mRNA was synthesized and HEK293T cells were transfected with Lipofectamine 2000 and then monitored for eGFP fluorescence using an IncuCyte S3 live-cell analysis system. Expression was reported as the maximum eGFP fluorescence over a 20.5-h time window (n = 3, mean ± s.e.m.). Even though they were picked as they covered a wide range of expression, and the most

poorly translated sequence showed 15-fold-less fluorescence than the most strongly translated sequence, the correlation between the fluorescent level and predicted MRL had r-squared of 0.87 suggesting the MRL prediction matched the actual protein expression well (Figure 4.11).

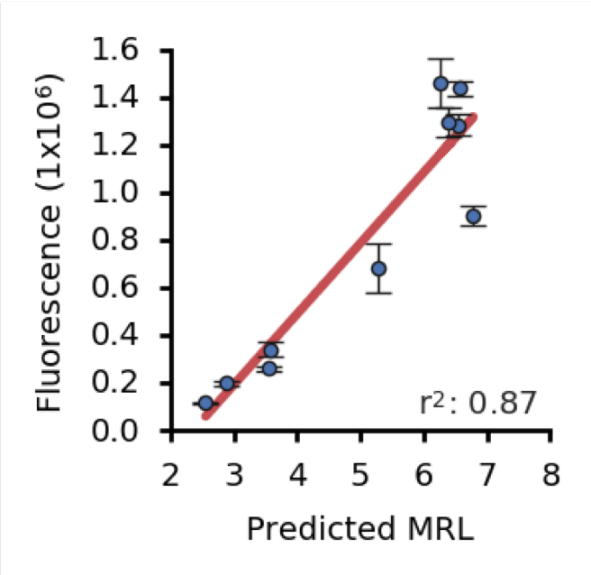


Figure 4.11. Fluorescence signal of eGFP expression for ten UTRs selected from the library was evaluated using IncuCyte live-cell imaging.

We also tested Optimus 5-Prime on 77 5' UTRs that were previously designed and experimentally characterized by Ferreira et al. [60]. Flow cytometry was performed after delivery of reporter constructs to six different cell lines, including human embryonic kidney cells (293T), mouse pre-B lymphocytes (PD31), human chronic myelogenous leukemia cells (K562), human colon cancer cells (HCT116), Chinese hamster ovary cells (CHO-K1) and mouse plasmacytoma (MPC11). UTRs were designed to result in a range of expression levels by inserting one or multiple uORFs. Only 5' UTRs upstream of coding sequence began with 'ATGG' were used for model testing. Sequences were dropped if fluorescence data was missing for some cell lines. UTR length ranged from 3 nt to 47 nt, and sequences were zero-padded on the 5' end to generate 50 nt input

sequences for our model. The MRL predictions from our model correlated very well with the independently reported fluorescence levels with r-squared values ranged from 0.73 to 0.85 shown in Figure 4.10, where UTRs with no upstream ATG (No uATG) are shown in red and UTRs with upstream ATG (uATG) are shown in blue.

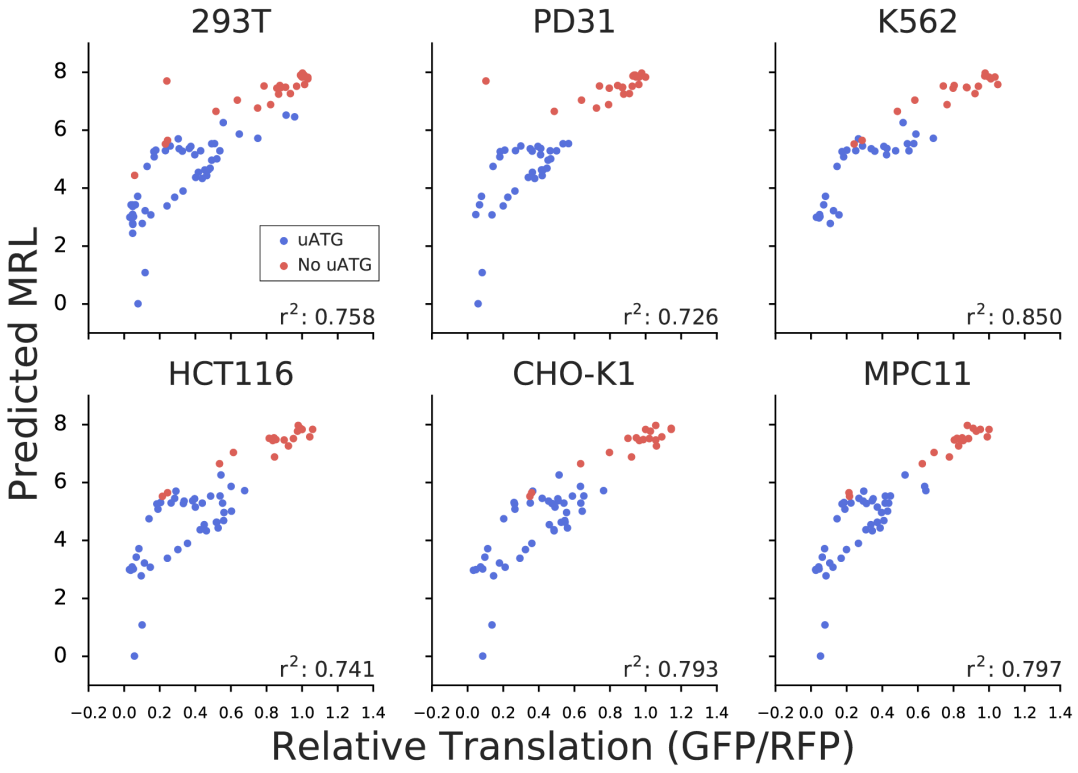


Figure 4.12. Optimus 5-Prime's performance on fluorescence data from another independent study by Ferreira et al. [60] in six cell lines.

4.4 Motif Visualization

One great advantage of using CNN is the ability to visualize the motifs learned by the model. To aid interpretation of the model, we applied visualization techniques to our model which was developed in computer vision and recently popularized in computational biology [[32], [47], [61]].

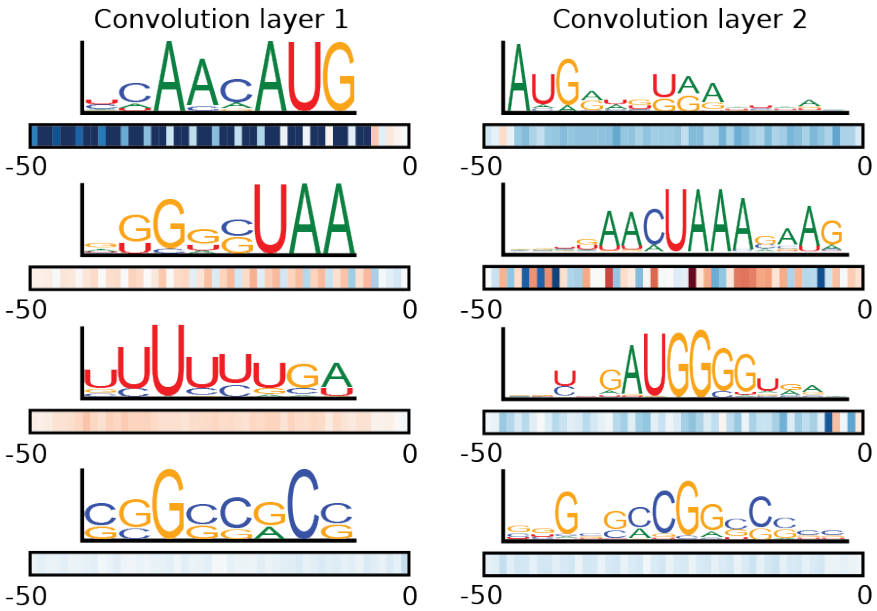


Figure 4.13. Select visualized filters from first and second convolutional layers with recognizable regulatory elements.

For each filter, 2,000 8-mers from the eGFP 5' UTR library that showed the highest activation were selected. From these, PWMs were calculated and used to visualize the sequence compositions that strongly activated each filter. Visualization of the second convolution layer involved a wider sequence window (15 bases) and PWMs were calculated with fewer *k*-mers (maximum 200). After visualizing the first and second convolution layer filters, recognizable motifs were apparent including strong TIS sequences (for example, ACCAUG), stop codons (UAA, UGA and UAG), uORFs, non-canonical start codons (CUG and GUG) and sequences composed of repeated CG or AU elements that were likely involved in the formation of secondary structure.

In Figure 4.13, four out of 120 filters in first convolutional layer were shown on the left and four out of 120 filters in second convolutional layer were shown on the right. Known features such as start codons, stop codons and uORFs could been seen. Boxes below showed correlation

(Pearson r) between filter activation and MRL at each UTR position. For a given filter, the filter's activation at each UTR position was assessed (only the top 100,000 UTRs in terms of total read counts were analyzed). These activations, position by position, were compared to UTR MRLs and a Pearson's *r* value was calculated. Negative values indicated a negative correlation between filter activation and MRL. Positive values indicated that filter activation and MRLs were positively correlated. If UTRs that showed high filter activation had low MRLs then the two were negatively correlated. Many filters did not fall into either of these categories and also did not match previously described position–weight matrices (PWMs) for RNA-binding proteins (Tomtom [62] and the *Homo sapiens* RBP database [63]), suggesting that there were also many unclear motifs that have not been classified and could be possible undiscovered regulatory interactions. See Appendix A for a full list of 120 filters from first convolutional layer and 120 filters from second convolutional layer.

4.5 Model Generalization and Specificity

In order to learn whether the model would be able to generalize to other CDS, we have built an independent mCherry library and performed polysome profiling on this library. The protocol for mCherry library polysome profiling was different from the protocol on other libraries and collected poorer quality of polysome profiles as described in Chapter 3.2 (Figure 3.3). The mCherry library shared identical transcript structure with eGFP library but having no common UTR sequences between these two libraries. Collected dataset on mCherry library again was sorted based on total number of reads – the same preprocessing as on eGFP library. The top 20k UTR sequences with most total number of reads were used as test set while the rest were used as training set. Since

eGFP library size was 300k and mCherry library size was 200k, the size of training set for mCherry library was slightly smaller than eGFP library training set.

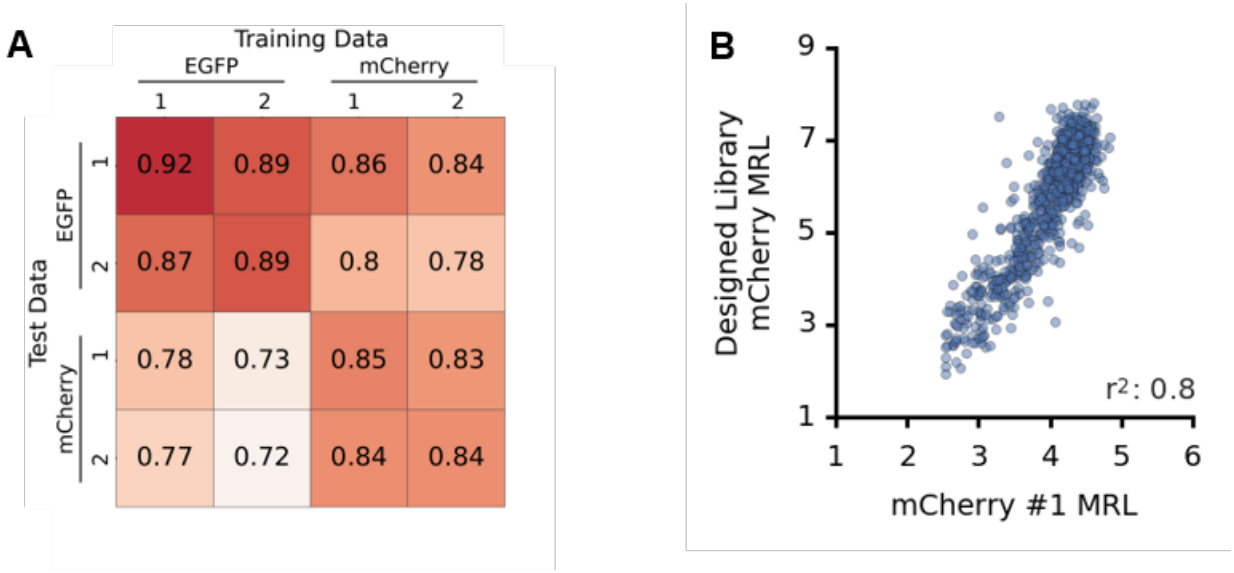


Figure 4.14. Model generalization between coding sequences – eGFP and mCherry.

Model performance was evaluated as which training set was used and tested on which test set and for each library, we had two replicates datasets (Figure. 4.14A). The r-squared values were in the range of 0.87 to 0.92, when model was trained on eGFP library training set and test on eGFP library test set, while for mCherry library, the r-squared values were in the range of 0.83 to 0.85 when training and testing from itself. The poor dataset quality was account for this accuracy lost. The quality of the mCherry data was poorer resulting in a noisier test set which made the eGFP model perform poorly on mCherry. However, the models trained on mCherry worked best when predicting the higher quality data from eGFP 1 and ePFP 2, even better than predicting on mCherry test set. We have tested another 2,000 UTRs from the eGFP library but with eGFP replaced with mCherry. Polysome profiling experiment showed that the MRL between the eGFP and mCherry

data were highly correlated with r-squared value as 0.8 (Figure 4.14B) suggesting that the decrease accuracy in mCherry data did due to the poor quality of data.

Next, we applied our MPRAs to modified RNA, more specifically, pseudouridine (Ψ) and 1-methylpseudouridine ($m^1\Psi$) (Figure 4.15A). RNA modifications may have unexpected interactions with RNA-binding proteins, alter recognition of 5' UTR motifs, and form secondary structures that could either enhance or diminish ribosome loading and translation initiation and efficiency. Nowadays, they were broadly used in RNA therapeutics because they could increase mRNA stability and help modulate the host immune response [[64], [65]].

Using the same T7 eGFP library double-stranded DNA as the template for *in vitro* transcription for unmodified eGFP IVT mRNA library, the pseudouridine and 1-methylpseudouridine eGFP IVT mRNA libraries were synthesized courtesy of Moderna. After performing the same MPRAs described for unmodified eGFP IVT mRNA library, datasets for two modified RNA libraries were collect. The method for evaluating model performance was the same for CDS generalization between eGFP and mCherry. Datasets were sorted based on total number of reads first. Since unmodified, Ψ and $m^1\Psi$ IVT mRNA libraries were from the same template, common sequences were selected for all three libraries to enhance accuracy of comparison. Then training sets and test sets were split as above for all three libraries. Overall, all training and testing sets combinations performed very well (Figure 4.15B) despite some differences. The unmodified models were the best when predicting at unmodified testing set, and the r-squared dropped to a range between 0.68 to 0.76 when tested on modified RNA data, while models trained on Ψ predicted better on Ψ and models trained on $m^1\Psi$ predicted better on $m^1\Psi$.

This could be due to the model learning information on some specific influences Ψ and $m^1\Psi$ bring such as more stable secondary structure which was lacking in unmodified dataset. In

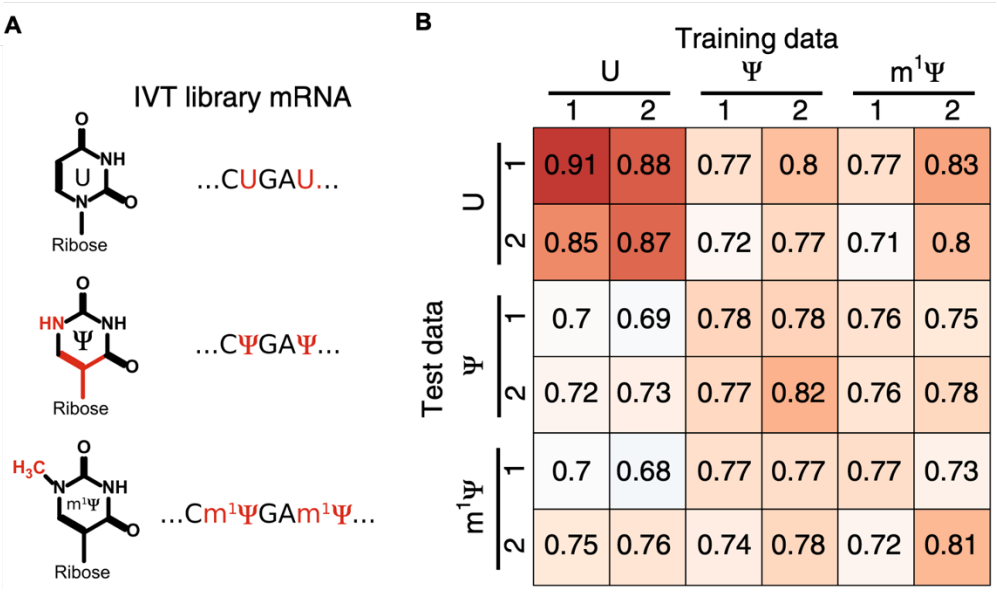


Figure 4.15. Model generalization on RNA modification.

