UTRGAN: LEARNING TO GENERATE 5' UTR SEQUENCES FOR OPTIMIZED TRANSLATION EFFICIENCY AND GENE EXPRESSION

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Sina Barazandeh

July 2024

UTRGAN: Learning to Generate 5' UTR Sequences for Optimized Translation Efficiency and Gene Expression

By Sina Barazandeh

July 2024

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

A. Ercüment Çiçek(Advisor)

Can Alkan

Gülden Olgun

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

UTRGAN: LEARNING TO GENERATE 5' UTR SEQUENCES FOR OPTIMIZED TRANSLATION EFFICIENCY AND GENE EXPRESSION

Sina Barazandeh
M.S. in Computer Engineering
Advisor: A. Ercüment Çiçek
July 2024

The 5' untranslated region (5' UTR) of the messenger RNA plays a crucial role in the translatability and stability of the molecule. Thus, it is an important component in the design of synthetic biological circuits for high and stable expression of intermediate proteins. Several UTR sequences are patented and used frequently in laboratories. We present a novel model, UTRGAN, which is a Generative Adversarial Network (GAN)-based model designed to generate 5' UTR sequences coupled with an optimization procedure to ensure a target feature such as high expression for a target gene sequence or high ribosome load and translation efficiency. We rigorously analyze and show that the model can generate sequences that mimic various properties of natural UTR sequences. Then, we show that the optimization procedure yields sequences that are expected to yield (i) up to 5-fold higher average expression on a set of target genes, (ii) up to 2-fold higher mean ribosome load on average, and (iii) a 34-fold higher average translation efficiency, compared to the initially generated UTR sequences. We compare our method with other approaches and find that UTRGAN-generated sequences contain motives that show higher sequence similarity to known regularity motives in various regions such as (i) internal ribosome entry sites, (ii) upstream open reading frames, (iii) G-quadruplexes, (iv) Kozak and initiation start codon regions. Finally, we show in-vitro that the UTR sequences we designed yield a higher translation rate for the human TNF-α protein compared to the human Beta Globin 5' UTR, which is a 5' UTR with high production capacity.

Keywords: Generative Models, 5' UTR design, Deep Learning, Gene Expression, Translation Efficiency.

ÖZET

UTRGAN: OPTIMIZASYONLU ÇEVIRI VERIMLILIğI VE GEN EKSPRESYONU IÇIN 5 UTR SEKANSLARđNđ ÜRETMEYI ÖğRENMEK

Sina Barazandeh
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: A. Ercüment Çiçek
Temmuz 2024

Mesajcı RNA'nın 5' untranslated bölgesi (5' UTR), molekülün çevrilebilirliği ve kararlılığı üzerinde kritik bir rol oynar. Bu nedenle, sentetik biyolojik devrelerin tasarımında, ara proteinlerin yüksek ve kararlı ifadesi için önemli bir bileşendir. Birçok UTR dizisi patentlidir ve laboratuvarlarda sıklıkla kullanılmaktadır. Bu tezde, yeni bir model olan UTRGAN'ı sunuyoruz. UTRGAN, Generatif Rekabetçi Ağlar (GAN) tabanlı bir model olup, belirli bir gen dizisi için yüksek ifade veya yüksek ribozom yükü ve çeviri verimliliği gibi hedef özellikleri sağlamak amacıyla optimize edilmiş 5' UTR dizileri üretmek üzere tasarlanmıştır. Modelin, doğal UTR dizilerinin çeşitli özelliklerini taklit eden diziler üretebildiğini titizlikle analiz ediyor ve gösteriyoruz. Ardından, optimizasyon prosedürünün, (i) hedef genler setinde ortalama ifadeyi 5 kata kadar artırması, (ii) ortalama ribozom yükünü 2 kata kadar artırması ve (iii) başlangıçta üretilen UTR dizilerine kıyasla ortalama çeviri verimliliğini 34 kata kadar artırması beklenen diziler ürettiğini gösteriyoruz. Yöntemimizi diğer yaklaşımlarla karşılaştırıyor ve UTRGAN tarafından üretilen dizilerin, (i) iç ribozom giriş bölgeleri, (ii) yukarı yönde açık okuma çerçeveleri, (iii) G-quadrupleksler, (iv) Kozak ve başlatma kodonu bölgeleri gibi çeşitli bölgelerde bilinen düzenleyici motiflere daha yüksek dizilim benzerliği gösteren motifler içerdiğini buluyoruz. Son olarak, in-vitro olarak tasarladığımız UTR dizilerinin, yüksek üretim kapasitesine sahip bir 5' UTR olan insan Beta Globin 5' UTR'sine kıyasla, insan TNF-α proteini için daha yüksek bir çeviri oranı sağladığını gösteriyoruz.

Anahtar sözcükler: Üretici Modeller, 5' UTR tasarımı, Derin Öğrenme, Gen İfadesi,

Çeviri Verimliliği.

Acknowledgement

First and foremost, I would like to thank my supervisor, Prof. A. Ercüment Çiçek. I'm very happy to have completed my master's under his supervision, and I'm very satisfied with my experience in CicekLab.

I'm also profoundly grateful to my family and friends for their patience and support. All my achievements were possible because of them.

I also thank Dr. Urartu Özgür Şafak Şeker and Ahmet Hincer for their help with the experimental validation.

Lastly, many thanks to jury members Prof. Can Alkan and Prof. Gülden Olgun for allocating their valuable time to read and assess my thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

RNA-based therapeutics require both tunability and a prolonged effect after administration. To achieve this, optimizing both mRNA molecules and carriers is essential to enhance their stability and promote tissue-specific tropism [1]. The level of RNA and protein expression resulting from mRNA therapeutics is critical in various applications, including protein replacement therapies, genome engineering, genetic reprogramming studies, vaccinations, and cancer immunotherapies [2]. Targeted mRNA expression levels are particularly important in specific protein replacement therapies for conditions like hemophilia A and cystic fibrosis. For example, Chen et al. demonstrated that to achieve the necessary factor VIII expression in mice using mRNA-laden lipid nanoparticles (LNPs), a concentration of 2 mg of mRNA per kilogram of mouse body weight is required [3]. Similarly, a study on cystic fibrosis treatment found that a dosage of 0.1 mg/kg/day for two consecutive days was sufficient to restore the function of the CFTR gene in CFTR-knockout mice [4]. A phase 1 study on transthyretin amyloidosis used a dosage of 0.3 mg/kg in patients, resulting in the expression of Cas9 and the delivery of its single guide RNA to knock out

the transthyretin gene, representing a potential treatment for this disease [5]. Conversely, vaccination studies for viruses like Zika [6], Covid [7, 8], and influenza [9], or tumor-associated antigen expression in cancer immunotherapies [10], require a lower dosage of approximately 0.002-0.02 mg/kg to induce immunity against these viruses and tumor cells. Therefore, designing stable RNA molecules and being able to control expression levels are highly desirable [11–13].

Optimization of the 5' UTR is a preferable approach to control the expression level and the stability of the mRNA molecule, because it has been proved to be an important region in the sequence with respect to these features [14] [15]. While the untranslated region (UTR) does not encode protein, it is essential for governing the translation process as it harbors the ribosome binding site (RBS), which facilitates ribosome attachment and initiation of translation. Consequently, numerous UTR sequences have been patented and integrated into gene circuit design [16, 17].

The conventional method for optimizing UTR sequences involves introducing variations into existing target sequences and assessing their efficacy using basic algorithmic techniques. Von Niessen et al. developed a library of 3' UTR sequences derived from the human genome, focusing on enhancing transcript stability [17] [18] [19]. They utilized a genetic algorithm to generate synthetic sequences based on sequence attributes such as GC content, k-mer frequency, and free energy, selecting sequences based on their predicted translation efficiency [20]. Similarly, Lu et al. identified 5' UTRs influencing protein expression and discussed strategies for their optimal pairing [16]. Other studies have identified 3' and 5' UTRs that enhance gene expression, selecting samples from naturally occurring mRNA sequences in human tissues [21]. These models rely solely on machine learning algorithms to predict sequence characteristics and evaluate the generated samples. However, the full potential of generative modeling in the design and optimization of UTR sequences remains underexploited.

Numerous studies utilize Generative Adversarial Networks (GANs) to produce sequences that mimic natural DNA or RNA sequences. RNAGEN, for instance, facilitates the generation and refinement of synthetic piRNA sequences tailored to specific properties like protein binding affinity [22]. Zirmec et al. introduced Expression-GAN, a GAN-based framework designed for generating 1 kb-long regulatory DNA sequences encompassing promoter, 5'UTR, 3'UTR, and terminator regions [23, 24]. They applied this framework to optimize complete regulatory sequences, not limited to UTRs, leveraging a model that predicts yeast gene expression [25]. However, their approach imposes constraints on the length of generated regulatory regions (250 bp for 5' UTR and 1 kbp for the entire sequence), potentially limiting applicability to human genes with longer regulatory sequences such as the MECP2 gene, where the 3' UTR spans over 8 kb [26].

In a separate study, Castillo-Hair et al. explore the capabilities of integrating machine learning techniques and predictive models to enhance 5' UTR sequences [27]. As far as we are aware, no previous research has focused on generative models specifically designed to create and refine 5' UTRs according to diverse metrics.

This study introduces UTRGAN, the inaugural GAN-based framework designed for synthesizing novel human 5' UTR sequences and optimizing them to enhance Mean Ribosome Load (MRL), Translation Efficiency (TE), and overall gene expression levels. The model generates customizable 5' UTR sequences applicable to any target gene, natural or synthetic, while accommodating variable sequence lengths. UTRGAN operates exclusively on the coordinates of the Transcription Start Site (TSS) and 5' UTR regions to optimize gene expression or other specified objectives. Importantly, it achieves this without altering the underlying DNA/RNA sequence to which the UTR is appended. The generated sequences closely emulate natural 5' UTRs in terms of critical features such as GC content, k-mer distribution, and Minimum Free Energy (MFE) [28]. Through optimization iterations, UTRGAN significantly enhances MRL, mRNA expression by up to 53%, and gene expression by

up to 61% compared to baseline values for the designed sequences. Additionally, optimization for TE results in an average predicted increase of up to 34-fold. These results validate the efficacy of the optimization process. Depending on the application, the model can tailor the generated 5' UTRs for a specific DNA sequence or a set thereof, achieving an average 2.2-fold expression increase for single genes and up to 32-fold enhancement for the most effective synthetic 5' UTRs. Furthermore, the study assesses the similarity of UTRGAN-generated sequences to known regulatory motifs such as Internal Ribosome Entry Site (IRES) and Upstream Open Reading Frame (uORF) sequences. It demonstrates that UTRGAN-predicted sequences maintain essential motifs and regulatory elements akin to those found in natural sequences. Conversely, sequences generated by other methods, despite exhibiting high predicted MRL values, show much lower conservation of these motifs, implying potential non-functionality. In vitro experiments confirm the superior translation rate of UTRGAN-generated 5' UTR sequences compared to the native human -globin 5' UTR when fused with the TNF- gene. These findings underscore the practical utility and biological fidelity of UTRGAN in enhancing gene expression through tailored 5' UTR sequence optimization.

UTRGAN's capacity to create and refine 5' UTRs tailored to specific gene sequences promises to significantly advance genetic circuit design. We believe UTRGAN will play a pivotal role in facilitating mRNA-based therapeutics across diverse applications in the biotech industry, particularly in areas demanding precise control over gene expression, such as cancer immunotherapy.

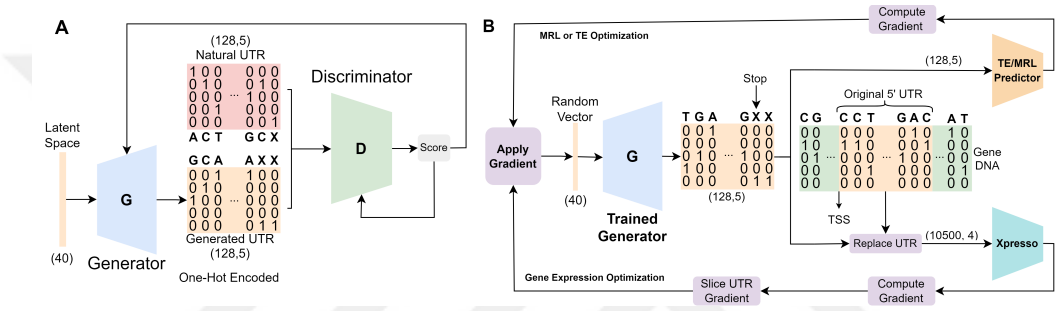

Fig. 1.1: **Generative and optimization architectures.** Shown on panel **A** is the training phase of the GAN model. The feedback resulting from the competition between the critic and the generator updates the weights of both components. The panel **B** shows the optimization procedure. The GAN is used to generate samples, and the input noise is updated using the Gradient Ascent algorithm to increase the predicted MRL, TE, or mRNA abundance of the generated 5' UTR sequences. The gene expression optimization procedure (Xpresso model) requires attaching the generated 5' UTR to a DNA sequence to get feedback. We slice and apply the relevant gradient to update the input to generate the UTR sequence (the random vector). The MRL and TE optimization require only the UTR sequence as their input.

Chapter 2

Methods

2.0.1 Datasets and the Experimental Setup

The natural 5' UTR sequence data utilized in this study is sourced from UTRdb 2.0 [29], a comprehensive database containing sequences from various organisms. Specifically, we focus on the Human 5' UTR samples, which range in length from 30 bp to 600 bp. Given that this wide length range poses challenges for convolutional neural network (CNN) models—where sequences must be padded to match the longest sequence length, often resulting in excessive padding—we constrain the training data to sequences between 64 bp and 128 bp. This subset comprises 33,250 natural 5' UTR sequences. Training involves using 90% of these sequences for training and the remaining 10% for validation, with each training batch consisting of 64 sequences. Additionally, DNA sequences of Human genes intended for optimization are obtained from the Ensembl Biomart tool [26].

We employ one-hot encoding to represent the data, where each UTR is formatted as $(128, 5)$, with 128 being the maximum sequence length and 5 representing the dimensionality of the one-hot vector for each nucleotide. The DNA nucleotides

are encoded as follows: A:(1,0,0,0,0), C:(0,1,0,0,0), G:(0,0,1,0,0), T:(0,0,0,1,0), and X:(0,0,0,0,1), where X denotes a padding character used when sequences are shorter than 128 bp. This padding ensures compatibility with the fixed dimensions required by the Critic, a convolutional neural network, given the variable lengths of natural sequences. To generate truly random sequences, we randomly select a length between 64 and 128 nucleotides and choose each nucleotide randomly from the DNA alphabet. Our analysis encompasses 2048 randomly selected sequences, 2048 generated sequences, and all 33,250 natural sequences unless otherwise specified. Further technical specifications, including hardware details, are provided in Supplementary Notes 2.1, while hyperparameter optimization details can be found in Supplementary Notes 2.2. The code repository for this project is accessible at https://github.com/ciceklab/UTRGAN.

2.0.2 The Model

Overall Architecture

The generative model employed in this study is a variation of a Generative Adversarial Network (GAN) [30]. In this setup, the GAN takes a random noise sample z as input and is trained to produce realistic 5' UTR sequences. The GAN effectively learns to transform the input noise into the space of 5' UTR sequences. Subsequently, we can refine the model to generate specific sequences by optimizing the input noise alone. This optimization process focuses solely on adjusting the input noise z without updating the generator's weights.

The sequences generated by the GAN undergo evaluation using selected scoring models. Three deep convolutional neural networks are employed for this purpose. The mRNA abundance scoring model [31] estimates the log TPM expression of a gene sequence, including its 5' UTR. Xpresso offers multiple models for predicting median expression across various cell types, specifically K562 erythroleukemia cells

and GM12878 lymphoblastoid cells [31]. For this study, we utilize the median prediction model, although optimization can target specific cell types using these models. Models assessing ribosome load and translation efficiency focus solely on the 5' UTR sequence as input [32] [33]. Specifically, the MTtrans 3R model predicts TE values as a proxy for translation rate [33]. All models, along with the generator, are differentiable, enabling optimization of the input noise z through Stochastic Gradient Ascent across multiple iterations. Sequences are fine-tuned over 3,000 iterations for optimizing expression and 10,000 iterations for optimizing MRL and TE. The performance of each iteration is recorded, and the version exhibiting the highest performance is selected as the final optimized iteration of that sequence.

We also attempted to train a VAE for designing UTRs, as VAEs offer a latent space that can be utilized for optimization. Initially, training a VAE with one-hot encoded input sequences, similar to UTRGAN, was unsuccessful as the model did not converge. In a subsequent attempt, we used byte-pair encoding (BPE) [34] to tokenize UTR sequences. However, the model struggled to learn the padding patterns and often failed to generate sequences with padding only at the end. Ultimately, our best efforts resulted in a model that produced sequences with semantically meaningless content or highly biased GC content, predominantly around 100 percent.

2.0.2.1 Generative Model Architecture

The GAN utilized in this study is based on the original Wasserstein GAN (WGAN) [35], which is a Convolutional GAN model [36] incorporating the Wasserstein loss [37] and Gradient-Penalty for enhanced training stability [38].

The deep learning model's layers were adapted to suit the task, considering the data dimensions: 128 nucleotides, each represented by a vector of length 5. The generator G includes a dense layer to expand the input noise dimension, followed by five residual convolution blocks with a uniform number of channels. The final residual block's output is fed into a convolution layer with five output channels, followed by

a softmax layer, resulting in a one-hot encoded sequence. Supplementary Table 6 details the generator's layer dimensions. The input to the generator is a random vector of length 40, sampled from a normal distribution with a mean of zero and a standard deviation of 1.0. The discriminator, technically referred to as the critic D, is a convolutional neural network with five residual blocks, culminating in a dense layer with a single output. Supplementary Table 7 provides details on the critic's layers. Unlike the original DCGAN model [36], there is no activation function at the critic's dense layer output, as the output is an unbounded score rather than a probability [35]. The critic's parameters are updated five times for each update of the generator, meaning the critic is updated five times more frequently than the generator.

The initial weights of the network are sampled from a normal distribution with a mean of 0 and a standard deviation of 0.1. In the loss function, as illustrated in Equation 2.1, the generator is trained to minimize the distance between the distribution of the critic's scores for natural and generated sequences. The first part of the loss function corresponds to the original WGAN model [35], while the second part incorporates the gradient penalty and its coefficient λ. In this context, x and $\tilde{x}$ represent the natural and generated samples, respectively, with $\hat{x} \sim P_{\tilde{x}}$ sampled uniformly along straight lines between pairs of points from the data distribution $\mathbb{P}r$ and the generator distribution $\mathbb{P}g$ for computing the gradient penalty. The gradient penalty enhances GAN training by enforcing the Lipschitz constraint, as suggested by the model [38], thereby allowing for a higher number of training iterations. In this study, the model is trained for 3,000 epochs.

$$L_{adv} = (\mathbb{E}_{\tilde{x}\sim P_g}[D_{\tilde{x}}] - \mathbb{E}_{x\sim P_r}[D_x]) + \lambda(\mathbb{E}_{\hat{x}\sim P_{\hat{x}}}\left[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2\right]) \tag{2.1}$$

2.0.2.2 Expression Prediction Optimization

The Xpresso model can predict the expression of a one-hot encoded DNA sequence with a fixed length of 10,500 nucleotides. It takes sequences comprising 3,500 nucleotides downstream and 7,000 nucleotides upstream of a Transcription Start Site (TSS) and predicts the log TPM expression for these sequences.

Let z_0 denote the initial latent vector. The generator uses this latent vector $G(z_0)$ to generate the 5' UTR, referred to as U_g. S represents the gene sequence for which we are designing the 5' UTR, with the original 5' UTR denoted as U_r. The expression prediction model is denoted as $E(S)$, where S is the input sequence.

To perform the optimization, the initial 5' UTRs are generated using a fixed seed for the random latent vector. These generated UTRs are then attached to the selected gene sequences, replacing the original 5' UTRs, and the initial expression is measured.

We define R as the function to replace the original 5' UTR U_r with the generated 5' UTR U_g while maintaining the length of S and the TSS position in the sequence (see Equation 2.2). In the DNA samples, the original TSS position is the 7000th nucleotide from the start. The length of a sequence is determined using the len function, and bracket notation is used to slice the sequences.

$$R(U_g, S) = S[0 : 7000] + U_g + S[7000 + len(U_g) : (10500 - (len(U_g) - len(U_r)))] \tag{2.2}$$

To perform Gradient Ascent on the input latent vector for optimization, we first calculate the gradient of the Xpresso model's output with respect to its input DNA sequences (Equation 2.3). We then extract the portion of the gradient corresponding to the 5' UTR, located between the 7000th and 7128th indices, depending on L, the

length of the generated 5' UTR. In the next step, the gradient of the generator is calculated with respect to the input noise vector.

$$\frac{\partial E(R(G(z_0), S))}{\partial z_0} = \frac{\partial E(R(G(z_0), S))}{\partial R(G(z_0), S)}[7000 : 7000 + L] \times \frac{\partial G(z_0)}{\partial z_0} \tag{2.3}$$

Updating the noise vector according to the gradient is expected to increase the expression in the following iteration. This optimization iteration is repeated 3,000 times through updating the latent vector $z_{new} = z_0 + \frac{\partial E(R(G(z_0),S))}{\partial z_0}$. We store the history of predicted expression values for each iteration and select the sequence with the best-predicted value as the optimized version. Although the optimization is conducted on a batch of sequences, the final sequences are chosen independently, as if the optimization had been performed separately for each element of the latent vector. This approach allows for batch optimization rather than optimizing individual UTRs one at a time.