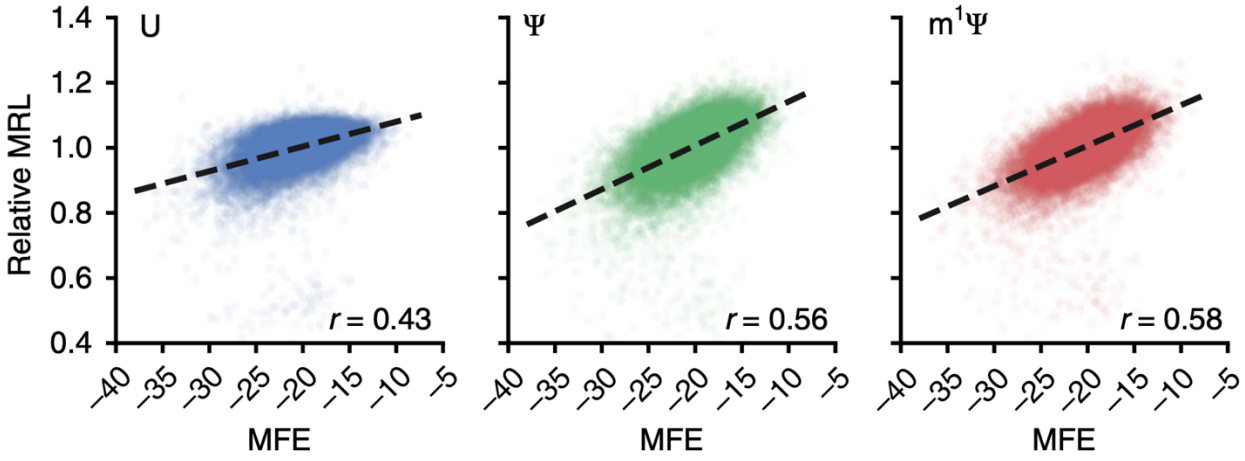


Figure 4.16. Correlation Between MRL and MFE for Three RNA Libraries.

fact, MRL was more positively correlated with a UTR's predicted MFE for Ψ ($r^2 = 0.56$) and $m^1\Psi$ ($r^2 = 0.58$) than unmodified U ($r^2 = 0.43$) (Figure. 4.16). The minimum free energy (MFE) values for 20,000 UTRs from the eGFP library were calculated using Nupack [57] and compared to the

MRLs from the uridine, Ψ and m^1 Ψ datasets. Interestingly, the Ψ and m^1Ψ models performed equivalently well on unmodified test set. This could be due to the high quality for unmodified eGFP data set which was the same result for CDS generalization, or it could suggest that the learned features from modified RNA did not harm the model performance on unmodified RNA.

Chapter 5. APPLICATIONS OF OPTIMUS 5-PRIME

As demonstrated in Chapter 1, we had the strong motivation to build a 5' UTR model because there were a variety of potential applications for this model. From Chapter 3 and 4, with the built Optimus 5-Prime, a CNN based model that could predict mean ribosome loading for any given 50 nt long (upstream of CDS) 5' UTR sequence, we would apply it to multiple fields in this chapter.

In this chapter, we first used Optimus 5-Prime to design new sequences with targeted expression levels and explored the dynamic range of the model. Then, we tested Optimus 5-Prime on human native 5' UTR sequences and single nucleotide variants (SNVs) reported in ClinVar database [66]. The limitation of Optimus 5-Prime that only 50 nt of 5' UTR could be investigated was further demonstrated that by extending the model to a varying length version of Optimus 5-Prime which could take sequence lengths from 25 nt to 100 nt. An online Optimus 5-Prime webtool was built to provide users a fairly convenient way to use our model with no coding experience required, which would expand the potential users to a boarder range.

5.1 FORWARD ENGINEERING ON 5' UTR

The models we have discussed were all built up upon random libraries and predicted arbitrary sequences from the subsets of that random libraries. While we have shown that the model can explain up to 93% of variance for random sequences, we were more interested in whether it could be used to engineer completely new functional 5' UTRs. A tool capable of designing 5' UTRs to obtain specific levels of protein expression would be a valuable asset for mRNA therapeutics and metabolic engineering. While there has been some success in this effort in prokaryotes, yeast and

even mammalian cells [[60], [67]–[69]], a fully rational approach to designing functional 5' UTRs has not yet been implemented.

We developed a genetic algorithm that iteratively edited a randomized 50-mer (not contained in the library of 280,000 sequences) that used Optimus 5-Prime to build a 5' UTR that would load an intended number of ribosomes and thus showed an intended level of translation efficiency. When starting with randomized sequences, over a set number of iterations, a single randomly selected base, or two with a 50% probability, were introduced and the fitness was evaluated using the model. If the new sequence scored higher, or closer to the target mean ribosome load then it was accepted, otherwise, it was rejected, and the unchanged sequence was selected.

We have evolved two sets of UTRs for different purposes, first for hitting target ribosome loads, second for observing the evolution paths and testing the model's dynamic range. For the first set of UTRs, we have evolved three distinct sets of targeted expression: sequences without upstream AUGs (uAUGs) and upstream stops, sequences where uAUGs and upstream stops were allowed, and sequences where uAUGs were not allowed but upstream stops were. Each set evolved initially random sequences to hit mean ribosome loads of 3, 4, 5, 6, 7, 8, 9, and maximum. 200 sequences for 3–7 and 1000 sequences for 8, 9, maximum were selected. In total, including the three sequence conditions, 12,000 sequences were synthesized and tested via polysome profiling.

For the second aim where we wanted to see the evolution path and observe the model's dynamic range, we recorded the sequences for all steps that generating new sequences and tested their performance relative to the model prediction. Four distinct conditions were used and 20 UTRs for each were evolved, totaling 80 UTR stepwise evolution examples. The first two were evolved to the highest ribosome load over 800 iterations; one allowed for uAUGs and the other did not. The third condition evolved sequences to the lowest ribosome load over 800 iterations and then

changed the selective pressure for highest ribosome load over 800 iterations while allowing uAUGs. The fourth condition was the same as the third except that uAUGs were not permitted. In total, beginning with 20 sequences for each condition, 7,526 UTRs were generated for analysis. We observed that sequences containing uAUGs and those without uAUGs could both span the full MRL range. With all the designed sequences which were ordered oligos from CustomArray, we constructed the design IVT mRNA library as we described in Chapter 3. The same workflow of polysome profiling based MPRA was performed on this library (Figure 5.1A).

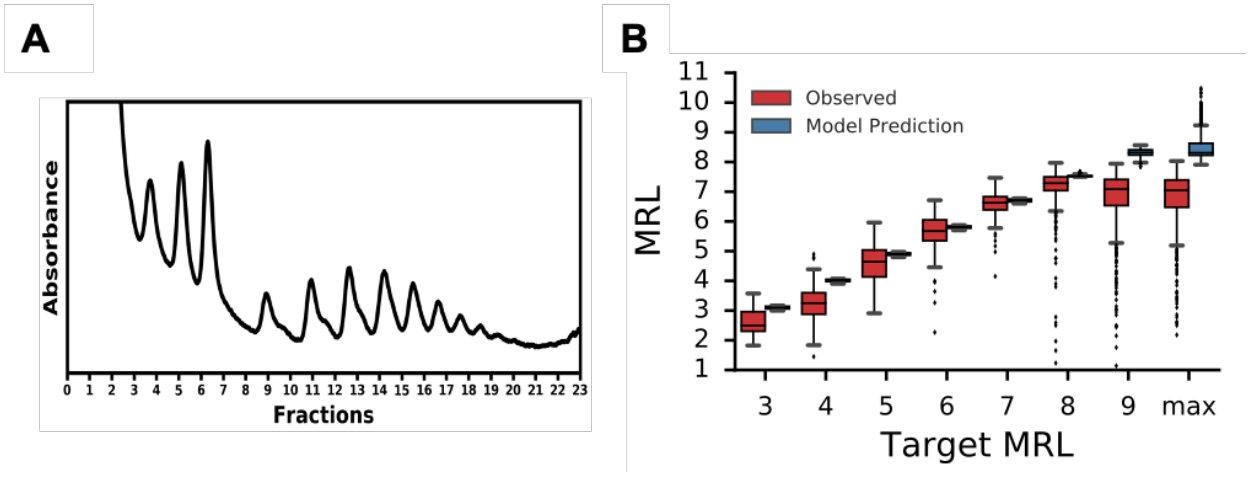


Figure 5.1. Polysome profile for the designed library and observed MRL correlated with model's target prediction well.

For the first set of UTRs, we found that the observed performance was showing the expected order from low to high MRL (Figure 5.1B). The sequences between 5 to 7 were especially accurate, but the trend broke with UTRs that were designed to have MRLs of very high number - 9 and maximum. The reason for this discrepancy would be explained later. For the second set of UTRs, we saw that the observed MRL trend as the sequence evolved matched with what the model suggested very well till the end they diverged, where the model thought it should go higher but the

real performance dropped down. In Figure 5.2, four examples were shown for sequences evolution from original model, retrained model and observed performance. These four examples were from the set where sequences initialized randomly, then iterated for lowest expression then towards maximum expression. The top two examples were no uATGs allowed, and the bottom two examples were uATGs allowed. Performance predicted by original model (green line) closely matched with observed data until to the tails where the original model predicted continuing increase in MRL but the observed performance saturated. See Appendix B for a full list figures of 80 UTR stepwise evolution examples with 20 UTRs for four distinct designing conditions.

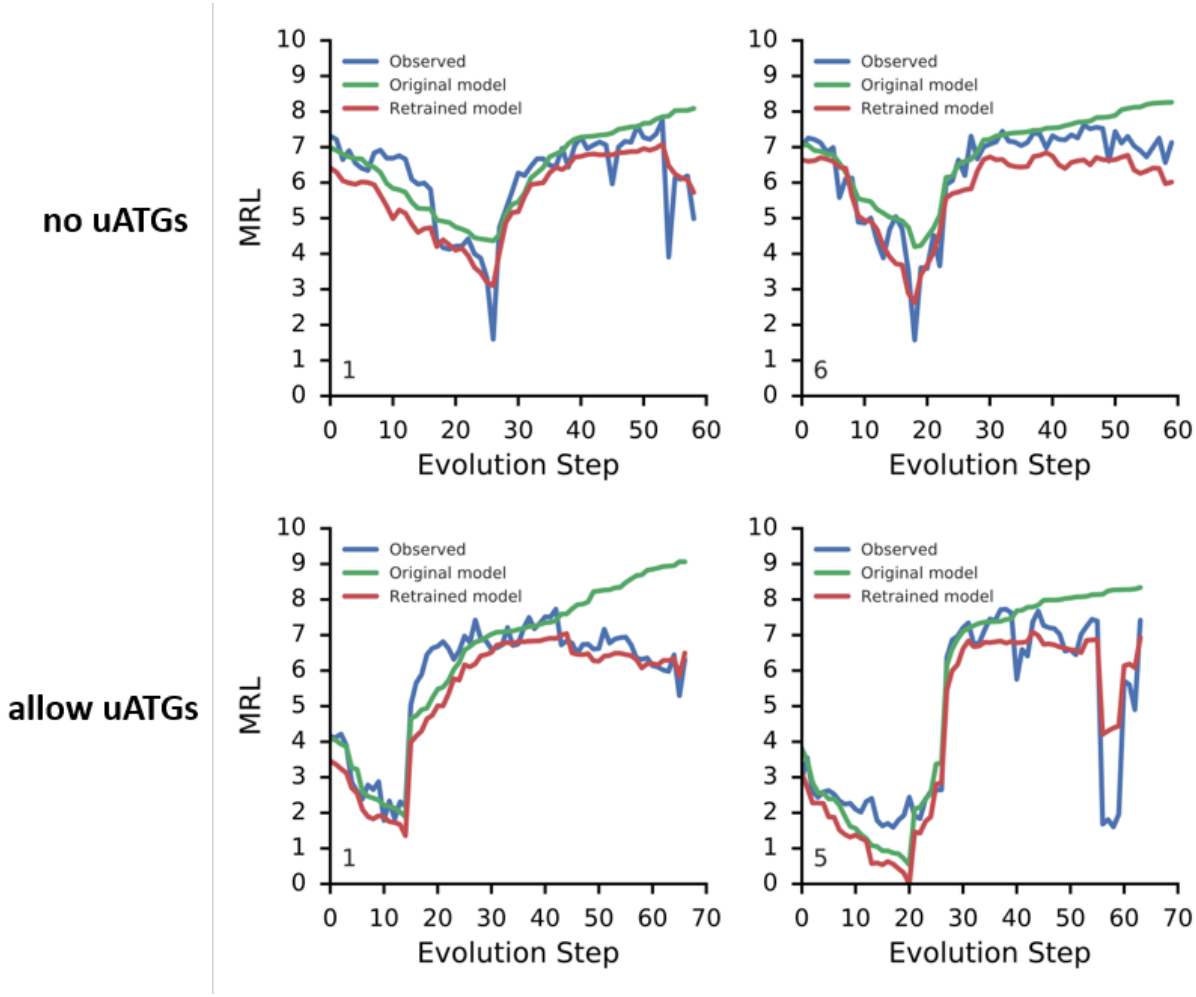


Figure 5.2. Four examples from designed library with different designing constrains.

We highly suspected that this inaccuracy might be a reflection of the unusual sequence composition of the maximally evolved UTRs which often contained multiple long stretches of poly-U. Since the model was trained on random unmodified eGFP library, the chances for the model to see some extreme structures were very rare, thus the model only learned that to have multiple U in sequence was good for higher MRL but could not be able to capture the harm that poly-U brought. With this in mind, we attempted to improve the model by retraining it on sequences from the designed library that had a much wider range of single nucleotide stretches, i.e. sequences having extremely low chance to appear in a random library. This retrained model corrected the original model's limits and increased the accuracy obviously when visualizing the stepwise evolution traces (Figure 5.2 red line).

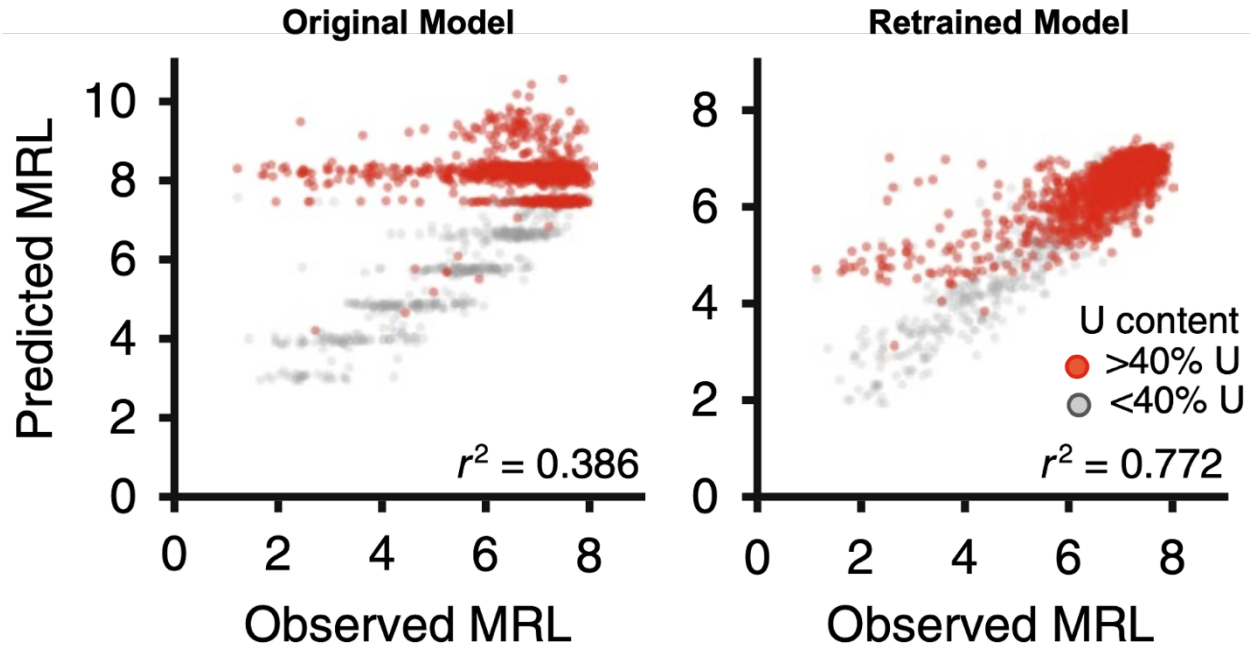


Figure 5.3. The accuracy of retrained model was better than that of the original model when predicting MRL for sequences with a high frequency of poly(U) stretches.

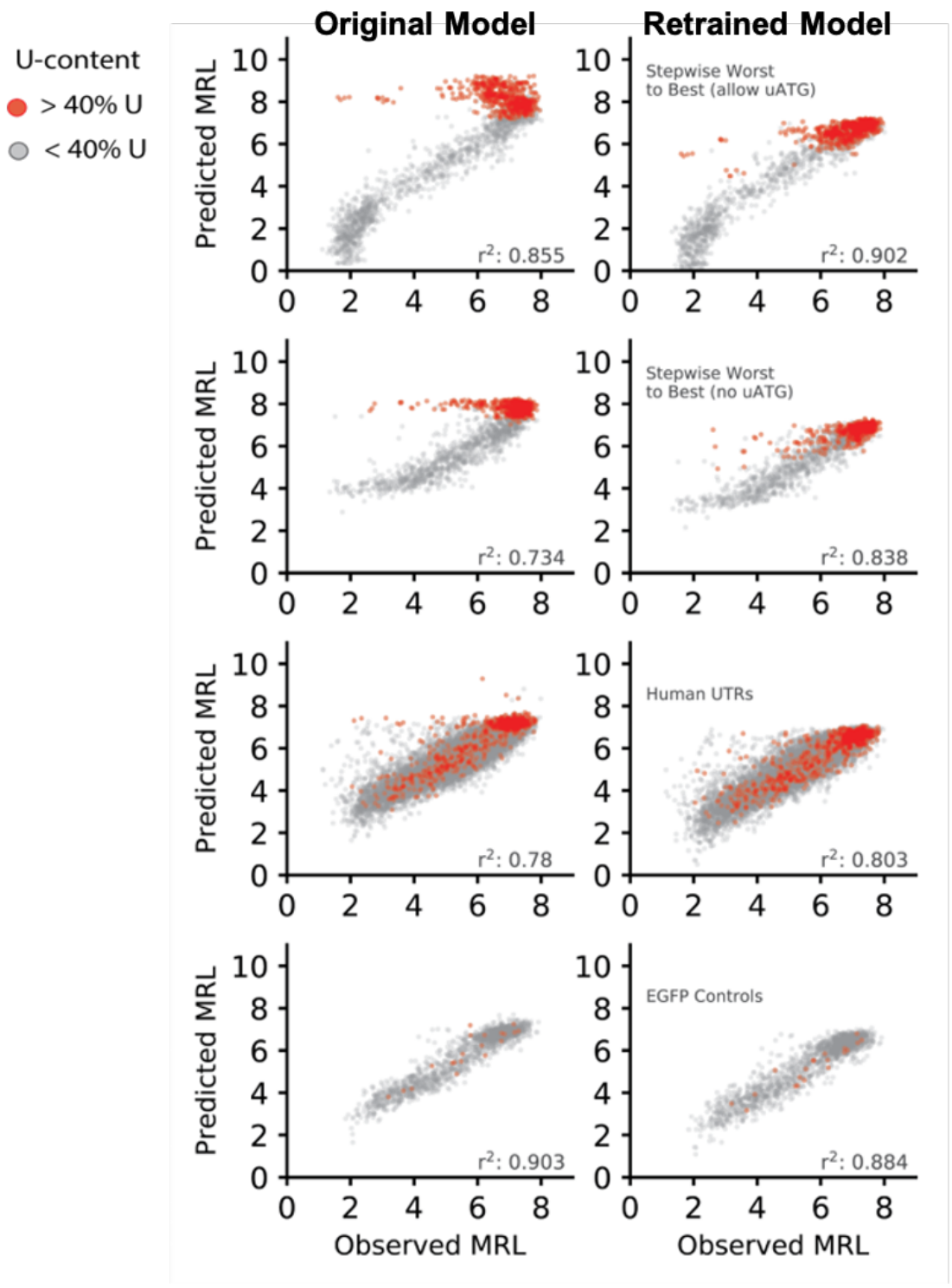


Figure 5.4. Comparing the performance of the original model to the retrained model on sub-libraries.

Quantifiably, the retrained model had a better performance than original model by improving r-squared from 0.386 to 0.772 (Figure 5.3). The most noticeable correction that the retrained model made was that data with high U composition (>40%) was corrected instead of predicting them having high MRL without discrimination. Thus, using this expanded dataset, we retrained the Optimus 5-Prime model, which showed increased accuracy with all sub-libraries shown in Figure 5.4 and the retrained model was used to reevaluate the unmodified eGFP data, and it showed very similar remarkable performance with r-squared at 0.9. Therefore, from this point on, all models referred in this dissertation are the retrained Optimus 5-Prime model.

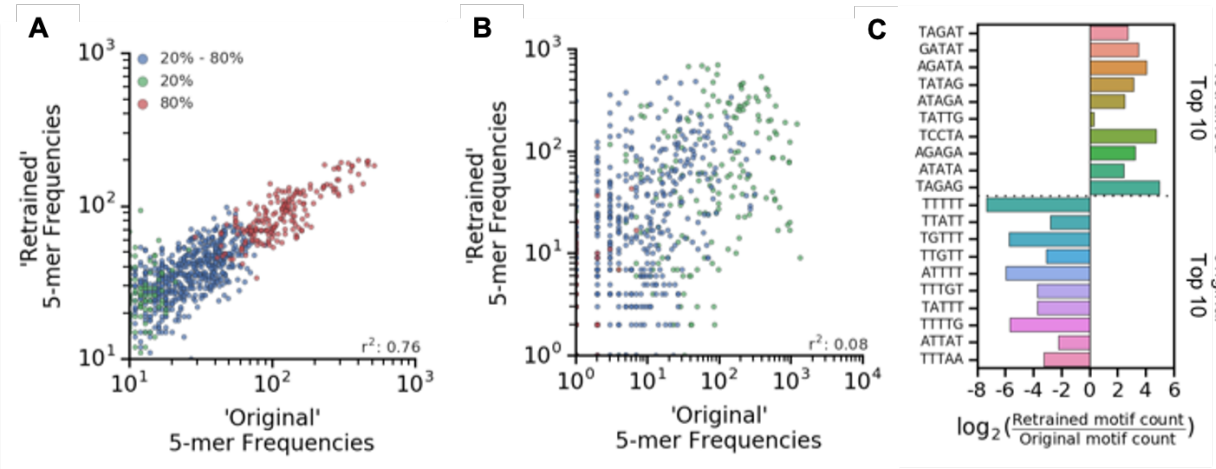


Figure 5.5. Retrained model showed significant different feature than the original model.

To look at motif utilization differences between the original and the retrained model, we evolved 1000 sequences using both models. Figure 5.5A shows that when targeting an MRL of 5, 5-mer usage between the two were correlated with an r-squared of 0.76. Blue dots indicated GC content in 5-mers was in the range of 20% to 80%. Green dots indicated GC content below 20% and red dots indicated GC content above 80%. However, there was basically no 5-mer frequencies

correlation when evolving UTRs for maximum MRL shown in Figure 5.5B. When evolved for maximum expression, Figure 5.5C shows the top 10 5-mer frequencies for retrained model differed from original model. UTRs from the original model used 'UUUUU' nearly 256 times more than the retrained model and all ten of the top motifs in the original model contained three or more T's while only one was found in the top ten 5-mers of the UTRs from the retrained model. This supported the idea that the original model without ever have seen poly-U content in the training data overestimated the positive effect that poly U generated but ignored the harm it might bring. The retrained model learned this point and corrected the poly-U inaccurate prediction.

What we should learn from this is that the deep learning model trained on large random dataset can pick up features that help us understand the regulatory rules, but when using the model for new sequences design, we need pay attention to those rare structures that may be hard to be seen in the random training set, but may harm the new sequence generation due to some unknown effects that the model may strongly bias to, like the poly-U artifacts that was introduced in the original Optimus 5-Prime model.

5.2 HUMAN 5' UTR AND SNVS PREDICTION

The ultimate goal of investigating 5' UTR functions was to learn the mysteries happening in life, and by mastering the rules of its operation, we could know how to control the protein expression level for many applications such as disease therapeutics. This was especially important for single nucleotide variants (SNVs) which played a major role in gene expression variation between individuals [70]. All libraries we have introduced were synthesized libraries including totally random sequences and evolved sequences. We wondered how our model was going to perform on native human 5' UTR sequence and associated SNVs. Assessing model performance on

endogenous transcripts was challenging owing to the confounding contributions of 3' UTRs and CDSs. As an alternative approach to serve this goal, we took the first 50 nucleotides preceding the annotated transcripts which included transcript isoforms, as well as SNVs sequences from the ClinVar databases [66] that occurred within these regions. All human 5' UTR transcripts from the human genome, as annotated by Ensembl [71], were retrieved using Biomart [72]. The first 50 nucleotides upstream of the annotated TISs were selected for synthesis, totaling 35,212 sequences. All sequence variants in the ClinVar database [66] occurring in the selected UTR regions were synthesized, totaling 3,577 sequences. All sequences were ordered through CustomArray Inc. together with designed sequences introduced in Chapter 5.1 and integrated into IVT mRNA library and ran through the same MPRA described in Chapter 3.

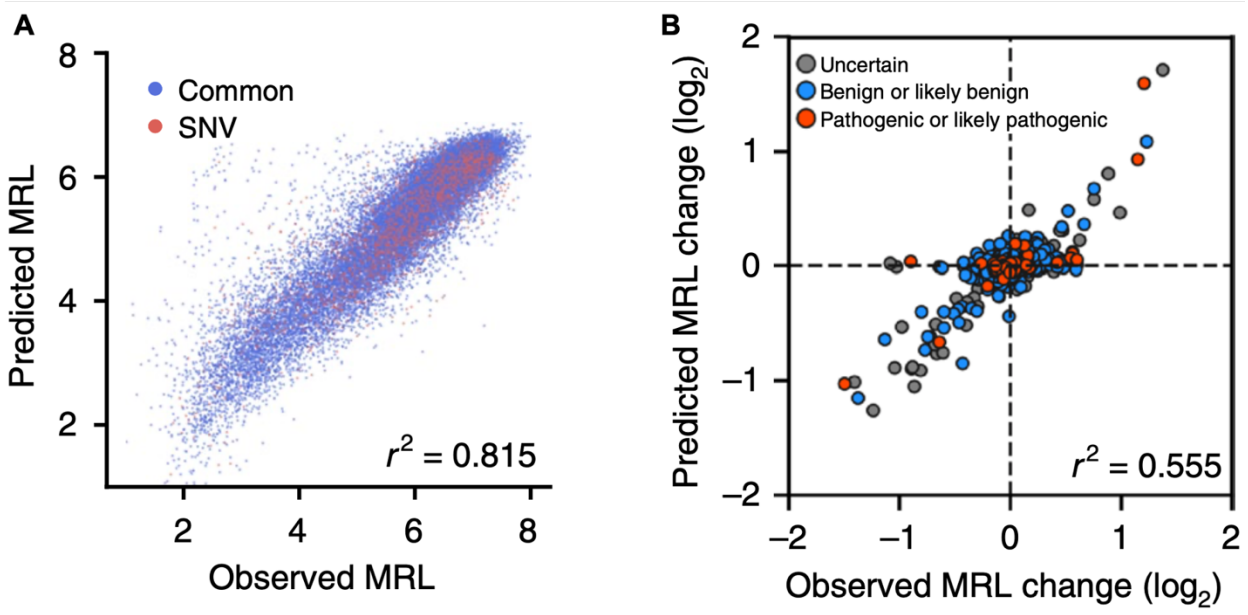


Figure 5.6. Optimus 5-Prime prediction on human native 5' UTRs and SNVs.

After sequencing, the top 25,000 sequences by read coverage, which includes 22,747 common and 2,253 variant sequences, were used for downstream analysis. Using the retrained Optimus 5-

Prime model, we were able to predict 82% MRL variance of this dataset including native sequences (blue dots labeled as 'Common') and SNVs (red dots) (Figure 5.6A), suggesting that despite being trained on random sequences, the model was able to learn the *cis*-regulatory rules of human 5' UTR sequences that lay directly upstream of a CDS.

Genetic variants play a major role in phenotypic differences between individuals [70], and how these sequences affect translation is only beginning to be understood [[73], [74]]. However, existing approaches to this problem, such as quantitative trait locus analysis and genome-wide association studies are limited to common variants and cannot scale to the enormous number of rare 5' UTR variants occurring in the human population. In contrast, a model-based approach can in principle be used to score the impact of any 5' UTR variant on translation. With this in mind, we investigated the ability of the Optimus 5-Prime model to predict the effects of disease-relevant variants by testing its performance when predicting the difference in MRL for pairs of wild-type ('common') and SNV-containing 5' UTR sequences. We assessed this by taking the log2 difference in the observed MRL between an SNV sequence and its corresponding common sequence and compared it to the predicted log2 difference between the two. The majority of SNVs had little to no effect, but 45 had log2 differences greater than 0.5 or less than -0.5 (Table 5.1), and 30 variants out of 45 causing significant MRL changes were due to insertion or deletion of uAUG while the other 15 variants impacted ribosome loading through other mechanisms.

Table 5.1 provides detailed information for the identified 45 SNVs with gene name, phenotype, significances provided by ClinVar dataset, the experimental validated MRL differences between common and SNVs in log2 scale and the dbSNP for reference. Overall, the r-squared for predicting log2 differences was 0.555 (Figure 5.6B). The model could accurately predict the direction of change for 64% of the variants. The relatively lower predictive accuracy

as compared to direct prediction of variant effects was a consequence of the increased noise that from comparing two measurements. Moreover, a majority of variants did not affect translation, resulting in a large cluster of variants for which the difference in MRL change was close to zero where measurements were dominated by noise. Importantly, the model could explain 77% of the variance for variants with measured log2 MRL differences of greater than 0.5 or less than -0.5 in comparison to the common sequence (Figure 5.7A). We also identified 2,308 additional SNVs resulting from errors in oligonucleotide synthesis and found that 103 of them showed log2-transformed MRL changes of greater than 0.5 or less than -0.5 (Figure 5.7B).

Gene	**Phenotype**	**Significance**	**Obs.Diff. (log_2)**	**dbSNP**
GNPAT	Rhizomelic chondrodysplasia punctata	Uncertain	-1.41	rs201907247
CSTB	Unverricht-Lundborg syndrome	Uncertain	0.56	rs776181852
GALT	Deficiency of UDPglucose-hexose-1-phosphate uridylyltransferase	Pathogenic/ Likely	0.56	rs111033654
MLH1	not specified	Uncertain	-0.9	rs1016433173
MSH6	not specified	Uncertain	-0.81	rs1064793670
BBS7	Bardet-Biedl syndrome	Uncertain	0.62	rs757523715
GALT	Deficiency of UDPglucose-hexose-1-phosphate uridylyltransferase	Pathogenic/ Likely	0.54	rs111033656
GALT	Deficiency of UDPglucose-hexose-1-phosphate uridylyltransferase	Pathogenic/ Likely	0.6	rs111033656
COA6	not specified	Benign/ Likely	-0.62	rs73099933

PDE6C	Achromatopsia, Cone-Rod Dystrophy, Recessive	Uncertain	-0.72	rs374900090
TCTN3	not specified	Benign/ Likely	-0.81	rs41291572
HSPB1	Charcot-Marie-Tooth, Type 2, Distal hereditary motor neuronopathy	Uncertain	0.59	rs372833436
TARS2	not specified	Benign/ Likely	1.23	rs201336268
ZMPSTE24	Lethal tight skin contracture syndrome, Mandibuloacral dysplasia	Uncertain	-0.67	rs200527699
SMAD4	not specified	Benign/ Likely	-0.61	rs1057523754
CHRNA4	Autosomal dominant nocturnal frontal lobe epilepsy	Benign/ Likely	-0.77	rs200259564
TMEM127	Pheochromocytoma	Pathogenic/ Likely	-1.5	rs121908813
PNPO	Pyridoxal 5'-phosphate-dependent epilepsy	Uncertain	-0.98	rs886053100
SMAD3	Loeys-Dietz syndrome 3	Pathogenic/ Likely	1.21	rs587776882
LZTR1	not specified	Benign/ Likely	-1.14	rs370616172
CTSA	Combined deficiency of sialidase AND beta galactosidase	Benign/ Likely	-1.38	rs116893852
TP53	Sarcoma	Uncertain	-0.64	rs137852791
CPOX	Hereditary coproporphyria	Uncertain	-0.89	rs867711777
MPDU1	Congenital disorder of glycosylation	Uncertain	0.75	rs370389790
PEX12	not specified	Uncertain	-1.04	rs727504080

MLH1	Hereditary cancer-predisposing syndrome, Lynch syndrome	Uncertain	-0.68	rs587779001
SMAD3	not specified	Uncertain	1.37	rs863223756
SLX4	Fanconi anemia	Uncertain	-0.61	rs113023461
POLE	not specified	Uncertain	-0.68	rs1064796567
MAP2K2	not specified	Benign/ Likely	0.66	rs1057520422
NBN	not specified	Benign/ Likely	-0.75	rs730881843
SYNE2	Emery-Dreifuss muscular dystrophy	Benign/ Likely	0.52	rs199566869
UQCRB	not specified	Benign/ Likely	0.75	rs373747569
PDHX	Pyruvate dehydrogenase complex deficiency	Benign/ Likely	-0.6	rs2956112
HSPB1	Charcot-Marie-Tooth, Type 2, Distal hereditary motor neuronopathy	Uncertain	0.51	rs199602956
RPL5	Diamond-Blackfan anemia	Uncertain	-0.87	rs376208311
ETHE1	Ethylmalonic encephalopathy, not specified	Uncertain	0.98	rs138958351
FOXRED1	not specified	Uncertain	-1.24	rs778239850
MKKS	Bardet-Biedl syndrome, McKusick Kaufman syndrome	Uncertain	-1.03	rs886056499
TTC19	not provided	Pathogenic/ Likely	-0.9	rs769078093
PHEX	Familial X-linked hypophosphatemic vitamin D refractory rickets	Uncertain	-1.09	rs1057515841

C19orf12	Neurodegeneration with brain iron accumulation 4	Uncertain	0.88	rs186970109
SMPD1	Niemann-Pick disease, type A	Pathogenic/ Likely	-0.64	rs875989837
PRPH2	Choroidal Dystrophy, Cone-Rod Dystrophy, Fundus albipunctatus, Vitelliform macular dystrophy	Benign/ Likely	0.59	rs114062933
ACADM	Medium-chain acyl-coenzyme A dehydrogenase deficiency	Pathogenic/ Likely	1.15	rs1057516778

Table 5.1. 45 ClinVar [66] variants with MRL changes in log2-transformed values greater than 0.5 or less than -0.5.