2.0.2.3 Mean Ribosome Load and Translation Efficiency Prediction Optimization

The FramePool model was trained on UTR sequences up to 100 bp in length. It uses the Frame-slice layer, which enables it to predict the MRL value for 5' UTR sequences of any length [32]. Frame-slicing divides the sequence into 3-mers or codons, known as frames, and applies global pooling on the indices, allowing it to handle sequences of any size. For optimization with this model, we again use the Gradient Ascent algorithm. Slicing is not needed because the model inputs only the UTR, and the derivatives are taken directly on the 5' UTR sequence. With the MRL prediction model denoted as M and the initial latent vector as z_0, the gradient is calculated as shown in Equation 2.4.

$$\frac{\partial M(G(z_0))}{\partial z_0} = \frac{\partial M(G(z_0))}{\partial G(z_0)} \times \frac{\partial G(z_0)}{\partial z_0} \tag{2.4}$$

We update the input vector by applying the gradient as $z_{new} = z_0 + \frac{\partial M(G(z_0))}{\partial z_0}$. This process is repeated for 10,000 iterations, after which the optimized sequences are selected in the same manner as described for expression maximization in Section 2.0.2.2. Similarly, TE optimization using MTtrans *3R* follows the same procedure, with up to 10,000 iterations, and the best-performing sequences are selected as described previously.

Chapter 3

Results

3.0.1 Overview of UTRGAN

The UTRGAN model is a deep generative adversarial network (GAN) designed to generate 5' UTR sequences that resemble natural ones. This model is a variant of the WGAN-GP architecture [38], tailored for one-hot encoded DNA sequences. It offers improved training stability compared to the original Convolutional GAN (DCGAN) models [36, 38] and is less susceptible to overfitting than the original WGAN architecture [35].

Figure 1.1A provides an overview of the architecture. The model takes a noise vector as input. Both the generator and the critic are convolutional neural networks trained in tandem. The generator uses transpose convolutions to upsample the input vector and generate 5' UTR sequences, while the critic employs convolutions and dense layers to differentiate between natural and synthetic 5' UTRs. The feedback from the critic enables the generator to improve its output, producing 5' UTRs that more closely mimic natural sequences. The optimization pipeline is illustrated in Figure 1.1B. We use pre-existing deep-learning models for optimization. To enhance

the gene expression of a target RNA sequence, we utilize Xpresso [31], which predicts the expression of the entire sequence, including the UTR region. For optimizing the MRL of a UTR sequence, we use FramePool [32], which predicts this value based solely on the UTR sequence, independent of the RNA sequence it is part of. Similarly, MTrans [33] is employed to predict the translation efficiency of a UTR sequence. Optimization is achieved by updating the initial input to the GAN model, generating UTR sequences until the generated sequence reaches its maximum predicted value, independent of other sequences in the batch. Specifically, gradient ascent is applied to the generator's input based on feedback from the target feature predictor (MRL, translation efficiency, or gene expression).

In the subsequent sections, we perform various analyses to demonstrate that the generated 5' UTR sequences resemble natural 5' UTRs and compare our performance with other methods. UTRGAN is the sole generation tool used in our approach, so all generated sequences are produced by UTRGAN. These sequences are optimized by our method and are referred to as UTRGAN-optimized sequences (with expression, MRL, or TE as the target feature). Additionally, we compare our approach with a randomized sequence optimization algorithm, Optimus 5-Prime. For this comparison, we truncate our generated sequences to the 50 bp input length required by this method and use them as the initial input sequences. Sequences optimized using this method are referred to as Optimus 5-Prime-optimized.

3.0.2 Levenshtein distances are similar in natural and generated UTRs

A metric for assessing similarity between two sets of sequences is to examine the distribution of distances to the nearest sequences in a target set, such as the natural UTR set. The Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one

sequence into another [39]. In this analysis, we compare three sets: (i) UTRGAN-generated sequences (no optimization, n=1204, (ii) 1024 UTRGAN-generated sequences optimized for MRL (n=1024), and (iii) sequences generated and MRL-optimized by Optimus 5-Prime [32] (n=1024).

We use the natural 5' UTR dataset (n=33,250) as the target set. Specifically, we calculate the distribution of Levenshtein distances from sequences in each of the above sets to the nearest sequence in the natural UTR dataset. This results in three distributions, one for each sequence class: UTRGAN-generated, UTRGAN-optimized, and Optimus 5-Prime-optimized sequences. For comparison, we also calculate the distribution for the natural UTR set against itself to establish a baseline. Any natural sequences with unusually low distances to the natural UTR set are excluded from the plots. Figure 3.1A illustrates that the distributions for UTRGAN-generated sequences and natural sequences are nearly identical. Additionally, the UTRGAN-optimized sequences retain their proximity to the natural sequences while preserving a broad range of sequence lengths (see Supplementary Figure 5).

In contrast, sequences optimized using Optimus 5-Prime have a fixed length of 50 bp by design, resulting in less diversity in their distances from natural UTR sequences.

3.0.3 The generative model maintains the 4-mer distribution of the sequences

Another metric for assessing similarity between two sets of sequences is the distribution of k-mers within the sequences. We calculate the frequency of 4-mers for each sequence, following methods outlined in the literature [40]. For each sequence in the natural, generated, UTRGAN-optimized, and Optimus 5-Prime-optimized sets, we identify the closest sequence in the entire set of natural 5' UTRs based on Euclidean distance, excluding identical sequences (so natural sequences match their

nearest neighbor). This process yields a distribution for each group. Sequences with unusually low distances to the natural set are excluded from the analysis.

Figure 3.1B shows that the distance distributions for the generated sequences and natural sequences are very similar, indicating that the generated sequences retain structural similarity to natural sequences while still being distinct in terms of their specific sequences, as discussed in the previous subsection. Conversely, sequences optimized by Optimus 5-Prime exhibit a different 4-mer distribution, despite being similar to natural sequences in sequence content.

3.0.4 GC content distribution of the generated 5' UTRs resemble natural 5' UTRs

The GC content of a sequence is crucial for molecular stability [41]. We observe that the mean GC content in the generated sequences is very similar to that in natural sequences, with the generated sequences displaying a diverse range of GC content values (Figure 3.1C). The UTRGAN-optimized sequences exhibit a distribution similar to natural sequences but with a lower mean GC content, suggesting a connection between GC content stability and MRL. In contrast, the GC content of sequences optimized by Optimus 5-Prime follows a substantially different distribution.

3.0.5 Predicted Mean Ribosome Load and translation efficiency for the natural and generated sequences have similar distributions

While the previously mentioned metrics provide valuable insights into the model's performance in generating structurally similar samples to natural sequences, they do not convey functional information. MRL, a metric based on the ribosome count associated with an mRNA molecule, is considered a proxy for translation rate [32]. We use MRL to assess the functional similarity of the sequences.

To estimate the MRL of a given 5' UTR sequence, we employ a convolutional neural network model (FramePool) that predicts the MRL of a 5' UTR sequence [32]. This model uses a Frame-Pooling layer to process frames of three consecutive nucleotides (codons). Figure 3.1 D shows that the MRL distributions of the natural and generated sequences are almost identical. We also present the predicted MRL distribution of the UTRGAN-optimized and Optimus 5-Prime-optimized sequences. As expected, both sets of sequences have a very high average MRL, as they both use the FramePool [32] model to guide the optimization. However, Optimus 5-Prime fails to improve the expected MRL for all sequences, as evidenced by the very long low tail in the distribution. This method introduces random mutations in the UTR sequence, and unlike our gradient-based optimization, this may degrade the MRL for many sequences.

Similarly, we use the MTtrans model, trained on ribosome profiling datasets, to predict the translation efficiency of UTR sequences [33]. Translation efficiency is another metric representing the rate at which a cell translates mRNA. MTtrans is trained on various translation profiling datasets, including (i) three massively parallel assay (MPRA) polysome profiling datasets and (ii) three ribosome profiling datasets. The model trained on MPRA datasets is referred to as *3M*, the model trained on ribosome profiling datasets is called *3R*, and the model trained on all datasets is named *3M3R*. We use the MTtrans 3R version. As shown in Figure 3.1E, the predicted TE for generated and UTRGAN-optimized sequences falls within the same range as the natural sequences. In contrast, the sequences optimized by Optimus 5-Prime have a lower average predicted TE and a distribution that differs from the natural sequences.

3.0.6 Generated and natural samples have similar Minimum Free Energy distributions

The minimum free energy (MFE) of a 5' UTR sequence, as part of the RNA sequence, is another important characteristic related to its stability. MFE depends on the number, composition, and arrangement of nucleotides in the mRNA sequence [28], specifically within the 5' UTR. Although MFE alone may not fully characterize a 5' UTR sequence, we expect to see similar MFE distributions across a large sample set. We calculate the MFE of the sequences using two packages: Nupack [42] and ViennaRNA [43, 44], and present the results from ViennaRNA, as both yield similar outcomes.

As shown in Figure 3.1F, the MFE distribution in generated sequences closely matches that in natural sequences, whereas sequences optimized by Optimus 5-Prime show a markedly different distribution. While the UTRGAN-optimized sequences do not exactly match natural sequences in MFE distribution, they exhibit a much broader range of MFE values compared to those optimized by Optimus 5-Prime.

3.0.7 Optimization Yields Sequences with Higher Expected Expression

The optimization procedure, as detailed in Section 1, involves iteratively updating the input noise of the GAN to generate 5' UTR sequences with higher predicted expression. We use a convolutional neural network-based model (Xpresso [31]) to predict log TPM expression for the input DNA sequences. The Xpresso model's mean saliency scores [31] indicate which nucleotides around the TSS most influence predicted expression, with the most important ones being downstream of the TSS where the 5' UTR is located. Based on this, we focus on optimizing the 5' UTR to enhance expression values.

As shown in Figure 3.2A, optimizing 100 generated 5' UTR sequences for maximizing average expression in eight randomly selected human genes (*MYOC*, *TIGD4*, *ATP6V1B2*, *TAGLN*, *COX7A2L*, *IFNGR2*, *TNFRSF21*, *SETD6*) results in increased average predicted mRNA expression in 80% of the samples with optimized 5' UTRs. To calculate a single score for each 5' UTR, we average the expression over all eight genes with the generated 5' UTRs replacing the original ones. Although limited to updating the 5' UTR alone, we observe up to a 61% increase in average predicted gene expression compared to the initial values. In 20% of cases, optimization results in reduced expression compared to the initial generated sequence, allowing users to retain the initially generated 5' UTR if necessary. In the Discussion Section, we discuss possible reasons for these decreases. Additionally, our optimized 5' UTRs yielded a 38% increase in the average predicted expression of another set of eight randomly selected genes (*ANTXR2*, *NFIL3*, *UNC13D*, *DHRS2*, *RPS13*, *HBD*, *METAP1D*, *NCALD*), demonstrating the optimization's ability to generalize 5' UTRs for higher average gene expression.

3.0.8 Mean Ribosome Load and Translation Efficiency increases after optimization

Unlike mRNA abundance, the MRL is not DNA-specific and does not require a target DNA or mRNA sequence for optimization. The FramePool model [32] predicts the MRL value for 5' UTR sequences of any length without needing the UTR sequence to be attached to an mRNA sequence. We used this model to optimize the initially generated 64 5' UTRs over 10,000 iterations. The results indicate an increase in MRL for more than 95% of the generated 5' UTRs (see Figure 3.3A). Although the number of sequences can be increased, we expect similar results since each sequence is optimized independently and gradients are not combined. Supplementary Figure 2 further demonstrates that the model learns a meaningful latent space regarding the predicted MRL values for both generated and optimized sequences.

We also use the MTtrans model [33] to optimize our generated 5' UTRs for higher translation efficiency (TE). Translation efficiency, another proxy for the rate at which a cell translates mRNA, is predicted by MTtrans, which is trained on various translation profiling datasets, including (i) 3 massively parallel assay (MPRA) polysome profiling datasets and (ii) 3 ribosome profiling datasets. The model trained on MPRA datasets is called *3M*, the model trained on ribosome profiling datasets is called *3R*, and the model trained on all datasets is called *3M3R*. Here, we use the MTtrans *3R* model for optimizing generated 5' UTRs for higher TE. The results in Figure 3.3B show a significant increase in TE values post-optimization. To combine the optimizations for mRNA abundance and translation efficiency, one can first optimize sequences for gene expression and then select those with higher MRL or translation efficiency. Our two-step optimization results are presented in Section 3.0.11.

Additionally, the MTtrans 3M model extracts important motifs from the MPRA datasets used for training [45]. It outputs 256 7-mers as motifs that influence the predicted MRL either positively or negatively. During MRL optimization using the FramePool model, we observed a 53% average reduction in the number of motifs negatively affecting MRL, while positive motifs were mostly preserved or increased. This indicates that our model's optimization eliminates negative motifs based on the gradients backpropagated from the MRL prediction model.

To further analyze the optimization results, we compare different characteristics of the generated sequences and the optimized ones for both TE (see Supplementary Figure 3) and MRL (see Supplementary Figure 4) optimizations. Both TE and MRL optimizations favor longer UTR sequences and lower GC content. In terms of MFE, both optimizations slightly reduce the average absolute MFE of the sequences. However, we do not see a clear correlation between the predicted TE and MFE values in either of the two cases.

3.0.9 Optimization improves the expected expression for specific target genes

As discussed earlier, our optimization procedure increases gene expression when optimizing multiple generated 5' UTRs for multiple gene sequences, based on the average. However, we can also target a single gene. To optimize 5' UTRs for a specific gene, we generate 64 sequences using a fixed seed and optimize them for higher expression when placed in the target gene's UTR.

We applied this method to the *TLR6*, *IFNG*, *TNF*, and *TP53* genes from the human genome. The *TLR6* gene, discussed by Noreen and Arshad [46], plays a role in regulating both innate and adaptive immunity, thus affecting the immune response. The *IFNG* (*IFN-*γ) gene encodes interferon-gamma, crucial for immune response [47, 48]. Similarly, *TNF* or *TNF-*α encodes cytokines essential for the immune system's inflammatory response. The *TP53* gene is a key tumor suppressor encoding the P53 protein [49, 50]. Controlling these genes' expression has significant implications in immunology and oncology. These examples demonstrate the model's capacity to enhance the expression of specific target genes. This method can be used to optimize the expression of any gene, either increasing or decreasing it.

Figure 3.4 shows that optimizing these genes improves gene expression for more than 95% of the generated 5' UTRs, resulting in a 4-fold, 37%, 4.2-fold, and 63% increase in expression on average for each gene, respectively. These results indicate successful optimization on average, allowing users to select the top sequence for their application. The highest expression increases for each gene are 8-fold, 3-fold, 32-fold, and 3.3-fold, respectively. Overall, this optimization leads to a 2.2-fold increase in the average expression of the four genes. The sequences of the best-performing 5' UTRs are provided in Supplementary Table 1.

Additionally, we observe that the GC content of sequences increases with gene

expression optimization, with the best-performing sequences exhibiting GC content as high as 85%. Some natural sequences even exceed this number. Although high GC content can boost mRNA expression levels [51] [52], it may not always be desirable due to stability concerns [41]. To address this, we incorporate a control mechanism in our optimization process, allowing us to limit the GC content of the best-performing sequences. For results with a maximum GC content of 65%, see Supplementary Figure 1. This approach still significantly increases the expected expression of all four genes.

3.0.10 Synthetic 5' UTRs yield higher predicted expression for target genes compared to their natural 5' UTRs

Our primary objective in this study is not to replace the existing 5' UTRs of natural genes. The experiments conducted in previous sections, whether focused on a single gene or a set of target genes, demonstrate that our optimization mechanism can adjust the initially generated (synthetic) 5' UTR sequences to achieve a specific goal (e.g., maximizing expression) by using natural gene sequences as templates or examples when their UTR sequences are substituted with the generated and optimized UTR sequences. Instead, our approach is aimed at creating novel and synthetic DNA molecules with specific desired properties, such as binding to another protein post-translation.

Nevertheless, in this section, we examine whether the 5' UTRs we generate result in higher predicted expression compared to natural 5' UTRs. For each of the *TLR6*, *IFNG*, *TNF*, and *TP53* genes, we generate 64 5' UTR sequences, select the one that maximizes the expected expression, and compare this value with the expected expression of the gene using its natural 5' UTR. We find that our generated sequences yield increases in expected expression by 8-fold, 3-fold, 32-fold, and 3.3-fold, respectively.

3.0.11 Joint optimization results in both higher translation efficiency and gene expression

As demonstrated in Sections 3.0.7 and 3.0.8, independent optimizations for translation efficiency (TE) and mRNA abundance each yield positive results. However, optimizing one metric does not necessarily improve the other. Here, we address the case of optimizing 5' UTR sequences for both higher translation rate and mRNA abundance for a target gene, using the same four genes discussed in Section 3.0.9. To achieve this, we employ a sequential optimization procedure. Since TE optimization is gene-agnostic and mRNA abundance optimization is gene-specific, we first optimize the sequences z for higher translation efficiency over 1,000 iterations. Then, we use the TE-optimized z as the starting point for mRNA abundance optimization.

Initially, the optimization increases the average TE from negative to positive values, specifically from 0.27 to 30.9. Subsequently, the TE-optimized sequences undergo further optimization for mRNA abundance specific to each target gene over another 1,000 iterations. While this second optimization step slightly reduces the average translation efficiency from 1.49 to just above zero for certain genes, both TE and mRNA abundance remain higher compared to the initially generated sequences, as shown in Figure 3.5. Additionally, our two-step optimization outperforms simultaneous optimization for these two metrics. The sequences of the best-performing 5' UTRs after joint optimization for these genes are provided in Supplementary Table 2.

3.0.12 Regulatory Elements are Conserved in UTRGAN-generated and -optimized Sequences

One of the key regulatory elements in 5' UTRs are upstream open reading frames (uORFs), which serve various functions such as translation reinitiation [53–55]. We acquired the set of known human uORF sequences from uORFdb [56] (n = 2,422,112)

and searched for uORFs ranging from 7bp to 64bp, identifying 402,140 sequences within our set of interest. Figure 3.6A illustrates the number of uORFs found in each sequence set. Since the set of natural UTRs is larger, the counts are normalized for comparison. We found a substantial number of uORF elements in the UTRGAN-generated and optimized sequences, whereas very few were retained in the sequences optimized by Optimus 5-Prime. The figures in Figure 3.6 represent the number of these elements per 1,024 sequences. Given that all sequence sets contain 1,024 sequences, except for the natural 5' UTRs, we scaled the element count for the natural 5' UTRs to ensure comparability.

Internal ribosome entry site (IRES) sequences are another set of RNA elements that facilitate cap-independent translation during protein synthesis [57]. These sequences are predominantly found in 5' UTR regions and are used by cells to enhance the translation rate of specific genes [58]. We sourced human IRES elements from the IRESBase database [59] and compared our sequence sets against them using maximum pairwise alignment scores. Due to the relatively long length of IRES sequences (over 128bp), we used alignment for comparison. Figure 3.6D demonstrates that UTRGAN-optimized sequences exhibit significantly higher alignment scores compared to the generated sequences and those optimized by Optimus 5-Prime.

Additionally, we examined whether the generated sequences are enriched with elements such as Kozak sequences and G-quadruplexes (G4) within the sequence sets of interest. G4s are structural elements in G-rich regions of the 5' UTR that influence mRNA stability [60, 61]. We identified G4 structures using the G4Boost model [62]. Kozak sequences are typically found in the non-coding region upstream of the translation initiation site [63]. Besides the consensus Kozak sequence (GCCGCCRCCAUGG), alternative sequences are known to have a similar regulatory function [64, 65]. However, many natural 5' UTRs do not contain any Kozak sequences, as shown in Figure 3.6C. We searched for exact matches of these alternative Kozak sequences with various initiation start codons ('AUG', 'AUA', 'CUG',

and 'GUG') within the sequences [65, 66]. As depicted in Figure 3.6A and Figure 3.6B, respectively, UTRGAN-generated and optimized sequences contain numerous instances of G-quadruplexes and Kozak sequences, while these elements are rarely found in sequences optimized using Optimus 5-Prime.

3.0.13 Motif Analyses Provide Insights on Learned Regulatory Patterns

We conducted a motif analysis on 1,024 UTRGAN-optimized sequences to identify novel motifs. Using the MEME suite [67, 68], which identifies enriched motifs within a sequence set, we retrained the top 50 motifs ranging from 7 to 15 base pairs in length for each set of interest (UTRGAN-generated/optimized and Optimus 5-Prime-optimized). This analysis revealed that some motifs either disappear or change during the optimization process, as the de novo motif identification yields different motifs for the generated and optimized sequences.