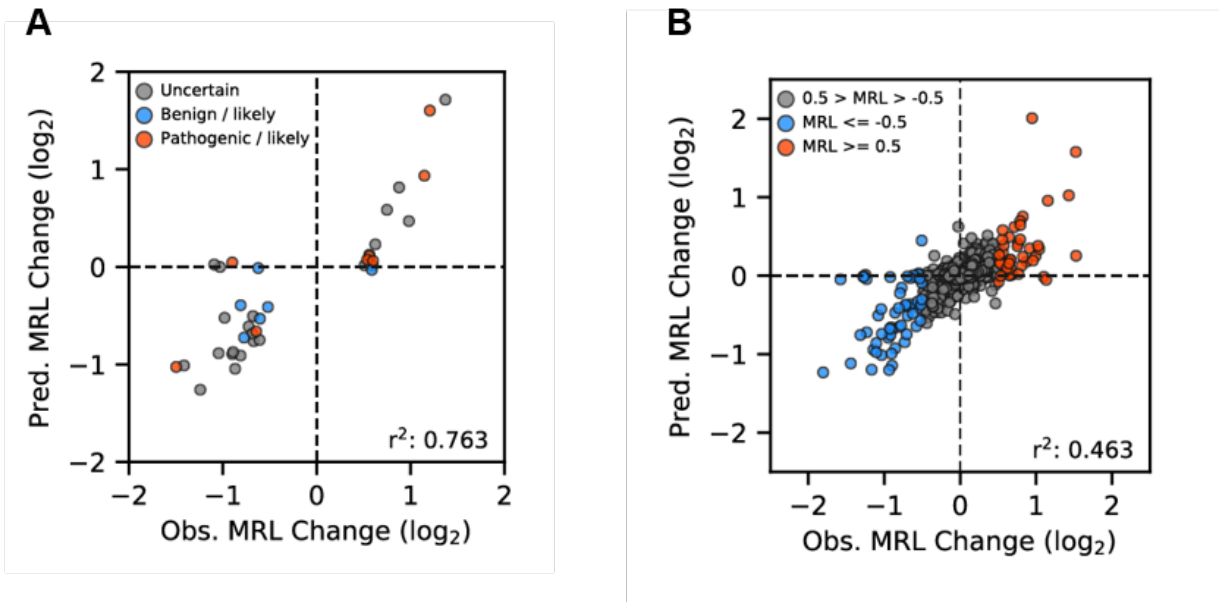


Figure 5.7. Optimus 5-Prime prediction on differences of pairs of common and SNVs.

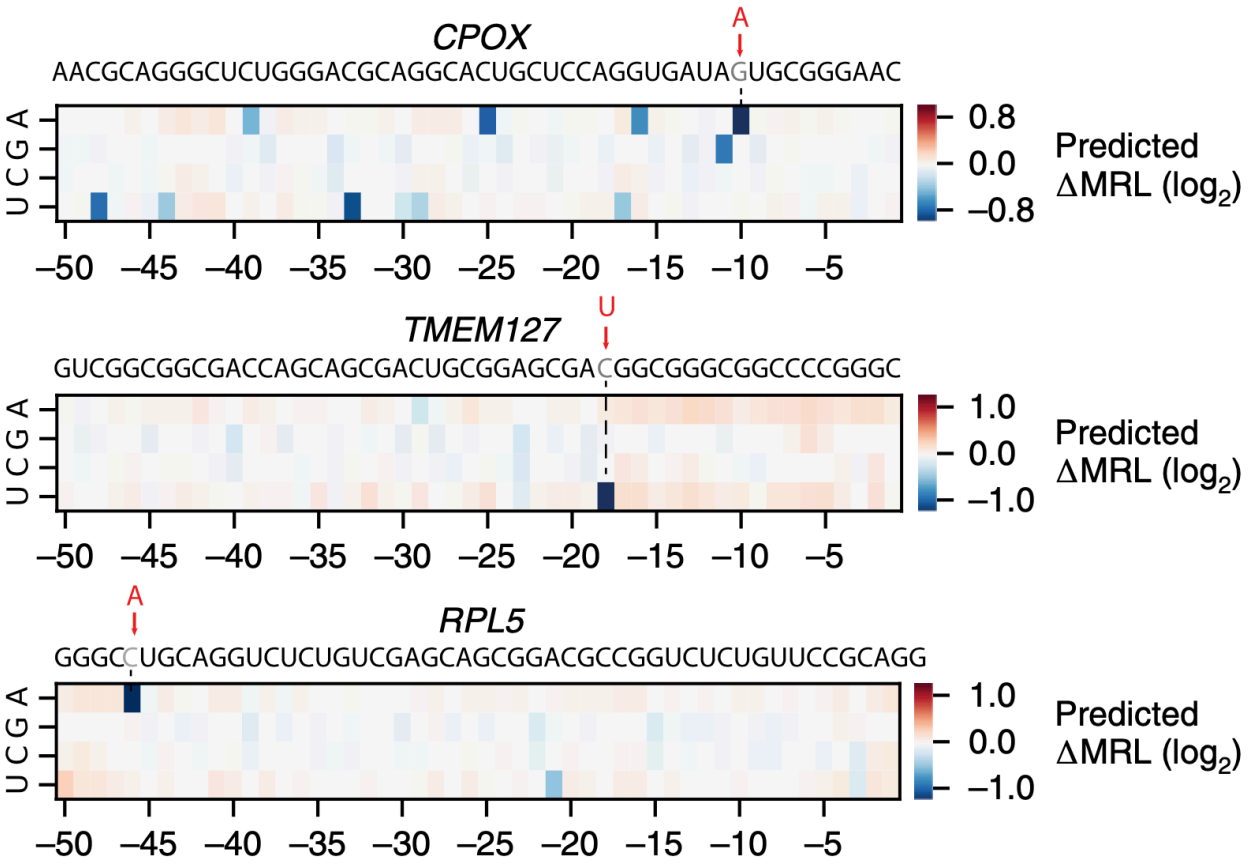


Figure 5.8. *In silico* saturation mutagenesis for gene *CPOX*, *TMEM127* and *RPL5*.

Using Optimus 5-Prime, *in silico* saturation mutagenesis and model prediction of the difference in MRL could be performed for all 5' UTR variants of interest, which could help us understand rare 5' UTR variants that occurred in human population that might not have enough patient data to work with yet. Figure 5.8 shows three examples of *in silico* saturation mutagenesis analysis for all 5' UTR variants of *CPOX*, *TMEM127* and *RPL5* in heatmaps. The reference sequence was placed on top, and each position could be mutated to the other 3 nucleotides and the log2 MRL differences was calculated by Optimus 5-Prime. Positive change which meant the introduced SNV promoted the translation, was denoted in red boxes and negative change which meant the introduced SNV repressed the translation, was denoted in blue boxes. In the three

examples shown in Figure 5.7, the three annotated ClinVar variants rs867711777 (*CPOX*, G>A), rs121908813 (*TMEM127*, C>U) and rs376208311 (*RPL5*, C>A), all of which introduced an upstream start codon, were predicted to have the most dramatic effect on ribosome loading. Our assay represented a molecular basis for these variants.

As an example, one of the ClinVar variants with sizeable differences in MRL, rs867711777, was found in the 5' UTR of the *CPOX* gene and showed a log2 difference of -0.89. Depletion of *CPOX* reduces heme biosynthesis and is the cause of hereditary coproporphyria [75]. The large MRL difference suggested that this SNV, labeled as uncertain in the ClinVar database, could be pathogenic. The rs376208311 variant lay in the 5' UTR of the ribosomal subunit gene *RPL5* and showed a log2 difference in MRL of -0.87. This variant is associated with Diamond–Blackfan anemia, which can be caused by disruption or downregulation of *RPL5* [76]. Another SNV in the 5' UTR of *TMEM127*, rs121908813, is implicated in familial pheochromocytoma, a condition characterized by tumors found in the neuroendocrine system that secrete high levels of catecholamines [77]. In our assay, the variant 5' UTR showed a log2 difference in MRL of -1.5 as compared to the wild-type 5' UTR sequence. *TMEM127* acts as a tumor suppressor, and decreased expression of it could explain the observed pathogenicity of this variant. This shows the potential of our assay to help identify new or rare SNVs that happens in human 5' UTR in the future clinical research.

5.3 Modeling 5' UTR of Varying Length

Human 5' UTR sequences vary in length from tens to thousands of nucleotides with a median length of 218 nucleotides [[11], [12]]. Because only 13% of human 5' UTRs are shorter than 50 nucleotides and can be covered by Optimus 5-Prime, we next asked whether the approach

introduced here could be extended to longer 5' UTRs. To this end, we first created a 5' UTR library where the length of the random sequence upstream of the start codon ranged from 25 to 100 nucleotides, which would increase the coverage of human 5' UTRs to 29%. After the experimental workflow demonstrated in Chapter 3 including IVT mRNA library synthesis, polysome profiling and RNA sequencing, we retained 83,919 distinct 5' UTRs spanning the entire length distribution from 25 to 100 nucleotides.

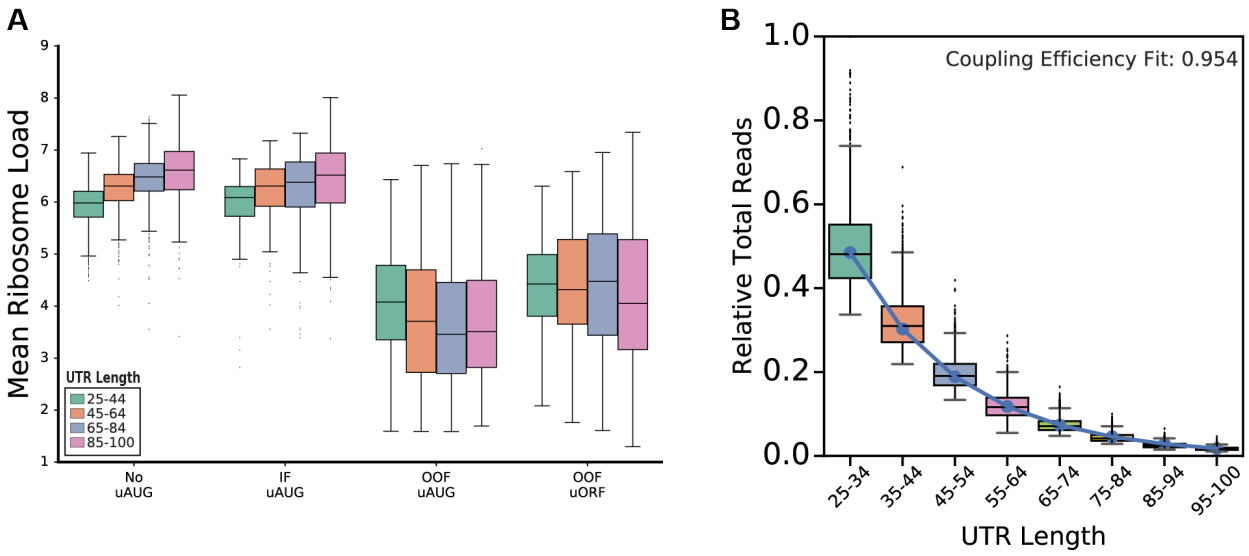


Figure 5.9. Upstream AUGs, UTR length and read depth effects on varying length 5' UTR data.

As observed with the 50-nucleotide library, out-of-frame upstream open reading frames (OOF uORFs) and upstream AUGs (uAUGs) caused reduced loading of ribosomes while in-frame (IF) do not. However, when sequences did not contain any uAUGs (No uAUGs) or have in-frame uAUGs (IF uAUGs), longer UTR sequences showed increased ribosome loading, likely because longer transcripts could accommodate more ribosomes (Figure 5.9A). We then retrained our model to capture and predict the impact of both sequence and length on MRL.

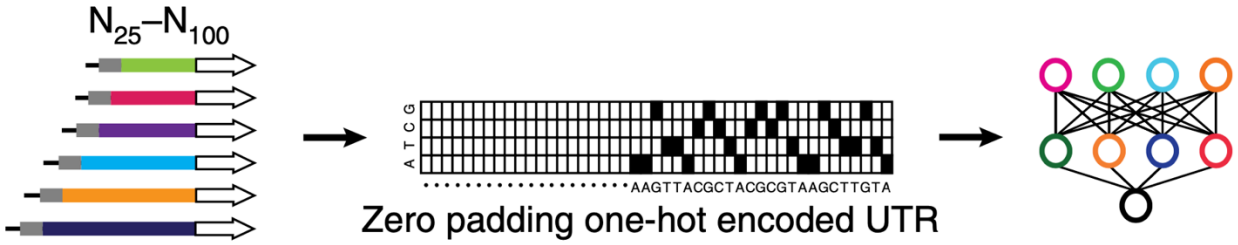


Figure 5.10. Generalized model structure on varying length 5' UTR.

To accommodate sequences up to 100 nucleotides in length, we increased the width of the input layer but otherwise retained the same network architecture as before which was trained on 50 nt library. Random sequences (n = 76,319) with lengths ranging from 25 to 100 nucleotides were used for training. The input space was expanded to 100 nucleotides using one-hot encoding (for sequences shorter than 100 nucleotides, zero padding was used) (Figure 5.10). To ensure that 5' UTRs of all lengths would be represented equally, the top 100 sequences at each length (~10% of the library), as measured by total read counts per UTR, rather than using the top 10% of the entire population, were used to test the model's accuracy, resulting in a test set of 7,600 random 5' UTRs. The remaining 90% of UTRs were used for training. In fact, we found that the average number of sequencing reads per library member rapidly decreased with increasing UTR length, likely because of the decreasing yield of full-length sequences for longer 5' UTRs shown in Figure 5.9B, where RNA sequencing read depth decreased with longer UTRs. All read counts were normalized to the maximum read count observed. All oligos used for 5' UTR construction were synthesized on the same array and the observed decrease was consistent with a coupling efficiency $c < 1$. A fit to the expression N^c where N was the length and c was the coupling efficiency resulted in $c = 0.954$ as a reasonable estimate of the coupling efficiency in oligo array synthesis (blue line).

We also created a second test set consisting of 7,600 human 5' UTRs in a similar way, corresponding to 100 UTRs for each length from 25 to 100 nucleotides: of 15,555 human UTRs

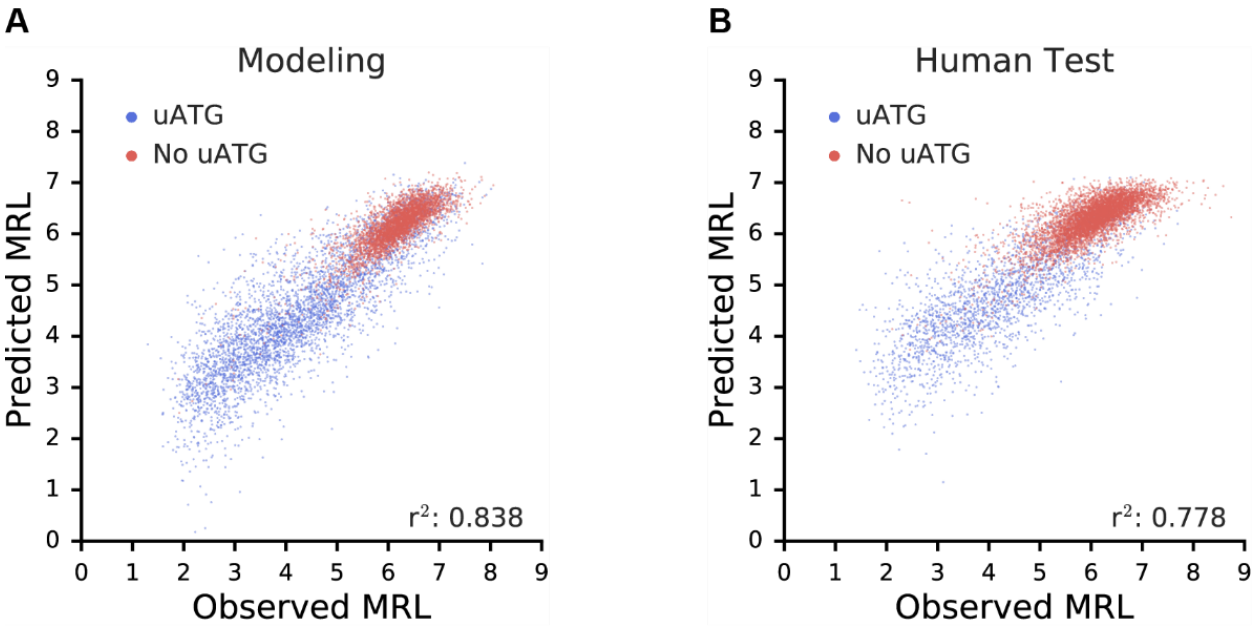


Figure 5.11. Generalized model performance on random and human sequences.

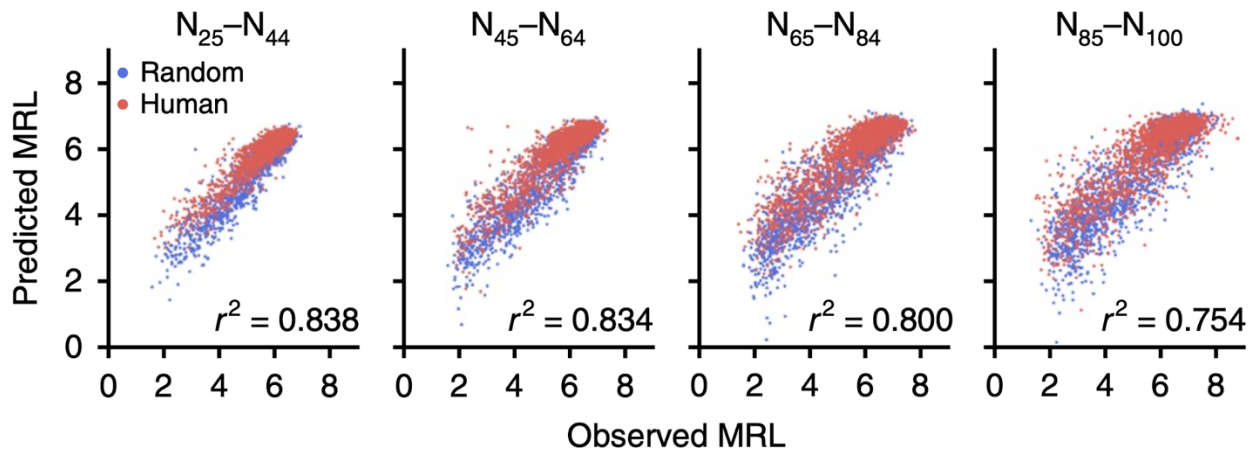


Figure 5.12. Random and human sequences were tested shown in length ranges.

that were detected in the sequencing data, the top 100 UTRs by read count at each length were used as part of the test set. The generalized Optimus 5-Prime model performed well on both the human (r^2 = 0.78) and random (r^2 = 0.84) sequences (Figure 5.11) and on 5' UTRs of any length (Figure 5.12). The slight decrease in performance observed for longer 5' UTRs was due to lower

read coverage for longer sequences and the concomitant decrease in the quality of the test set. These results suggested that the approach we developed here was not limited to fixed-length UTRs and could be extended even beyond a 100-nucleotide window by synthesizing correspondingly longer 5' UTRs for model training.

5.4 Optimus 5-Prime Webtool

With the strong scoring capacity that Optimus 5-Prime had, and the potential that it could be used to score any 5' UTR of interest, we raised a wide range of interests from scholars and researchers from molecular biology and clinical research. Although we have made all our data that supported all the work and findings described in previous sessions publicly available from Gene Expression Omnibus under accession GSE114002, and all code that needed to build Optimus 5-Prime was also available at Github (https://github.com/pjsample/human_5utr_modeling). There were still strong needs for lots of researchers without decent coding background and might find it difficult and challenging and time-consuming to use Optimus 5-Prime from scratch.

We sought to build an online webtool that was based on Optimus 5-Prime (https://optimus5.cs.washington.edu) using Python and HTML. This webtool provided three ways of inputs for users to query 5' UTR sequences (up to 100 nt) very easily, after getting input query sequence, MRL prediction would be provided, and *in silico* saturation mutagenesis was computed in real time and visualized for users to be able to find their variants of interests.

We provided three ways of input 5' UTR sequence as shown in Figure 5.13, which was a snapshot from the online webtool. The first one was to make input sequence with chromosome coordinates (GRCh38). After entering chromosome number, + or – strand, start position, and the end position or directly clicking on +100 button that would automatically calculate the end position

based on the start plus a hundred nucleotides, after clicking on ‘Search’ button, the corresponding sequence with the input chromosome coordinate would be showed up in the textbox on the bottom. The second way was to input 5’ UTR sequence with gene symbol and transcript ID, by selecting gene name from the list and the corresponding transcripts would show up (there might be multiple transcripts to one gene because of transcript isoforms), the user would need to select one transcript ID to identify which 5’ UTR sequence was their interest. The third way was simply to type in the sequence that the user was interested in. The online webtool was embedded with the generalized varying length Optimus 5-Prime model, so it could take sequence length from 25 nt to 100 nt. If the input sequence was shorter than 25 nt, then the textbox would show up alert message. If the input sequence was longer than 100 nt, and since the model was explicitly looking at the region just upstream of annotated start codon, the input sequence would be truncated to 100 nt from 5’ end (keep the 3’ end fixed) because the default direction of the input sequence was from 5’ to 3’. For sequences shorter than 100 nt but longer than 25 nt, we would pad ‘N’s to the 5’ end which was essentially doing zero-padding into the model prediction.

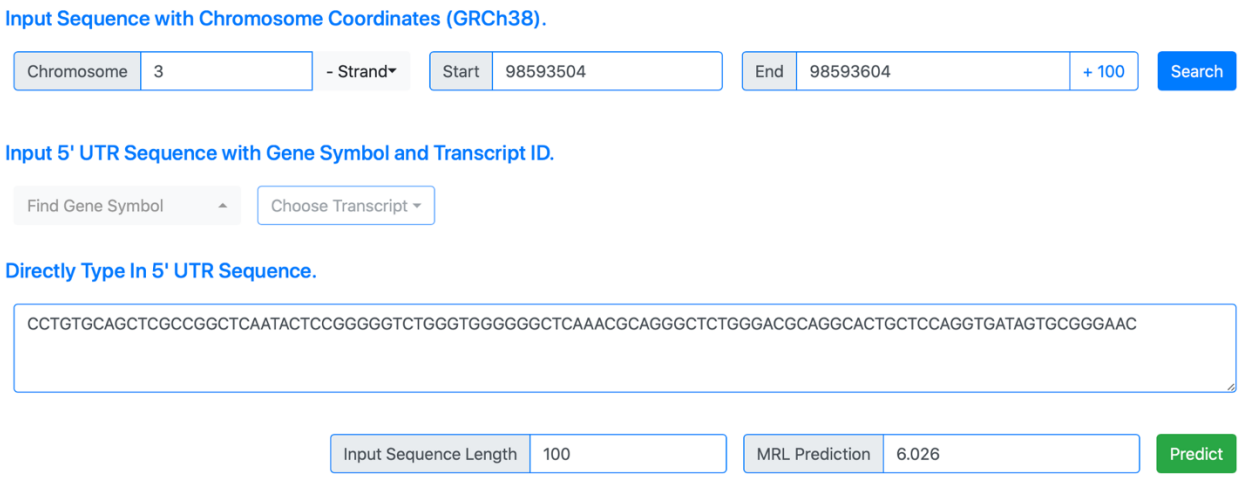


Figure 5.13. Three ways of inputs for Optimus 5-Prime webtool.

The webtool had two main functions, the first one was mean ribosome load (MRL) prediction, where the webtool would provide an MRL prediction based on the input of the 5' UTR sequence (Figure 5.13). In the meantime, the webtool would do *in silico* saturation mutagenesis in real time, where each position was mutated to the other three nucleotides and log2 MRL change was computed based on the MRL of reference sequences. The position label was considered as counting from the start codon, so the most 3' end position would be position 1 which presumably should be followed by a start codon. The largest positive log2 fold change value and position and mutation would be displayed in red font and the largest negative log2 fold change value, position and mutation would be displayed in blue font. Although in Chapter 5.2, when we were looking into human native 5' UTR and SNVs, only absolute value of log2 fold change of native and SNV sequences pair which was greater than 0.5 would be considered as significant variants. Here, we provided the freedom for users to choose what thresholds they were interested in to select out a list of SNVs from this *in silico* saturation mutagenesis. Thus, users could pick direction first, which meant what direction of fold changes they were interested in, if they would like to investigate SNVs that brought the MRL to a higher level, then they could select 'high than' and type in the numerical values as thresholds, or if SNVs that repressed the translation are what they were looking for, then they could select 'lower than' and type in the numerical values as the thresholds to select all variants brought significant repressing effects to the reference 5' UTR sequence. The generated list would be sorted by absolute values of log2 fold change in descending order by default but could also be changed to sorted by the SNV position (Figure 5.14).

The second function this webtool had was 'ClinVar Matching', which meant that it would match ClinVar Reports with Optimus 5-Prime's prediction. The main purpose of designing this function was that after getting a list of SNVs fulfilling certain thresholds from *in silico* saturation

mutagenesis, researchers were wondering whether these variants have been reported and recorded in datasets especially for those disease related variants. Therefore, this function could show if the query 5' UTR sequence had reported ClinVar SNVs. In order to match with established dataset, we only provided the first two ways of inputs for users – by chromosome coordinates and transcript ID. The resulting SNV list would show details of SNVs that were also recorded in the ClinVar dataset in the window of reference 5' UTR sequence. The list included information such as chromosome number, position, accession, relative position in this 100 nt window, original nucleotide, mutated nucleotide, significance recorded by ClinVar and the color code in the heatmap (Figure 5.15).

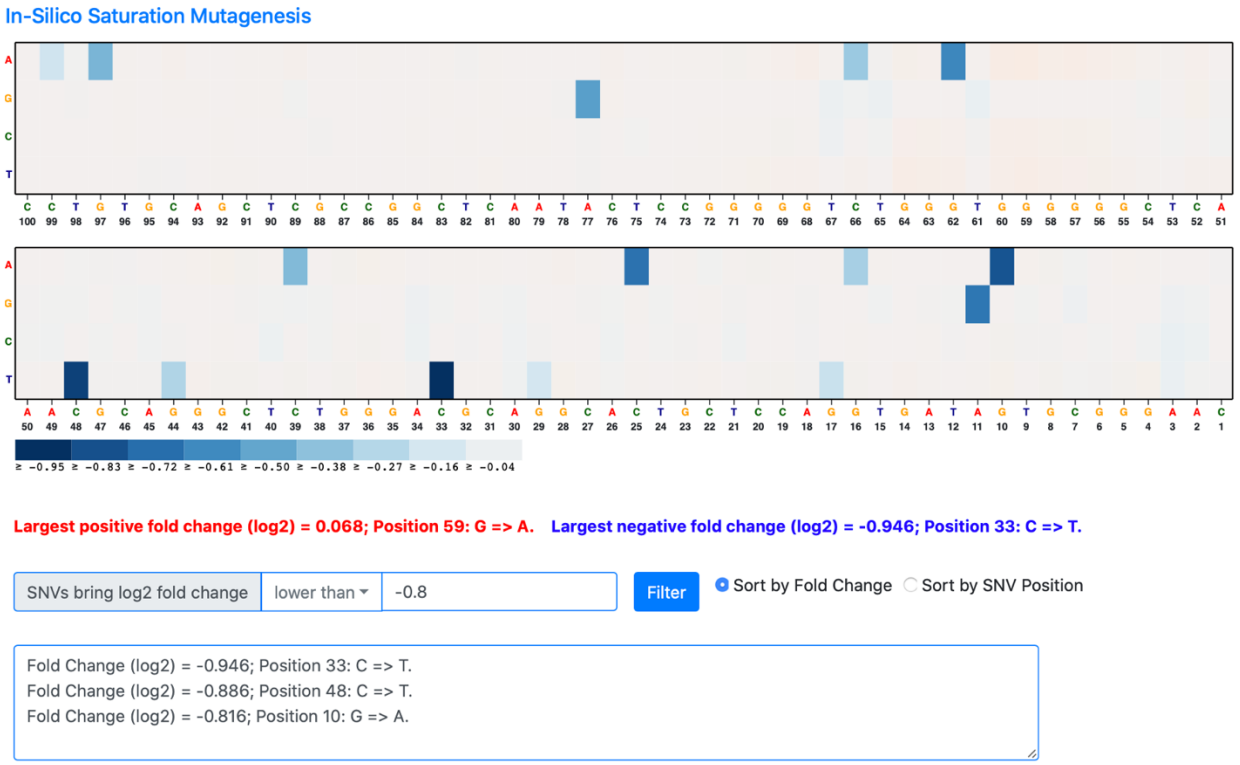
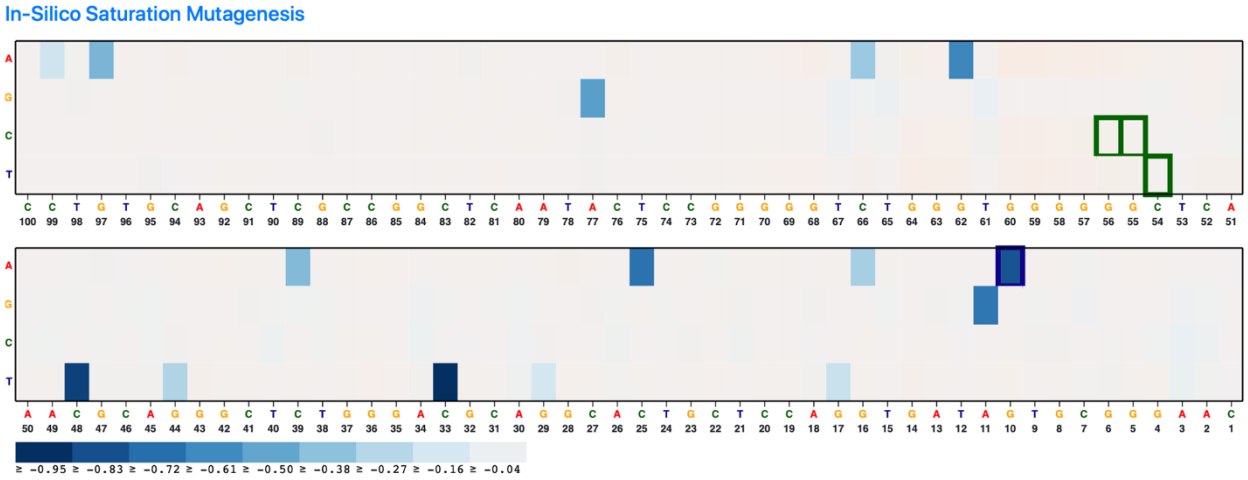


Figure 5.14. *In silico* saturation mutagenesis and SNV candidates list.



Clinvar SNVs Details

Chromosome	Position	Accession	Relative Position	Original	Mutated	Clinical Significance	Color Code
chr3	98593514	VCV000346997	10	G	A	Uncertain significance	Blue
chr3	98593558	VCV000346999	54	C	T	Likely benign	Green
chr3	98593559	VCV000347000	55	G	C	Likely benign	Green
chr3	98593560	VCV000347001	56	G	C	Likely benign	Green

Figure 5.15. *In silico* saturation mutagenesis and ClinVar Matching.

This Optimus 5-Prime webtool provides users a very easy and straight forward way to use our model, especially for those researchers and scholars do not have extensive practice in coding skills and potentially expand the users of our model to a much boarder range. The webtool gives users multiple ways for input sequences and provide MRL prediction and *in silico* saturation mutagenesis in a very fast time manner and provides users a list of SNVs that could possibly be disease related variants for further downstream research and analysis.

Chapter 6. EXTENSION OF 5' UTR MODEL

In previous chapters, we have shown that we have successfully built the Optimus 5-Prime which could assess the 5' UTR regulatory control on translation with originally 50 nt fixed length sequence length and later got expanded to varying length with range from 25 nt to 100 nt. The methodology we have proposed and showed that could be easily extended to other regions in the gene of interest, as well as other aspects of applications.

In this chapter, we will propose three future directions of this methodology and show some preliminary data of the work. First, as all the work we have conducted were in the HEK293T cells, we were wondering if we could transfer the study into other cell types such as T cells and hematopoietic stem cells (HSCs) to study 5' UTR regulatory effects in different cell types contexts and potentially find out if there are certain cell-type-specific motifs in 5' UTR regions. Second, the current library design had a 5' defined constant 25 nt long region upstream of the random region on the transcript which would limit us the ability to study the region adjacent to 5' cap in 5' UTRs, so we came up with an alternative library design and corresponding different experimental protocols to study this region. Third, while human 5' UTRs could be thousands of base pairs long, we wanted to use machine learning algorithms to build new models that can assess arbitrary length of 5' UTR to get rid of current upper limit length of 100 nt.

6.1 CELL-TYPE-SPECIFIC MOTIFS

There are hundreds of cell types in human body sharing the same genome. Despite the same reference book each cell type is holding, the cell-type-specific gene expression is widely observed. Studies have shown enhancers encode a lot of regulatory code that drives cell-type-specific gene

expression [78]. Cell-type-specific translation studies have been also performed but mostly in central neuron system (CNS) cells, because cell types in brain are very diverse and cell-type-specific behaviors are widely discussed in brain [[79]–[81]]. Here, we devoted to studying the 5' UTR dependent translation using the methodology we have described in previous chapters and expanding it into other cell types such as T cells and HSCs in collaboration with bluebird bio, Inc. We wanted to investigate the 5' UTR regulatory code on translation in T cells and HSCs , find out the cell-type-specific motifs in the 5' UTR region and also try to forward engineering new 5' UTR sequences that would maximize the expression level of translation in different cells respectively.

Here, we reported that we have finished our first trial on transferring our studies from HEK293T cells to T cells, although not much very interesting cell-type-specific motifs between the two cell types were found, it showed the proof that we could transfer our studies to very different cell types and we foresee there would be more interesting results coming from HSCs.