We applied the same tool to identify the top 50 motifs in natural 5' UTRs, human uORF sequences, and IRES sequences. Subsequently, we used the TomTom motif comparison tool [69] to find the top three motif matches between these groups (e.g., UTRGAN-optimized vs. IRES). Figure 3.7 displays the top matches between natural sequences (5' UTRs, uORFs, and IRES) and UTRGAN-optimized sequences (with additional comparisons for UTRGAN-generated sequences shown in Supplementary Figure 6). Our findings indicate that while some motifs in generated sequences may disappear or be modified during optimization, UTRGAN effectively preserves significant motifs from natural sequences. Moreover, the optimization process leads to meaningful alterations in the regulatory elements present, enhancing the efficiency of the target sequences.

We also note that the UTRGAN-optimized sequences exhibit a greater abundance of 'U'-rich motifs compared to the initially generated sequences, which are

predominantly 'A'-rich. Research indicates that 'U'-rich sequences are often over-regulated in cancerous cells, and there is a correlation between increased 'U' content and higher translation rates in associated genes [70,71]. Consistent with this, our optimization approach targets motifs with elevated 'U' content to enhance translation efficiency.

3.0.14 The cytotoxic effect of TNF-α proteins containing synthetic UTRs exhibit higher translation

We evaluate the performance of the designed UTRs through in vitro experiments, comparing the translation efficiency of TNF-α protein using (i) the UTRGAN-generated/optimized 5' UTRs and (ii) the human -globin 5' UTR. TNF-α is a multifunctional cytokine that can induce cell death through cytotoxicity in certain target genes. The MCF-7/MX cell line, known for its sensitivity to TNF-α, serves as our model. We use the cytotoxic effects on MCF7 cells to gauge the translation efficiency of the synthetic UTRs. Details of the experimental procedure are outlined in Supplementary Note 3.

The 5' UTR sequences are optimized based on the mean ribosomal load (see Methods for specifics). To ensure accurate translation initiation, we manually include the consensus Kozak sequence (GCCGCCACCAUGG) at the 3' end of both synthetic and natural 5' UTRs. We then construct an mRNA that includes the 5' UTR and the complete protein sequence, encompassing the first 76 amino acids of TNF-α, which acts as a leader sequence for natural protein secretion [72]. This mRNA is transcribed in vitro and used to transfect HEK293T cells, which are proficient in protein production and have the necessary secretion elements, including ADAM10 and ADAM17 sheddases [73].

All experiments are conducted under rigorously controlled conditions to ensure that any observed effects are solely due to the 5' UTRs. Given TNF-α's ability to

induce cytotoxicity and inhibit proliferation in MCF7 cells, we design an assay to measure TNF-α production by comparing cell viability across the different mRNA samples. As shown in Figure 3.8, the UTRGAN-generated and optimized 5' MRL UTRs demonstrate a significantly higher cytotoxic effect compared to the human -globin 5' UTR, with a 90 percent confidence level in all cases and a 95 percent confidence level for three of the comparisons based on a one-tailed Student's t-test.

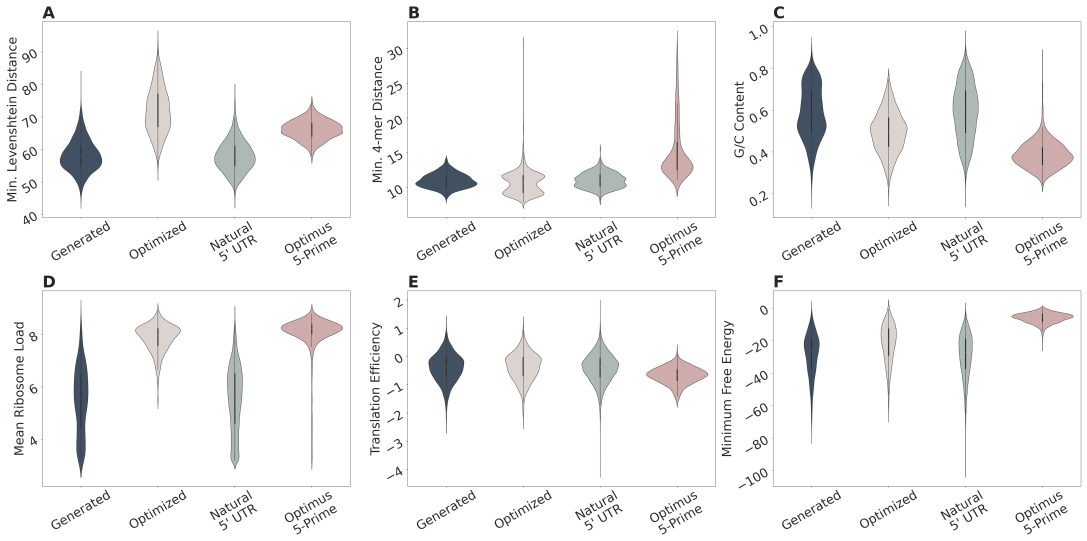

Fig. 3.1: **Comparison of UTRGAN-generated, natural, UTRGAN-optimized, and Optimus 5-Prime-optimized sequences.** **A**) The distributions of Levenshtein distances to the closest non-identical natural samples are shown for each group (UTRGAN-generated vs. natural: $p > 0.01$, effect size = 0.01, confidence interval = (0, 0); UTRGAN-optimized vs. natural: $p = 0.0$, effect size = 0.84, confidence interval = (8, 9); Optimus 5-Prime-optimized vs. natural: p = 0.0, effect size = 0.9, confidence interval = (13, 14)). **B**) The distributions of the 4-mer frequency distances to the closest non-identical natural sample are shown for each group (UTRGAN-generated vs. natural: p = 0.51, effect size = -0.02, confidence interval = (-0.18, 0.09); UTRGAN-optimized vs. natural: $p < 0.01$, effect size = -0.18, confidence interval = (-0.61, -0.31); Optimus 5-Prime-optimized vs. natural: $p < 0.01$, effect size = 0.89, confidence interval = (2.53, 2.92)). **C**) The GC content distributions are shown for each group (UTRGAN-generated vs. natural: p = 0.12, effect size = -0.02, confidence interval = (-0.01, 0.008); UTRGAN-optimized vs. natural: $p < 0.01$, effect size = -0.42, confidence interval = (-0.10, -0.09); Optimus 5-Prime-optimized vs. natural: p = 0.0, effect size = -0.81, confidence interval = (-0.21, -0.20)). **D**) The predicted MRL distributions are shown for each group (UTRGAN-generated vs. natural: p = 0.09, effect size = -0.03, confidence interval = (-0.14, -0.01); UTRGAN-optimized vs. natural: p = 0.0, effect size = 0.92, confidence interval = (2.14, 2.27); Optimus 5-Prime-optimized vs. natural: p = 0.0, effect size = 0.92, confidence interval = (2.37, 2.5)). **E**) The predicted translation efficiency distributions are shown for each group (UTRGAN-generated vs. natural: p = 0.92, effect size = 0.001, confidence interval = (-0.01, 0.01); UTRGAN-optimized vs. natural: p = 0.011, effect size = 0.04, confidence interval = (-0.61, -0.31); Optimus 5-Prime-optimized vs. natural: $p < 0.01$, effect size = -0.37, confidence interval = (0.006, 0.053)). **F**) The distributions of predicted MFEs are shown for each group (UTRGAN-generated vs. natural: p = 0.42, effect size = 0.01, confidence interval = (-0.5, 1.1); UTRGAN-optimized vs. natural: $p < 0.01$, effect size = 0.32, confidence interval = (6.29, 7.8); Optimus 5-Prime-optimized vs. natural: p = 0.0, effect size = 0.96, confidence interval = (20.69, 21.99)).

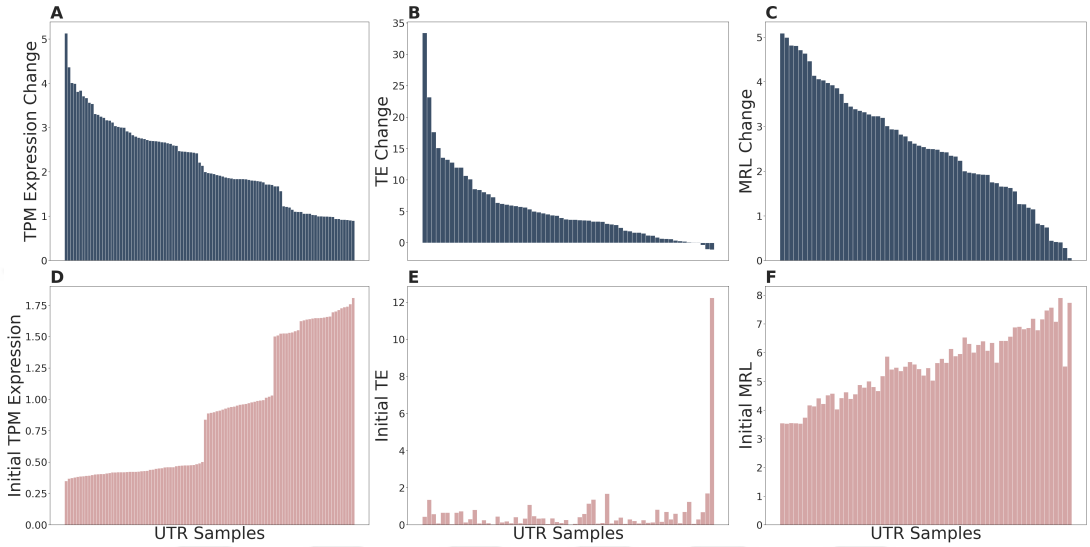

Fig. 3.2: **Overall performance of gene expression, TE, and MRL optimization A)** The expression change after 3,000 iterations of optimization for 8 DNA samples shows that the model successfully generates 5' UTR sequences and optimizes a majority of those to improve the original expression. The values are TPM Expression values, and the optimization yields higher predicted gene expression in 80% of the DNAs. **B)** Here is shown the translation efficiency change of 64 optimized generated 5' UTR sequences. In this case, a 97% majority of the sequences yield substantially higher TE after optimization. **C)** Similar to gene expression and TE, optimization is effective for the majority of the sequences and performs better for sequences with low initial MRL values. **D, E, and F)** Here, we show the initial gene expression, TE, and MRL values for the corresponding samples in panels D, E, and F, respectively. For all three panels, initial values tend to increase towards the right. We see that it is more likely for the optimization to degrade performance for samples with high initial TE or gene expression values. Unlike gene expression and TE, degradation in the predicted MRL of the optimized 5' UTRs is rare.