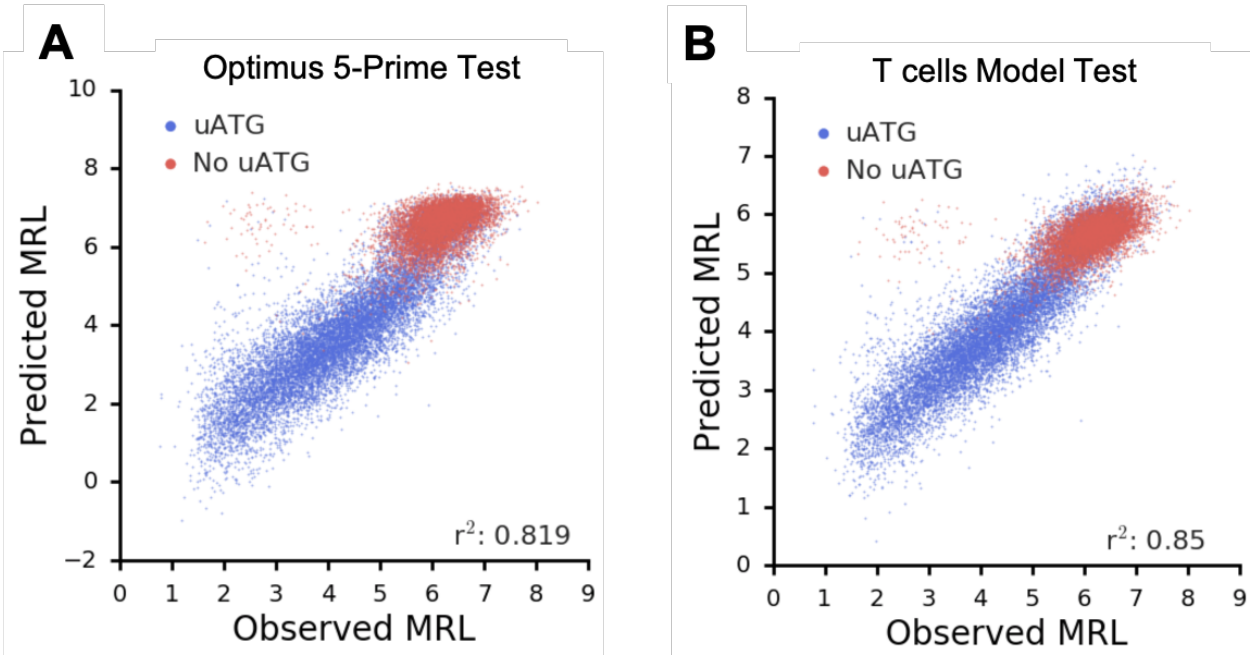


Figure 6.1. Model performance in T cells.

The protocol was optimized for primary cells which were not adherent and collected from donors. IVT mRNA library was electroporated into primary cells instead of transfection using lipofectamine and the incubation time after electroporation was reduced to 6 hours. The concentration of cycloheximide was 10 more times concentrated than original protocol. Each polysome profiling run for T cells used 10 million T cells in order to keep high and distinct peaks and better resolution in higher ribosomes fractions in the polysome profiles. After the same pipeline of high-throughput sequencing and data analysis that was performed as described in previous chapters in HEK293T cells, we have successfully collected the first trial run in T cells from the same IVT eGFP mRNA library. We first tested Optimus 5-Prime which was trained on the data from HEK293T cells on this data, and the performance was well with r-squared value 0.819 (Figure 6.1A). We also generated an independent T cell model which was trained on the 260,000-training dataset of its own and tested on the on-held test set, and the r-squared value was 0.85 (Figure 6.1B). Although we could observe a slightly increase on model accuracy when the model was trained on T cell data points, the improvement was not that significant, and we thought the lower r-squared value from Optimus 5-Prime test was due to the slightly different distribution of data collected between the two independent performed experiments in HEK293T and T cells, which did not strongly implicate cell-type-specific motifs finding.

Some low-throughput experiments that our collaborator has conducted showed that certain UTRs had distinct performance in HSCs, but much similar performances in T cells and HEK293T, which agreed with our high-throughput large scale experiments. However, the data and results from T cells were still from preliminary stage, and there would still be a large room to tune and improve, and we will also work on finding cell-type-specific motifs between HSC and HEK293T.

6.2 New Library Without Defined 5' Region

Optimus 5-Prime was developed to study 5' UTR dependent translation and focused on the region that adjacent to the CDS. However, the 5' m7G cap serves as a very unique module that will regulate many biological functions such as pre-mRNA processing, nuclear export and cap-dependent translation [82]. The region close to 5' cap and 5' cap itself are heavily involved in recruiting initiation factors and assembling ribosome [14]. Babendure et al. [83] designed a wide range of hairpins with multiple distances of 5' m7G cap (Figure 6.2A) and showed that shifting the position of hairpin relative to the 5' cap could modulate translation more than 50-fold (Figure 6.2B).

Figure 6.2. Shifting the position of hairpin relative to the 5' cap would modulate translation by Babendure et al [83].

Here, in order to investigate the region close to the 5' cap, we designed a new IVT mRNA library without the defined 5' end and modifying the protocol of RT-PCR after RNA extraction in order to amplify the signal of our library (Figure 6.3). We built two versions of random libraries with different lengths but identical structure, one was with random region 25 bp long (N25) and the other one was with random region 50 bp long (N50). After constructing the DNA plasmid

libraries, linear DNA fragment served as IVT template was generated using PCR and then *in vitro* transcription was performed to get IVT mRNA libraries (N50 and N25). After polysome profiling and RNA extractions to each desired fraction, a slightly different approach was taken because the transcript did not have the defined 5' to serve as PCR handle anymore. Instead, when performing reverse transcription, maxima H minus reverse transcriptase (ThermoFisher) was used which would add three more Cs to the 3' end of the cDNA, and a template switching oligo (TSO) was added in, which had three riboguanosines at 3' end that would bind to the 3 Cs on cDNA and make the RT to continue on the TSO sequence [84]. As a result, PCR could be performed to target the TSO sequence to amplify the library signal subject to next-generation sequencing. With the data collected from these two new designed 5' UTR, we believed it could provide us a comprehensive view to understand how the sequences close to 5' cap would modulate translation.

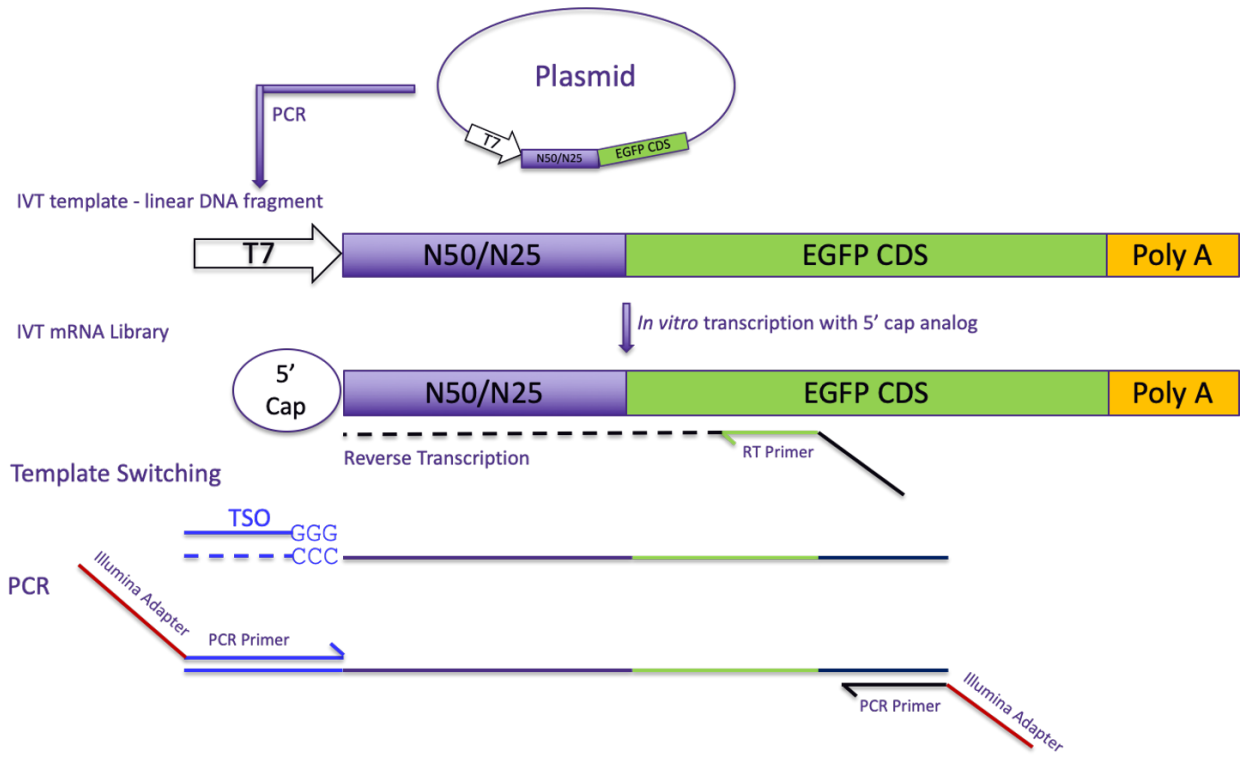


Figure 6.3. New library design and new protocol incorporating with template switching.

6.3 ULTIMATE 5' UTR MODEL

As what has been discussed in Chapter 5.3, human 5' UTRs span a large range of length distribution from tens to thousands with a median of 218 nucleotides [[11], [12]]. Although we have shown in Chapter 5 that Optimus 5-Prime was originally built based on random 50 nt library and could be expanded to varying length model covering 25 to 100 nt, it is not feasible to just naively extend the experimental assay to create one thousands of base pairs long library in order to build model to predict long UTRs. Instead, here we propose to use recurrent neural network (RNN) to build an ultimate 5' UTR model that can take in arbitrary length of 5' UTRs with current data.

Convolutional neural networks and *k*-mer features have been shown having very good performance but research indicates that RNN has better performance in time-series data [85]. RNN was first built with applications to speech recognition and language model but has been shown successfully applied to predict transcription factor binding sites [86] and DNA base modification [87]. 5' UTR dependent translation involves ribosome assembly around the 5' cap, scan through the sequence and translation initiation around the annotated start codon, so we believe that with the two version of random datasets we collect, one from the region very close to 5' cap and one from the region very close to the annotated start codon, together with deep RNN, we will be able to build an ultimate 5' UTR model that can score any human 5' UTR with arbitrary length.

BIBLIOGRAPHY

[1] S. Levy *et al.*, "The diploid genome sequence of an individual human," *PLoS Biol.*, vol. 5, no. 10, pp. 2113–2144, Oct. 2007, doi: 10.1371/journal.pbio.0050254.

[2] National Human Genome Research Institute, "The cost of sequencing a human genome." https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/ (accessed Apr. 20, 2020).

[3] J. Han *et al.*, "A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation," *PLoS Genet.*, vol. 4, no. 5, May 2008, doi: 10.1371/journal.pgen.1000074.

[4] D. N. Cooper and H. Youssoufian, "The CpG dinucleotide and human genetic disease," *Hum. Genet.*, 1988, doi: 10.1007/BF00278187.

[5] J. J. Shu, "A new integrated symmetrical table for genetic codes," *BioSystems*, 2017, doi: 10.1016/j.biosystems.2016.11.004.

[6] G. Elgar and T. Vavouri, "Tuning in to the signals: noncoding sequence conservation in vertebrate genomes," *Trends in Genetics*. 2008, doi: 10.1016/j.tig.2008.04.005.

[7] D. L. Black, "Mechanisms of Alternative Pre-Messenger RNA Splicing," *Annu. Rev. Biochem.*, 2003, doi: 10.1146/annurev.biochem.72.121801.161720.

[8] Y. Shen *et al.*, "Genome level analysis of rice mRNA 3′-end processing signals and alternative polyadenylation," *Nucleic Acids Res.*, 2008, doi: 10.1093/nar/gkn158.

[9] J. Guhaniyogi and G. Brewer, "Regulation of mRNA stability in mammalian cells," *Gene*. 2001, doi: 10.1016/S0378-1119(01)00350-X.

[10] C. U. T. Hellen and P. Sarnow, "Internal ribosome entry sites in eukaryotic mRNA molecules," *Genes and Development*. 2001, doi: 10.1101/gad.891101.

[11] F. Mignone, C. Gissi, S. Liuni, G. Pesole, and others, "Untranslated regions of mRNAs," *Genome Biol*, 2002, doi: 10.1186/gb-2002-3-3-reviews0004.

[12] K. Leppek, R. Das, and M. Barna, "Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them," *Nature Reviews Molecular Cell Biology*. 2018, doi: 10.1038/nrm.2017.103.

[13] P. R. Araujo *et al.*, "Before it gets started: Regulating translation at the 5′; UTR," *Comparative and Functional Genomics*. 2012, doi: 10.1155/2012/475731.

[14] R. J. Jackson, C. U. T. Hellen, and T. V. Pestova, “The mechanism of eukaryotic translation initiation and principles of its regulation,” *Nature Reviews Molecular Cell Biology*. 2010, doi: 10.1038/nrm2838.

[15] T. G. Johnstone, A. A. Bazzini, and A. J. Giraldez, “Upstream ORFs are prevalent translational repressors in vertebrates,” *EMBO J.*, 2016, doi: 10.15252/embj.201592759.

[16] D. R. Morris and A. P. Geballe, “Upstream Open Reading Frames as Regulators of mRNA Translation,” *Mol. Cell. Biol.*, 2000, doi: 10.1128/MCB.20.23.8635-8642.2000.

[17] M. Kozak, “Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes,” *Cell*, 1986, doi: 10.1016/0092-8674(86)90762-2.

[18] W. L. Noderer *et al.*, “Quantitative analysis of mammalian translation initiation sites by FACS-seq,” *Mol. Syst. Biol.*, 2014, doi: 10.15252/msb.20145136.

[19] A. G. Hinnebusch, “The Scanning Mechanism of Eukaryotic Translation Initiation,” *Annu. Rev. Biochem.*, 2014, doi: 10.1146/annurev-biochem-060713-035802.

[20] R. P. Smith *et al.*, “Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model,” *Nat. Genet.*, 2013, doi: 10.1038/ng.2713.

[21] Hentze MW, Muckenthaler MU, and Andrews NC., “Balancing acts: molecular control of mammalian iron metabolism.,” *Cell*, 2004.

[22] A. S. Y. Lee, P. J. Kranzusch, and J. H. D. Cate, “EIF3 targets cell-proliferation messenger RNAs for translational activation or repression,” *Nature*, 2015, doi: 10.1038/nature14267.

[23] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, “Benefits and limitations of genome-wide association studies,” *Nature Reviews Genetics*. 2019, doi: 10.1038/s41576-019-0127-1.

[24] A. Korte and A. Farlow, “The advantages and limitations of trait analysis with GWAS: A review,” *Plant Methods*. 2013, doi: 10.1186/1746-4811-9-29.

[25] A. Melnikov *et al.*, “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay,” *Nat. Biotechnol.*, 2012, doi: 10.1038/nbt.2137.

[26] R. P. Patwardhan, C. Lee, O. Litvin, D. L. Young, D. Pe’Er, and J. Shendure, “High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis,” *Nat.*

Biotechnol., 2009, doi: 10.1038/nbt.1589.

[27] E. Sharon *et al.*, "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters," *Nat. Biotechnol.*, 2012, doi: 10.1038/nbt.2205.

[28] P. Oikonomou, H. Goodarzi, and S. Tavazoie, "Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts," *Cell Rep.*, 2014, doi: 10.1016/j.celrep.2014.03.001.

[29] W. Zhao, J. L. Pollack, D. P. Blagev, N. Zaitlen, M. T. McManus, and D. J. Erle, "Massively parallel functional annotation of 3' untranslated regions," *Nat. Biotechnol.*, 2014, doi: 10.1038/nbt.2851.

[30] S. Ke *et al.*, "Quantitative evaluation of all hexamers as exonic splicing elements," *Genome Res.*, 2011, doi: 10.1101/gr.119628.110.

[31] A. B. Rosenberg, R. P. Patwardhan, J. Shendure, and G. Seelig, "Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences," *Cell*, 2015, doi: 10.1016/j.cell.2015.09.054.

[32] N. Bogard, J. Linder, A. B. Rosenberg, and G. Seelig, "A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation," *Cell*, 2019, doi: 10.1016/j.cell.2019.04.046.

[33] J. T. Cuperus *et al.*, "Deep learning of the regulatory grammar of yeast 5′untranslated regions from 500,000 random sequences," *Genome Res.*, 2017, doi: 10.1101/gr.224964.117.

[34] P. J. Sample *et al.*, "Human 5′ UTR design and variant effect prediction from a massively parallel translation assay," *Nat. Biotechnol.*, 2019, doi: 10.1038/s41587-019-0164-5.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, 2017, doi: 10.1145/3065386.

[36] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *31st International Conference on Machine Learning, ICML 2014*, 2014.

[37] Z. Chen and X. Huang, "End-To-end learning for lane keeping of self-driving cars," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2017, doi: 10.1109/IVS.2017.7995975.

[38] D. Silver *et al.*, "Mastering the game of Go without human knowledge," *Nature*, 2017,

doi: 10.1038/nature24270.

[39] Q. Kang and J. R. Pomerening, “Punctuated cyclin synthesis drives early embryonic cell cycle oscillations,” *Mol. Biol. Cell*, 2012, doi: 10.1091/mbc.E11-09-0768.

[40] M. J. Del Prete, R. Vernal, H. Dolznig, E. W. Müllner, and J. A. Garcia-Sanz, “Isolation of polysome-bound mRNA from solid tissues amenable for RT-PCR and profiling experiments,” *RNA*, 2007, doi: 10.1261/rna.79407.

[41] H. Chassé, O. Mulner-Lorillon, S. Boulben, V. Glippa, J. Morales, and P. Cormier, “Cyclin B translation depends on mTOR activity after fertilization in sea urchin embryos,” *PLoS One*, 2016, doi: 10.1371/journal.pone.0150318.

[42] J. Chen *et al.*, “Genome-wide analysis of translation reveals a critical role for deleted in azoospermia-like (Dazl) at the oocyte-to-zygote transition,” *Genes Dev.*, 2011, doi: 10.1101/gad.2028911.

[43] O. Larsson, B. Tian, and N. Sonenberg, “Toward a genome-wide landscape of translational control,” *Cold Spring Harbor Perspectives in Biology*. 2013, doi: 10.1101/cshperspect.a012302.

[44] S. N. Floor and J. A. Doudna, “Tunable protein synthesis by transcript isoforms in human cells,” *Elife*, 2016, doi: 10.7554/eLife.10921.

[45] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, “Deep learning for computational biology,” *Mol. Syst. Biol.*, 2016, doi: 10.15252/msb.20156651.

[46] K. Jaganathan *et al.*, “Predicting Splicing from Primary Sequence with Deep Learning,” *Cell*, 2019, doi: 10.1016/j.cell.2018.12.015.

[47] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nat. Biotechnol.*, 2015, doi: 10.1038/nbt.3300.

[48] N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman, “The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments,” *Nat. Protoc.*, 2012, doi: 10.1038/nprot.2012.086.

[49] V. Gandin *et al.*, “Polysome fractionation and analysis of mammalian translatomes on a genome-wide scale,” *J. Vis. Exp.*, 2014, doi: 10.3791/51455.

[50] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, 2011, doi: 10.14806/ej.17.1.200.

[51] L. Zhao, Z. Liu, S. F. Levy, and S. Wu, “Bartender: a fast and accurate clustering algorithm to count barcode reads,” *Bioinformatics*, 2017, doi: 10.1093/bioinformatics/btx655.

[52] X. Wang, J. Hou, C. Quedenau, and W. Chen, “Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals,” *Mol. Syst. Biol.*, 2016, doi: 10.15252/msb.20166941.

[53] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, and S.-B. Qian, “Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution,” *Proc. Natl. Acad. Sci.*, 2012, doi: 10.1073/pnas.1207846109.

[54] K. Reuter, A. Biehl, L. Koch, and V. Helms, “PreTIS: A Tool to Predict Non-canonical 5’ UTR Translational Initiation Sites in Human and Mouse,” *PLoS Comput. Biol.*, 2016, doi: 10.1371/journal.pcbi.1005170.

[55] M. Kozak, “Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes,” *Cell*, 1986, doi: 10.1016/0092-8674(86)90762-2.

[56] A. G. Hinnebusch, “The Scanning Mechanism of Eukaryotic Translation Initiation,” *Annu. Rev. Biochem.*, 2014, doi: 10.1146/annurev-biochem-060713-035802.

[57] J. N. Zadeh *et al.*, “NUPACK: Analysis and design of nucleic acid systems,” *J. Comput. Chem.*, 2011, doi: 10.1002/jcc.21596.

[58] S. R. Starck *et al.*, “Translation from the 5’ untranslated region shapes the integrated stress response,” *Science (80-.).*, 2016, doi: 10.1126/science.aad3867.

[59] M. Abadi *et al.*, “TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning,” *Proc 12th USENIX Conf. Oper. Syst. Des. Implement.*, 2016, doi: 10.1126/science.aab4113.4.

[60] J. P. Ferreira, K. W. Overton, and C. L. Wang, “Tuning gene expression with synthetic upstream open reading frames,” *Proc. Natl. Acad. Sci.*, 2013, doi: 10.1073/pnas.1305590110.

[61] D. R. Kelley, J. Snoek, and J. L. Rinn, “Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks,” *Genome Res.*, 2016, doi: 10.1101/gr.200535.115.

[62] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, “Quantifying similarity between motifs,” *Genome Biol.*, 2007, doi: 10.1186/gb-2007-8-2-r24.

[63] D. Ray *et al.*, “A compendium of RNA-binding motifs for decoding gene regulation,” *Nature*, 2013, doi: 10.1038/nature12311.

[64] K. Karikó *et al.*, “Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability,” *Mol. Ther.*, 2008, doi: 10.1038/mt.2008.200.

[65] B. R. Anderson *et al.*, “Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation,” *Nucleic Acids Res.*, 2010, doi: 10.1093/nar/gkq347.

[66] M. J. Landrum *et al.*, “ClinVar: Public archive of interpretations of clinically relevant variants,” *Nucleic Acids Res.*, 2016, doi: 10.1093/nar/gkv1222.

[67] S. W. Seo *et al.*, “Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency,” *Metab. Eng.*, 2013, doi: 10.1016/j.ymben.2012.10.006.

[68] M. K. Jensen and J. D. Keasling, “Recent applications of synthetic biology tools for yeast metabolic engineering,” *FEMS Yeast Research*. 2015, doi: 10.1111/1567-1364.12185.

[69] H. M. Salis, E. A. Mirsky, and C. A. Voigt, “Automated design of synthetic ribosome binding sites to control protein expression,” *Nat. Biotechnol.*, 2009, doi: 10.1038/nbt.1568.

[70] R. D. Hernandez, L. H. Uricchio, K. Hartman, C. Ye, A. Dahl, and N. Zaitlen, “Singleton Variants Dominate the Genetic Architecture of Human Gene Expression,” 2018. doi: 10.2139/ssrn.3151998.

[71] F. Cunningham *et al.*, “Ensembl 2019,” *Nucleic Acids Res.*, 2019, doi: 10.1093/nar/gky1113.

[72] D. Smedley *et al.*, “BioMart - Biological queries made easy,” *BMC Genomics*, 2009, doi: 10.1186/1471-2164-10-22.

[73] A. Battle *et al.*, “Impact of regulatory variation from RNA to protein,” *Science (80-.).*, 2015, doi: 10.1126/science.1260793.

[74] C. Cenik *et al.*, “Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans,” *Genome Res.*, 2015, doi: 10.1101/gr.193342.115.

[75] B. Wang and D. M. Bissell, *Hereditary Coproporphyria*. University of Washington,

Seattle, Seattle (WA), 2012.

[76] I. Boria *et al.*, “The ribosomal basis of diamond-blackfan anemia: Mutation and database update,” *Hum. Mutat.*, 2010, doi: 10.1002/humu.21383.

[77] Y. Qin *et al.*, “Germline mutations in TMEM127 confer susceptibility to pheochromocytoma,” *Nat. Genet.*, 2010, doi: 10.1038/ng.533.

[78] S. Heinz, C. E. Romanoski, C. Benner, and C. K. Glass, “The selection and function of cell type-specific enhancers,” *Nature Reviews Molecular Cell Biology*. 2015, doi: 10.1038/nrm3949.

[79] S. R. Thomson *et al.*, “Cell-Type-Specific Translation Profiling Reveals a Novel Strategy for Treating Fragile X Syndrome,” *Neuron*, 2017, doi: 10.1016/j.neuron.2017.07.013.

[80] D. Sapkota *et al.*, “Cell-Type-Specific Profiling of Alternative Translation Identifies Regulated Protein Isoform Variation in the Mouse Brain,” *Cell Rep.*, 2019, doi: 10.1016/j.celrep.2018.12.077.

[81] P. Shrestha *et al.*, “Cell-type-specific drug-inducible protein synthesis inhibition demonstrates that memory consolidation requires rapid neuronal translation,” *Nat. Neurosci.*, 2020, doi: 10.1038/s41593-019-0568-z.

[82] A. Ramanathan, G. B. Robb, and S. H. Chan, “mRNA capping: Biological functions and applications,” *Nucleic Acids Research*. 2016, doi: 10.1093/nar/gkw551.

[83] J. R. Babendure, J. L. Babendure, J. H. Ding, and R. Y. Tsien, “Control of mammalian translation by mRNA structure near caps,” *RNA*, 2006, doi: 10.1261/rna.2309906.

[84] A. Turchinovich, H. Surowy, A. Serva, M. Zapatka, P. Lichter, and B. Burwinkel, “Capture and Amplification by Tailing and Switching (CATS),” *RNA Biol.*, 2014, doi: 10.4161/rna.29304.

[85] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, “Recurrent neural network based language model,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[86] Z. Shen, W. Bao, and D. S. Huang, “Recurrent Neural Network for Predicting Transcription Factor Binding Sites,” *Sci. Rep.*, 2018, doi: 10.1038/s41598-018-33321-1.

[87] Q. Liu, L. Fang, G. Yu, D. Wang, C. Le Xiao, and K. Wang, “Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data,” *Nat. Commun.*, 2019, doi: 10.1038/s41467-019-10168-2.

APPENDIX A

Supplementary Material for Chapter 4.4

Visualization of 120 filters from the first convolution layer of the Optimus 5-Prime was shown in Figure A.1. For each filter we collected the top 2,000 8-mers in the eGFP library that showed maximal activation. These were then used to calculate position weight matrices and visualized as sequence logos. Some filters had fewer than 2,000 8-mers that showed activation.

Visualization of filters from the second convolution layer of the Optimus 5-Prime was shown in Figure A.2. For each filter we collected the top 2,000 8-mers in the eGFP library that showed maximal activation. These were then used to calculate position weight matrices and visualized as sequence logos. Some filters had fewer than 2,000 8-mers that showed activation or no activation at all.

For convolution layers one and two, the correlation between filter activation and MRL for each filter at each position of the 5′ UTRs was shown in Figure A.3. If UTRs that showed high filter activation had low MRLs then the two are negatively correlated. This showed the importance of each filter at each position for predicting MRL.

Figure A.1-1. Visualization of 120 filters (No.0-59) from the first convolution layer.

Figure A.1-2. Visualization of 120 filters (No.60-119) from the first convolution layer.

Figure A.2-1. Visualization of 120 filters (No.0-40) from the second convolution layer.

Figure A.2-2. Visualization of 120 filters (No.41-80) from the second convolution layer.

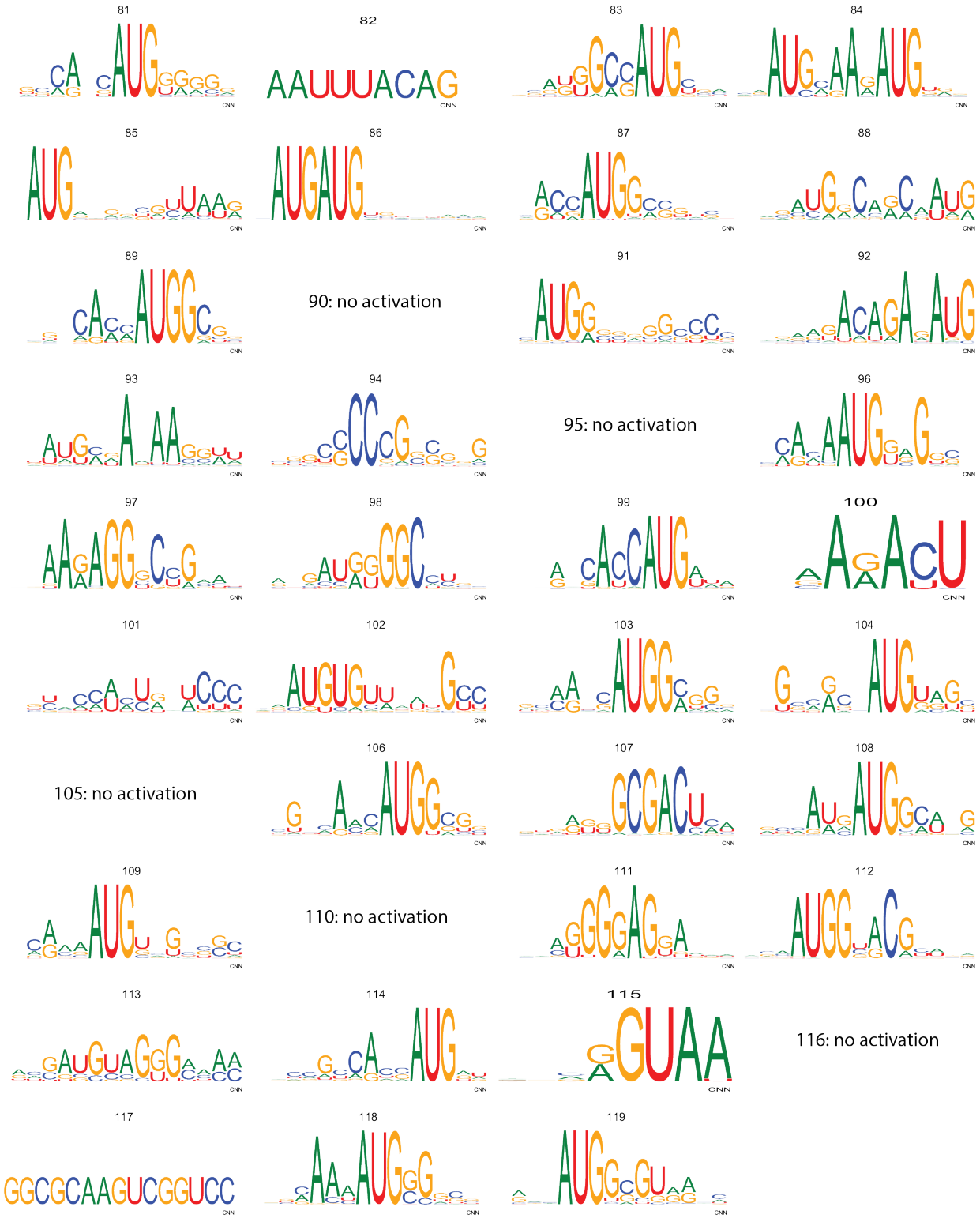


Figure A.2-3. Visualization of 120 filters (No.81-119) from the second convolution layer.

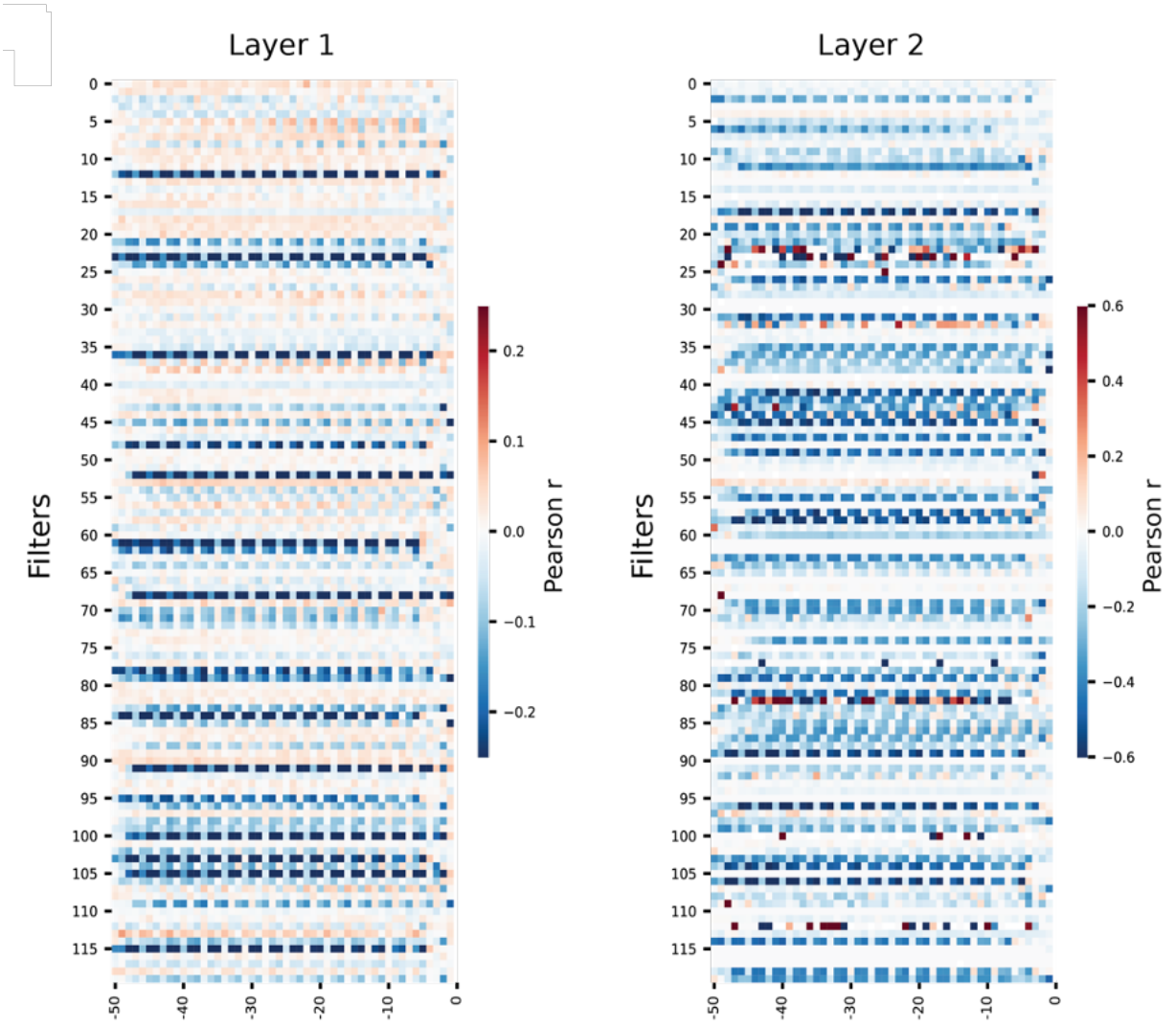


Figure A.3. Correlation between filter activation and MRL for each filter at each position of 5' UTR sequence for first convolutional layer and second convolutional layer.