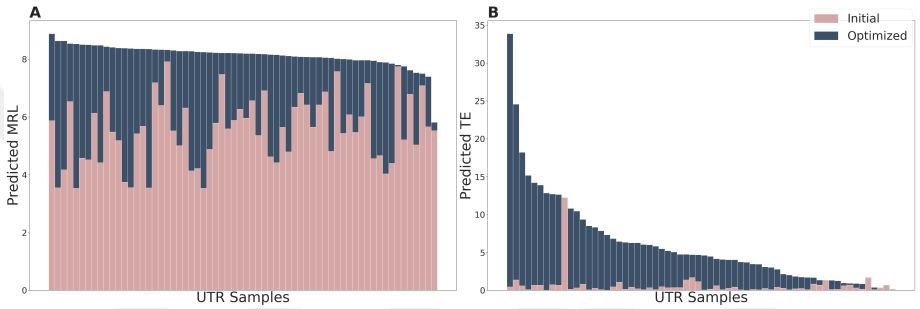

Fig. 3.3: **Overall performance of MRL and TE optimization. A)**. The optimization for higher MRL using FramePool results in sequences with a considerably high average MRL of 7.6. The optimization result may vary depending on the initial random sequences, and it optimizes almost all sequences to MRLs higher than 7. **B)**. Translation efficiency reaches a high value of 33 after optimization. The initial values are all around the average TE of the natural samples, and optimization increases the average more than 32-fold, and the highest optimized value is close to the maximum value of the natural sequences. The optimized values here are behind the initial values in all panels, and for the few sequences where the blue bar is not shown, the optimized value is smaller than the initial value.

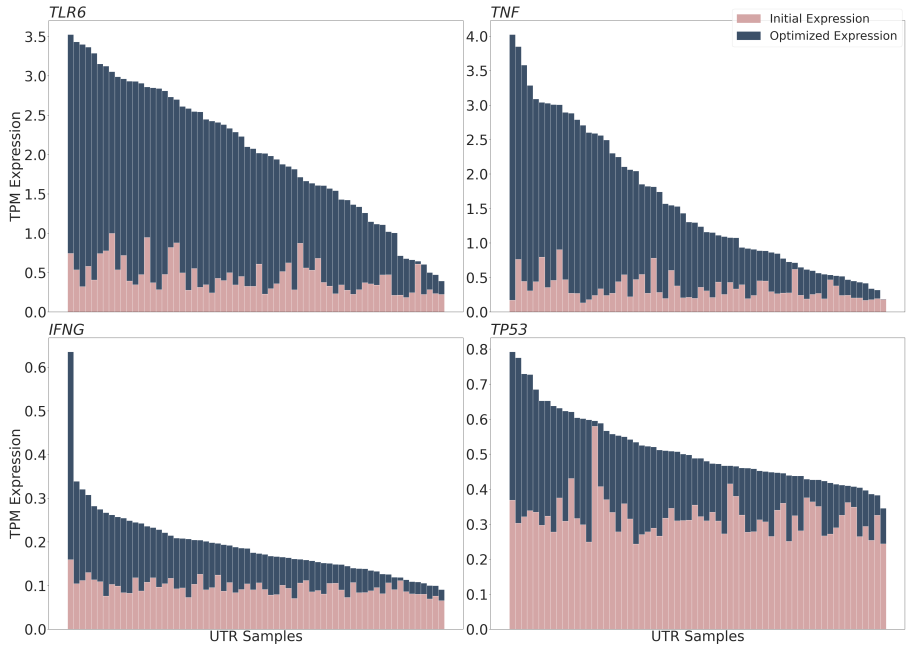

Fig. 3.4: **Expression optimization for specific target genes.** In addition to optimizing for random genes, it is possible to optimize many generated UTR sequences for specific genes. Here is shown the result of optimization for four specific genes. *TLR6*, *IFNG*, *TNF*, and *TP53* genes show a substantial increase in the expression for optimized sequences compared to initial predicted expression values. The optimization shows 4-fold, 37%, 63%, and 4.2-fold increase in expression on average for *TLR6*, *INFG*, *TNF*, and *TP53*, respectively. The optimized expression values are behind the initial values in all panels, and for the few sequences where the blue bar is not shown, the optimized value is smaller than the initial value.

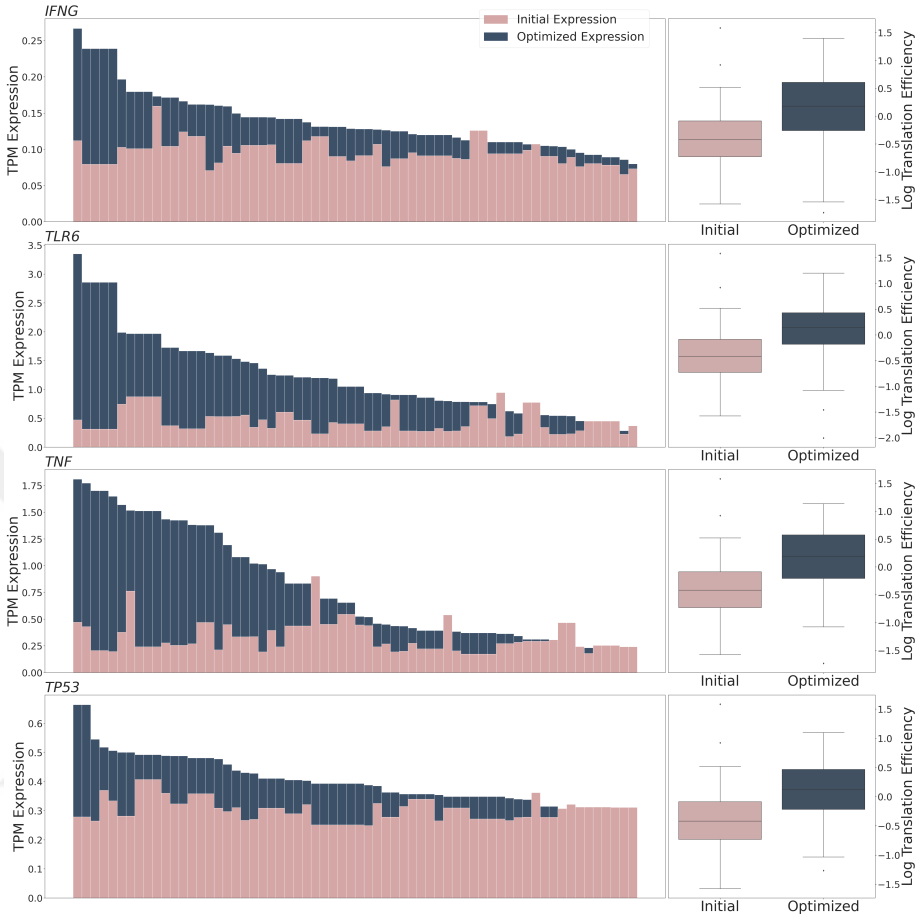

Fig. 3.5: **Jointly optimizing for TE and mRNA expression** The generated sequences are first optimized for high TE for 1,000 iterations, and then the sequences resulting from the optimization are used as the starting latent vector for gene expression optimization for another 1,000 iterations. As seen for the four genes, *TLR6*, *IFNG*, *TNF*, and *TP53*, the two-step optimization results in higher gene expression and translation efficiency. The predicted gene expression increases up to 12-fold, 1.8-fold, 10-fold, and 2.8-fold with respect to the natural 5' UTRs of the genes, respectively. The maximum predicted TE after optimization is at least 77% higher than the natural 5' UTR of the target genes. The optimized expression values are behind the initial values in all panels, and for the few sequences where the blue bar is not shown, the optimized value is smaller than the initial value. The box plot characterizes the TE values using the 25th, 50th, and 75th, also known as quartiles (Q1, median, Q3) and the interquartile range (IQR = Q3 - Q1), with whiskers extending to a maximum of 1.5 times the IQR. Outliers beyond the whiskers are plotted separately.

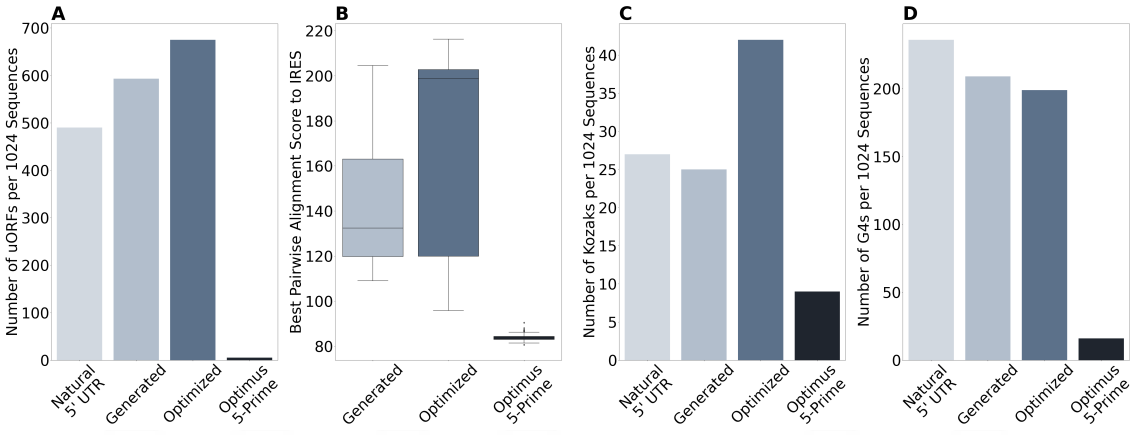

Fig. 3.6: **Regulatory element analysis of UTRGAN-generated, UTRGAN-optimized, Optimus 5-Prime-optimized and natural 5' UTR sequences.** The elements analyzed here include uORF, IRES, Alternative Kozak, and G4 sequences, which are shown on panels **A**,**B**,**C**, and **D**, respectively. We count the occurrence of each regulatory element in natural 5' UTR, generated, UTRGAN-optimized, and Optimus 5-Prime-optimized sequence sets. Note that all sets except the natural 5' UTR set have 1,024 sequences. So, we normalize the latter to show a number of hits for 1,024 sequences on average.