APPENDIX B

Supplementary Material for Chapter 5.1

All 80 sequences evolution of which 20 UTRs under four distinct conditions were shown in this appendix. Observed MRL experimental data in blue, predicted MRL by original model in green and predicted MRL by retrained model in red for each evolution process were shown. The retrained model closely matched the observed MRL and performed significantly better than the original model that was used for evolving the sequences. Following the sequence evolution of our genetic algorithm, the four distinct conditions for 20 UTRs of each were as following: Figure B-1 shows the 20 random starting UTRs were evolved to the lowest ribosome load over 800 iterations and then changed the selective pressure for highest ribosome load over 800 iterations allowing AUGs; Figure B-2 shows the 20 random starting UTRs were evolved to the lowest ribosome load over 800 iterations and then changed the selective pressure for highest ribosome load over 800 iterations without allowing the triplet AUGs in the sequences. Figure B-3 shows the 20 random starting UTRs were evolved to the highest ribosome load over 800 iterations allowing AUGs; Figure B-4 shows the 20 random starting UTRs were evolved to the highest ribosome load over 800 iterations without allowing the triplet AUGs in the sequences.

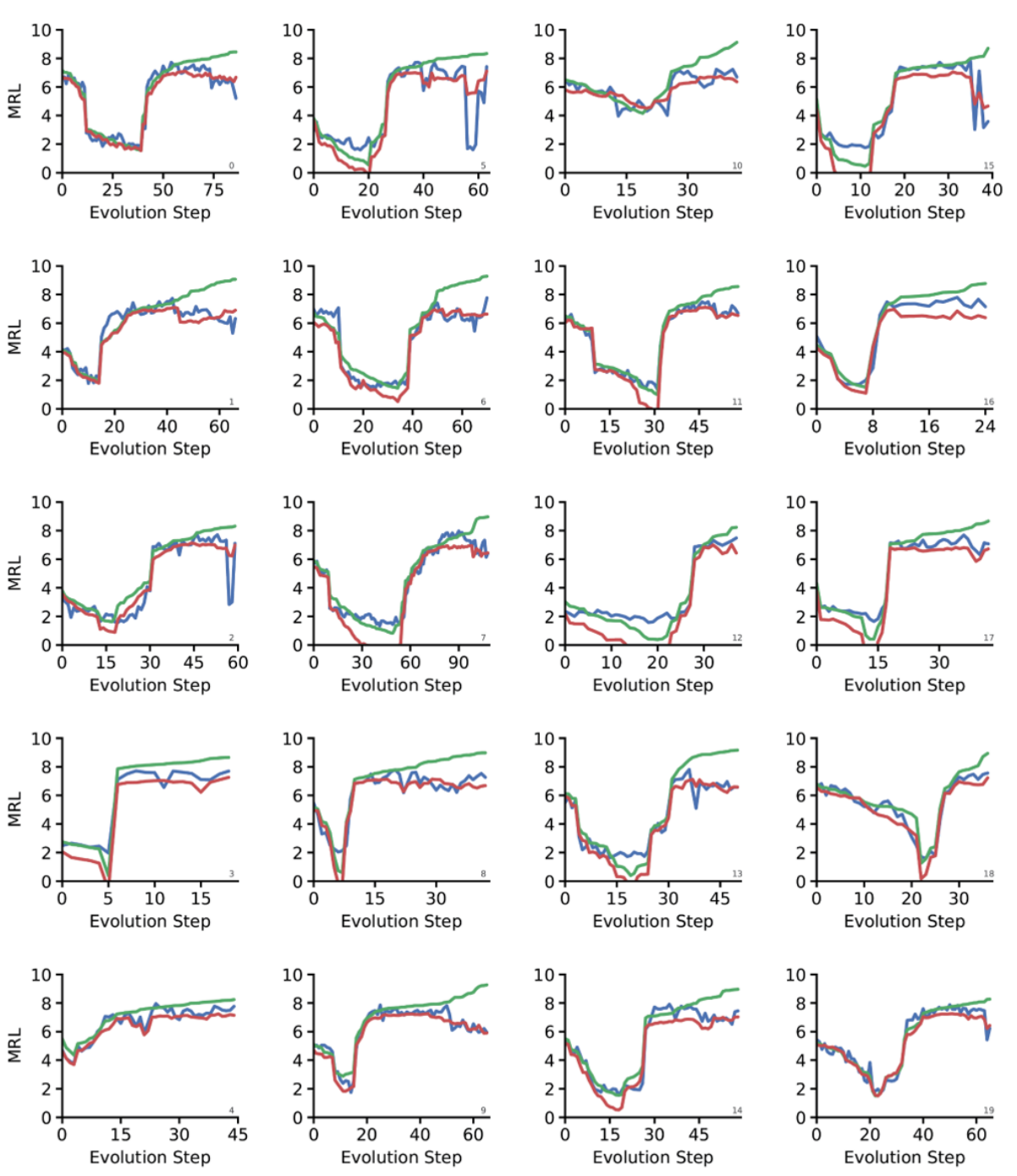


Figure B-1. UTRs were evolved to the lowest ribosome load over 800 iterations and then changed the selective pressure for highest ribosome load over 800 iterations allowing AUGs.

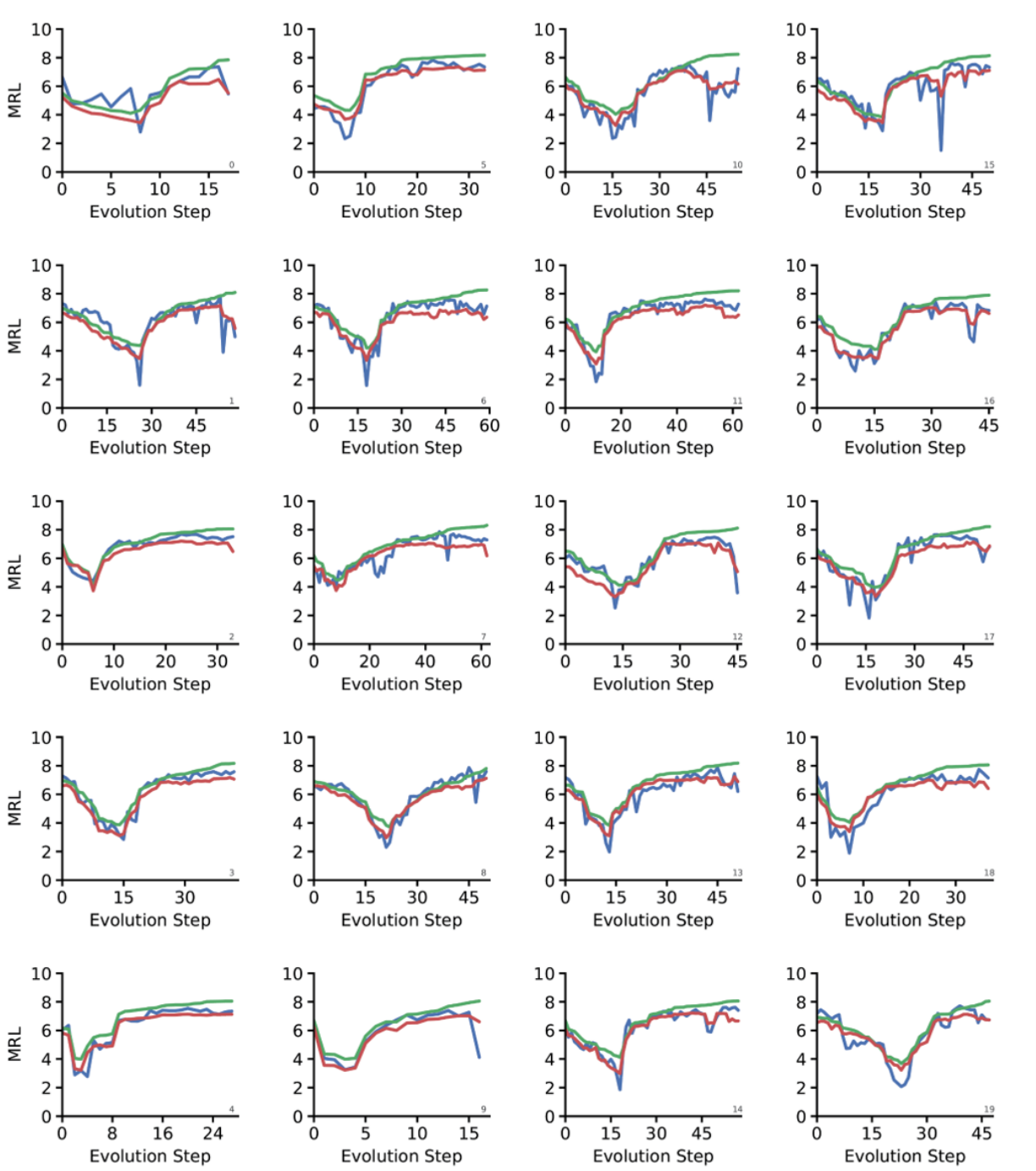


Figure B-2. UTRs were evolved to the lowest ribosome load over 800 iterations and then changed the selective pressure for highest ribosome load over 800 iterations without allowing AUGs.

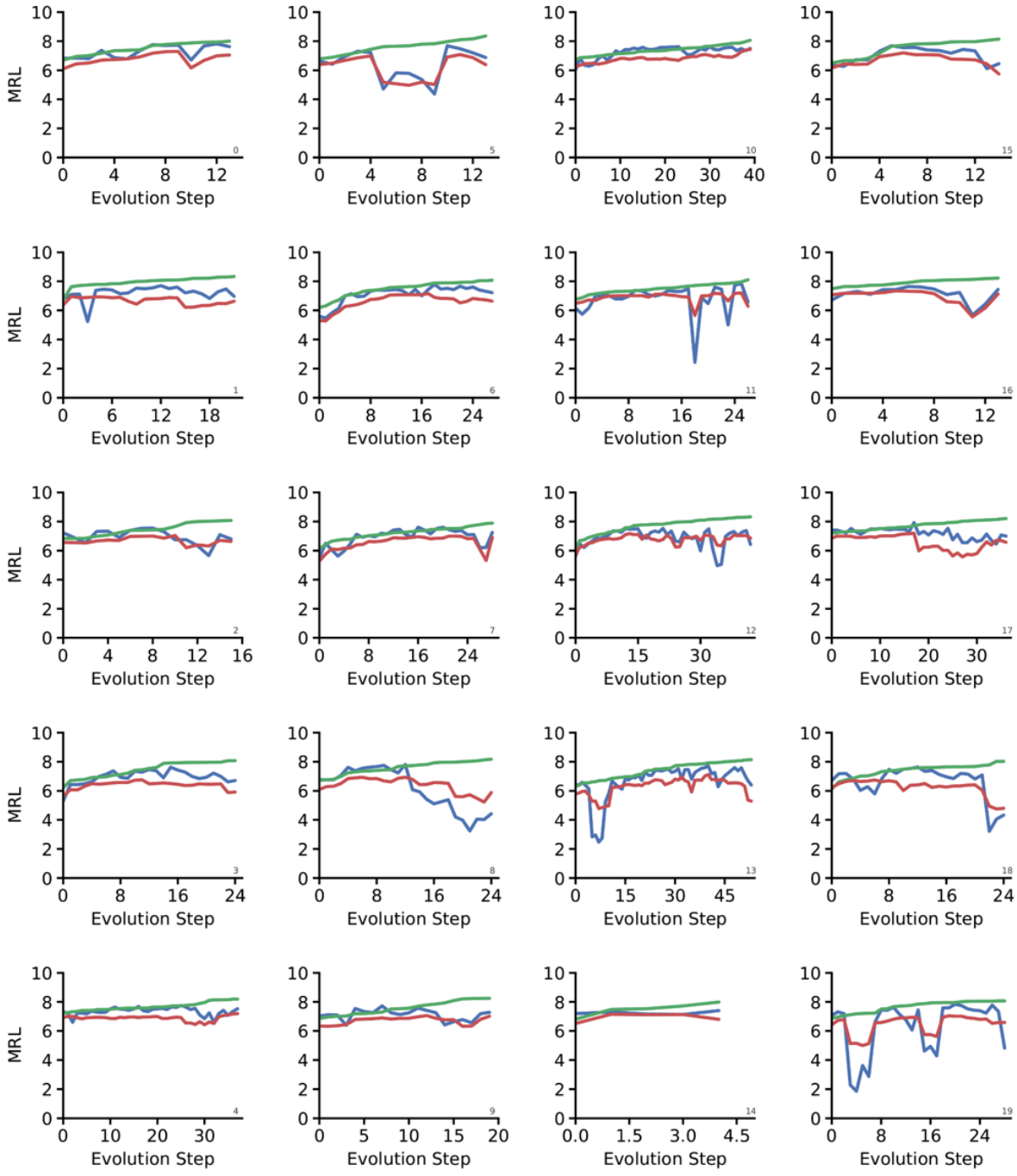


Figure B-3. UTRs were evolved to the highest ribosome load over 800 iterations allowing AUGs.

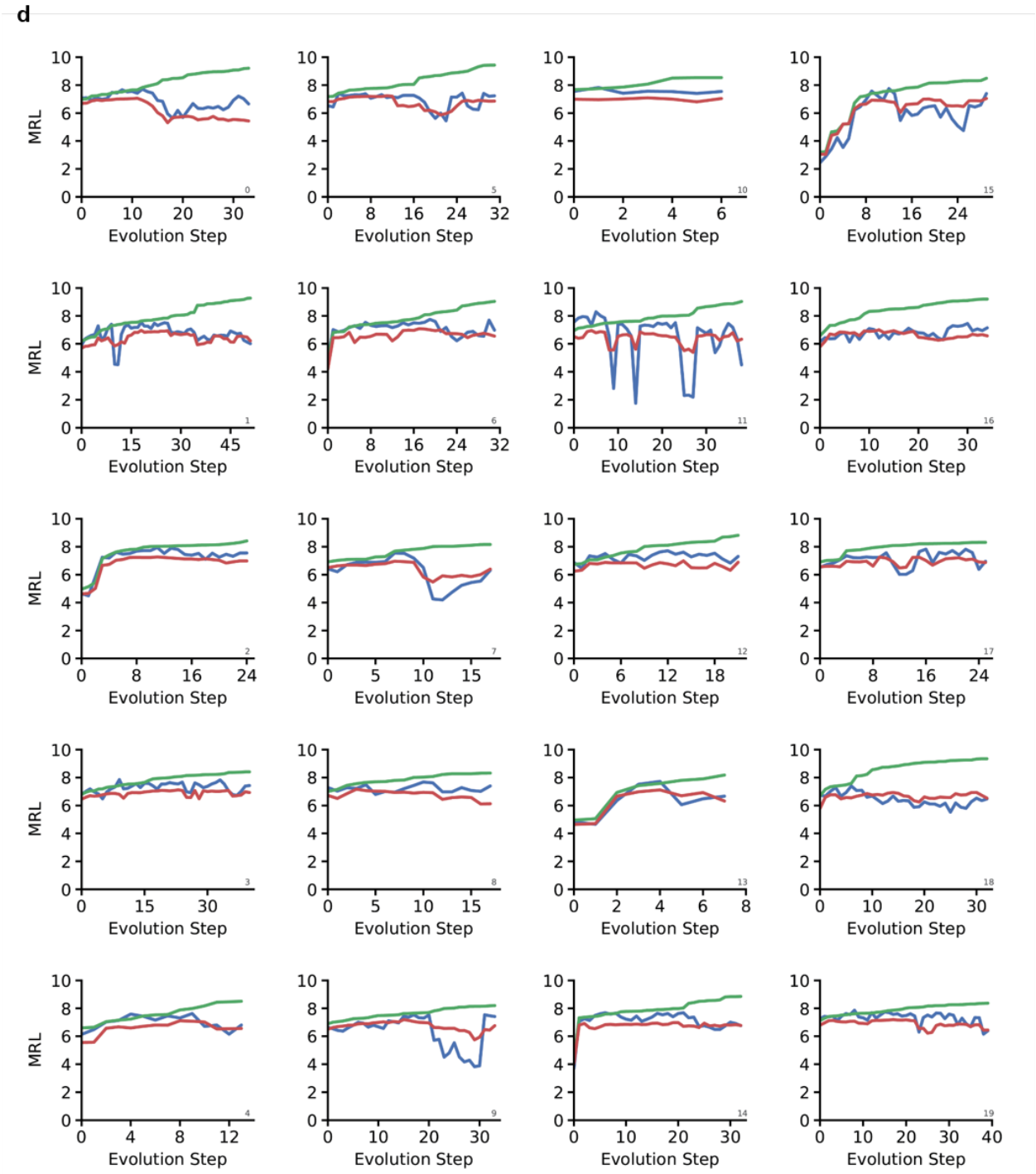


Figure B-4. UTRs were evolved to the highest ribosome load over 800 iterations without allowing AUGs.

VITA

Ban Wang received her Bachelor's degree in Engineering and Economics with a major in Measurement and Control Technology and Instrumentations and a minor in Economics at Xiamen University, Xiamen, Fujian Province, China in 2012. In her senior year as an undergraduate, she spent a quarter in University College London, London, UK as an exchange student. She received her Master's degree with a major in Electrical Engineering at Columbia University, New York City, NY in 2014. She started to get very interested in biological applications via an electrical engineer's view at Columbia, where her primary research focus was computational neuroscience and reversed engineering the fruit fly brain. Ban enrolled in the Electrical & Computer Engineering Ph.D. program at the University of Washington in 2014 and was advised by Prof. Georg Seelig. Her primary research interest is in developing high-throughput assay and applying machine learning algorithms to understand gene regulation.