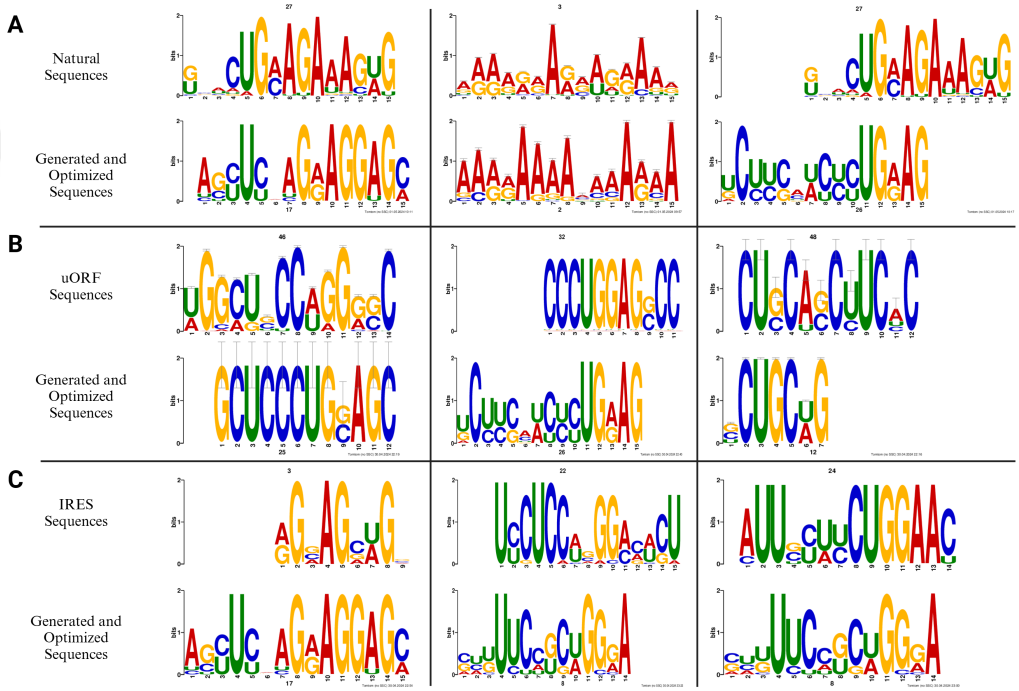

Fig. 3.7: **Top Matching Motives between UTRGAN-optimized and various natural sequences.** The top-3 motif matches between the sequences optimized using UTRGAN and the natural 5' UTR, uORF, and IRES sequences are shown on panels **A**,**B**, and **C**, respectively. The leftmost motif is the top match for each category.

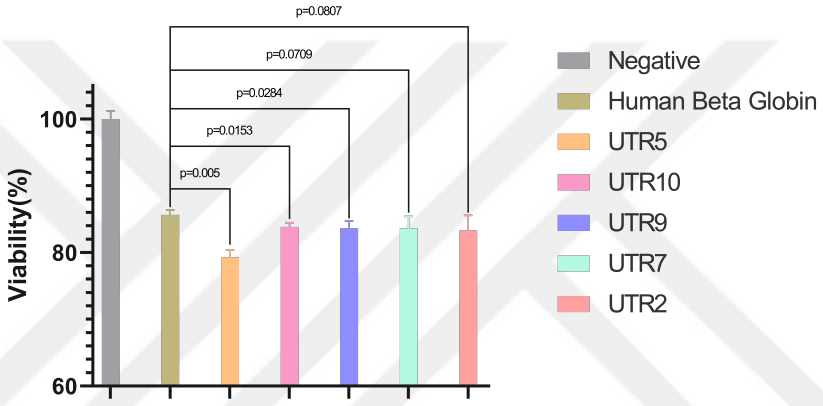

Fig. 3.8: **Cytotoxicity level of TNF-α with natural and UTRGAN-generated and then UTRGAN-optimized 5' UTRs.** Shown here are the effects of TNF-α proteins in transfecting MCF7 cells. These values show the level of translation using different 5' UTRs, including the human β-globin and 5 generated/optimized UTRs. The designed UTRs are more effective than the human β-globin UTR in upregulating the translation of the TNF-α mRNA.

Chapter 4

Discussion

We present UTRGAN, a pioneering framework for creating synthetic 5' UTR sequences tailored to a specific target feature. Our method allows for the production of sequences with characteristics akin to Human 5' UTR samples, circumventing the need to explore the vast and complex search space of DNA sequences up to 128 nucleotides in length. This represents a significant advancement over current state-of-the-art methods, which focus on optimizing existing UTRs through modifications or generating sequences from scratch. Additionally, our model is capable of generating sequences of varying lengths, thereby expanding the potential range of generated sequences.

UTRGAN is a powerful framework for engineering synthetic 5' UTR sequences, suitable for various applications in genetic circuit design. The model includes a GAN trained on a dataset of natural Human 5' UTRs and state-of-the-art prediction models for MRL, TE, and mRNA abundance [32] [33] [31]. Our model has the potential to optimize sequences for any given objective function, provided there is a differentiable prediction model to forecast the target metric for a 5' UTR sequence. It can

maximize or minimize the target feature. Additionally, by replacing the prediction model with a classifier, the framework can generate sequences belonging to a specific class.

Although UTRGAN can generate and optimize sequences with desirable characteristics, the optimization process can sometimes lead to decreased TE or mRNA levels. This occurs in less than 5% of cases for TE in the worst optimization scenarios and in 20% of cases for mRNA abundance when optimizing UTRs for multiple genes, while it rarely happens during MRL optimization (Figure 3.2 C). A possible reason for decreased expression is the limitation of the predictor model (Xpresso). Xpresso operates on the entire DNA (gene) sequence, and when back-propagating the gradient to update the sequence, only the gradient corresponding to the 5' UTR can be utilized. As a result, changing the 5' UTR without altering the rest of the sequence might cause the model to reach a suboptimal point in the loss space, and modifying the rest of the DNA is not desirable for the target application, which requires the rest of the DNA to remain unchanged. Additionally, in both mRNA expression and TE optimization, decreased scores tend to occur when the optimization begins with a sequence that initially yields a high score, as illustrated in Figure 3.2D and 3.2E. In such cases, users can opt to use the starting sequence and discard the optimization results.

We anticipate widespread adoption of data-driven sequences generated through deep learning methods in the near future. Creating 5' UTRs with this approach is practical and does not demand extensive computational resources. Although our focus is on 5' UTRs, the framework can easily be extended to generate 3' UTRs or other regulatory elements. Additionally, other generative models, such as variational autoencoders or diffusion models, can be employed. Generating sequences with regulatory functions allows researchers to conduct experiments more quickly, which is particularly advantageous for mRNA vaccine production and mRNA-based therapeutics.

Bibliography

[1] J. B. Ulmer and A. J. Geall, "Recent innovations in mrna vaccines," *Current opinion in immunology*, vol. 41, pp. 18–22, 2016.

[2] E. Rohner, R. Yang, K. S. Foo, A. Goedel, and K. R. Chien, "Unlocking the promise of mrna therapeutics," *Nature biotechnology*, vol. 40, no. 11, pp. 1586–1600, 2022.

[3] C.-Y. Chen, D. M. Tran, A. Cavedon, X. Cai, R. Rajendran, M. J. Lyle, P. G. Martini, and C. H. Miao, "Treatment of hemophilia a using factor viii messenger rna lipid nanoparticles," *Molecular Therapy-Nucleic Acids*, vol. 20, pp. 534–544, 2020.

[4] E. Robinson, K. D. MacDonald, K. Slaughter, M. McKinney, S. Patel, C. Sun, and G. Sahay, "Lipid nanoparticle-delivered chemically modified mrna restores chloride secretion in cystic fibrosis," *Molecular Therapy*, vol. 26, no. 8, pp. 2034–2046, 2018.

[5] J. D. Gillmore, E. Gane, J. Taubel, J. Kao, M. Fontana, M. L. Maitland, J. Seitzer, D. O'Connell, K. R. Walsh, K. Wood, *et al.*, "Crispr-cas9 in vivo gene editing for transthyretin amyloidosis," *New England Journal of Medicine*, vol. 385, no. 6, pp. 493–502, 2021.

[6] N. Pardi, M. J. Hogan, R. S. Pelc, H. Muramatsu, H. Andersen, C. R. DeMaso, K. A. Dowd, L. L. Sutherland, R. M. Scearce, R. Parks, *et al.*, "Zika virus protection by a single low-dose nucleoside-modified mrna vaccination," *Nature*, vol. 543, no. 7644, pp. 248–251, 2017.

[7] A. B. Vogel, I. Kanevsky, Y. Che, K. A. Swanson, A. Muik, M. Vormehr, L. M. Kranz, K. C. Walzer, S. Hein, A. Güler, *et al.*, "Bnt162b vaccines protect rhesus macaques from sars-cov-2," *Nature*, vol. 592, no. 7853, pp. 283–289, 2021.

[8] M. S. Gebre, S. Rauch, N. Roth, J. Yu, A. Chandrashekar, N. B. Mercado, X. He, J. Liu, K. McMahan, A. Martinot, *et al.*, "Optimization of non-coding regions for a non-modified mrna covid-19 vaccine," *Nature*, vol. 601, no. 7893, pp. 410–414, 2022.

[9] R. A. Feldman, R. Fuhr, I. Smolenov, A. M. Ribeiro, L. Panther, M. Watson, J. J. Senn, M. Smith, Almarsson, H. S. Pujar, *et al.*, "mrna vaccines against h10n8 and h7n9 influenza viruses of pandemic potential are immunogenic and well tolerated in healthy adults in phase 1 randomized clinical trials," *Vaccine*, vol. 37, no. 25, pp. 3326–3334, 2019.

[10] U. Sahin, P. Oehm, E. Derhovanessian, R. A. Jabulowsky, M. Vormehr, M. Gold, D. Maurus, D. Schwarck-Kokarakis, A. N. Kuhn, T. Omokoko, *et al.*, "An rna vaccine drives immunity in checkpoint-inhibitor-treated melanoma," *Nature*, vol. 585, no. 7823, pp. 107–112, 2020.

[11] M. Andronescu, A. P. Fejes, F. Hutter, H. H. Hoos, and A. Condon, "A new algorithm for rna secondary structure design," *Journal of molecular biology*, vol. 336, no. 3, pp. 607–624, 2004.

[12] A. Busch and R. Backofen, "Info-rna—a server for fast inverse rna folding satisfying sequence constraints," *Nucleic acids research*, vol. 35, no. suppl_2,

pp. W310–W313, 2007.

[13] S. Sav, D. J. Hampson, and H. H. Tsang, "Simard: A simulated annealing based rna design algorithm with quality pre-selection strategies," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, IEEE, 2016.

[14] Y. Liang, L. Huang, and T. Liu, "Development and delivery systems of mrna vaccines," *Frontiers in Bioengineering and Biotechnology*, vol. 9, 2021.

[15] N. Sultana, Y. Hadas, M. T. K. Sharkar, K. Kaur, A. Magadum, A. A. Kurian, N. Hossain, B. Alburquerque, S. Ahmed, E. Chepurko, *et al.*, "Optimization of 5′ untranslated region of modified mrna for use in cardiac or hepatic ischemic injury," *Molecular Therapy-Methods & Clinical Development*, vol. 17, pp. 622–633, 2020.

[16] T. K.-T. Lu, M. Kellis, J. Cao, E. M. N. Pardo, and Z. Zhang, "Synthetic 5'utr sequences, and high-throughput engineering and screening thereof," Feb. 27 2020. US Patent App. 16/441,647.

[17] A. O. Von Niessen, S. Fesser, B. Vallazza, T. Beissert, A. Kuhn, U. Sahin, and M. A. Poleganov, "3′-utr sequences for stabilization of rna," Nov. 8 2022. US Patent 11,492,628.

[18] A. G. O. von Niessen, M. A. Poleganov, C. Rechner, A. Plaschke, L. M. Kranz, S. Fesser, M. Diken, M. Löwer, B. Vallazza, T. Beissert, *et al.*, "Improving mrna-based therapeutic gene delivery by expression-augmenting 3′ utrs identified by cellular library screening," *Molecular Therapy*, vol. 27, no. 4, pp. 824–836, 2019.

[19] A. O. Von Niessen, S. Fesser, B. Vallazza, T. Beissert, A. Kuhn, U. Sahin, and M. A. Poleganov, "3'-utr sequences for stabilization of rna," Mar. 7 2019. US Patent App. 15/763,709.

[20] J. Cao, E. M. Novoa, Z. Zhang, W. C. Chen, D. Liu, G. C. Choi, A. S. Wong, C. Wehrspaun, M. Kellis, and T. K. Lu, "High-throughput 5′ utr engineering for enhanced protein production in non-viral gene therapies," *Nature communications*, vol. 12, no. 1, p. 4138, 2021.

[21] D. H. Burkhardt, R. R. Subramanian, and C. Cobaugh, "Heterologous utr sequences for enhanced mrna expression," July 19 2022. US Patent 11,389,546.

[22] F. Ozden, S. Barazandeh, D. Akboga, U. O. S. Seker, and A. E. Cicek, "Rnagen: A generative adversarial network-based model to generate synthetic rna sequences to target proteins," *bioRxiv*, pp. 2023–07, 2023.

[23] J. Zrimec, F. Buric, M. Kokina, V. Garcia, and A. Zelezniak, "Learning the regulatory code of gene expression," *Frontiers in Molecular Biosciences*, vol. 8, 2021.

[24] J. Zrimec, X. Fu, A. S. Muhammad, C. Skrekas, V. Jauniskis, N. K. Speicher, C. S. Börlin, V. Verendel, M. H. Chehreghani, D. Dubhashi, *et al.*, "Controlling gene expression with deep generative design of regulatory dna," *Nature communications*, vol. 13, no. 1, p. 5099, 2022.

[25] J. Zrimec, C. S. Börlin, F. Buric, A. S. Muhammad, R. Chen, V. Siewers, V. Verendel, J. Nielsen, M. Töpel, and A. Zelezniak, "Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure," *Nature communications*, vol. 11, no. 1, p. 6141, 2020.

[26] F. Cunningham, J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. Amode, I. Armean, O. Austine-Orimoloye, A. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui,

T. El Naboulsi, R. Fatima, C. G. Giron, T. Genez, J. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. C. Marugán, S. Mohanan, A. Mushtaq, M. Naven, D. Ogeh, A. Parker, A. Parton, M. Perry, I. Piližota, I. Prosovetskaia, M. Sakthivel, A. Salam, B. Schmitt, H. Schuilenburg, D. Sheppard, J. Pérez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M.-M. Suner, M. Szpak, A. Thormann, F. F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S. Hunt, G. IIsley, J. Loveland, F. Martin, B. Moore, J. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. Trevanion, S. Dyer, P. Harrison, K. Howe, A. Yates, D. Zerbino, and P. Flicek, "Ensembl 2022," *Nucleic Acids Research*, vol. 50, pp. D988–D995, 11 2021.

[27] S. M. Castillo-Hair and G. Seelig, "Machine learning for designing next-generation mrna therapeutics," *Accounts of Chemical Research*, vol. 55, no. 1, pp. 24–34, 2021.

[28] E. Trotta, "On the normalization of the minimum free energy of rnas by sequence length," *PloS one*, vol. 9, no. 11, p. e113380, 2014.

[29] C. Lo Giudice, F. Zambelli, M. Chiara, G. Pavesi, M. A. Tangaro, E. Picardi, and G. Pesole, "Utrdb 2.0: a comprehensive, expert curated catalog of eukaryotic mrnas untranslated regions," *Nucleic Acids Research*, vol. 51, no. D1, pp. D337–D344, 2023.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[31] V. Agarwal and J. Shendure, "Predicting mrna abundance directly from genomic

sequence using deep convolutional neural networks," *Cell reports*, vol. 31, no. 7, p. 107663, 2020.

[32] A. Karollus, Ž. Avsec, and J. Gagneur, "Predicting mean ribosome load for 5'utr of any length using deep learning," *PLoS computational biology*, vol. 17, no. 5, p. e1008982, 2021.

[33] W. Zheng, J. H. Fong, Y. K. Wan, A. H. Chu, Y. Huang, A. S. Wong, and J. W. Ho, "Translation rate prediction and regulatory motif discovery with multi-task learning," in *International Conference on Research in Computational Molecular Biology*, pp. 139–154, Springer, 2023.

[34] P. Gage, "A new algorithm for data compression," *The C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.

[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[36] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[37] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a wasserstein loss," *Advances in neural information processing systems*, vol. 28, 2015.

[38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[39] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, pp. 707–710, Soviet Union, 1966.

[40] L. V. Vinh, T. V. Lang, L. T. Binh, and T. V. Hoai, "A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads," *Algorithms for Molecular Biology*, vol. 10, no. 1, pp. 1–12, 2015.

[41] Ö. Konu and M. D. Li, "Correlations between mrna expression levels and gc contents of coding and untranslated regions of genes in rodents," *Journal of Molecular Evolution*, vol. 54, no. 1, pp. 35–41, 2002.

[42] M. E. Fornace, J. Huang, C. T. Newman, N. J. Porubsky, M. B. Pierce, and N. A. Pierce, "Nupack: Analysis and design of nucleic acid structures, devices, and systems," 2022.

[43] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Viennarna package 2.0," *Algorithms for molecular biology*, vol. 6, pp. 1–14, 2011.

[44] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, P. Schuster, *et al.*, "Fast folding and comparison of rna secondary structures," *Monatshefte fur chemie*, vol. 125, pp. 167–167, 1994.

[45] P. J. Sample, B. Wang, D. W. Reid, V. Presnyak, I. J. McFadyen, D. R. Morris, and G. Seelig, "Human 5′ utr design and variant effect prediction from a massively parallel translation assay," *Nature biotechnology*, vol. 37, no. 7, pp. 803–809, 2019.

[46] M. Noreen and M. Arshad, "Association of tlr1, tlr2, tlr4, tlr6, and tirap polymorphisms with disease susceptibility," *Immunologic Research*, vol. 62, no. 2,

pp. 234–252, 2015.

[47] Y. Sun, B. Hu, G. Stanley, Z. M. Harris, S. Gautam, R. Homer, J. L. Koff, and G. Rajagopalan, "Interferon-γ is protective in cytokine release syndrome-associated extrapulmonary acute lung injury.," *American Journal of Respiratory Cell and Molecular Biology*, 2022.

[48] L. N. Benson, Y. Liu, K. Deck, C. Mora, and S. Mu, "Ifn-γ contributes to the immune mechanisms of hypertension," *Kidney360*, vol. 3, no. 12, pp. 2164–2173, 2022.

[49] Y.-J. Lee, H. W. Seo, J.-H. Baek, S. H. Lim, S.-G. Hwang, and E. H. Kim, "Gene expression profiling of glioblastoma cell lines depending on tp53 status after tumor-treating fields (ttfields) treatment," *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[50] T. Ozaki and A. Nakagawara, "Role of p53 in cell death and human cancers," *Cancers*, vol. 3, no. 1, pp. 994–1013, 2011.

[51] S. Arhondakis, O. Clay, and G. Bernardi, "Gc level and expression of human coding sequences," *Biochemical and Biophysical Research Communications*, vol. 367, no. 3, pp. 542–545, 2008.

[52] G. Kudla, L. Lipinski, F. Caffin, A. Helwak, and M. Zylicz, "High guanine and cytosine content increases mrna levels in mammalian cells," *PLoS biology*, vol. 4, no. 6, p. e180, 2006.

[53] S. van der Horst, T. Filipovska, J. Hanson, and S. Smeekens, "Metabolite control of translation by conserved peptide uorfs: the ribosome as a metabolite multisensor," *Plant physiology*, vol. 182, no. 1, pp. 110–122, 2020.

[54] H.-H. Chen and W.-Y. Tarn, "uorf-mediated translational control: recently elucidated mechanisms and implications in cancer," *RNA biology*, vol. 16, no. 10, pp. 1327–1338, 2019.

[55] C. Barbosa, I. Peixeiro, and L. Romão, "Gene expression regulation by upstream open reading frames and human disease," *PLoS genetics*, vol. 9, no. 8, p. e1003529, 2013.

[56] F. Manske, L. Ogoniak, L. Jürgens, N. Grundmann, W. Makałowski, and K. Wethmar, "The new uorfdb: integrating literature, sequence, and variation data in a central hub for uorf research," *Nucleic Acids Research*, vol. 51, no. D1, pp. D328–D336, 2023.

[57] I. N. Shatsky, I. M. Terenin, V. V. Smirnova, and D. E. Andreev, "Cap-independent translation: what's in a name?," *Trends in biochemical sciences*, vol. 43, no. 11, pp. 882–895, 2018.

[58] H. A. King, L. C. Cobbold, and A. E. Willis, "The role of ires trans-acting factors in regulating translation initiation," 2010.

[59] J. Zhao, Y. Li, C. Wang, H. Zhang, H. Zhang, B. Jiang, X. Guo, and X. Song, "Iresbase: a comprehensive database of experimentally validated internal ribosome entry sites," *Genomics, Proteomics and Bioinformatics*, vol. 18, no. 2, pp. 129–139, 2020.

[60] T. Endoh and N. Sugimoto, "Mechanical insights into ribosomal progression overcoming rna g-quadruplex from periodical translation suppression in cells," *Scientific reports*, vol. 6, no. 1, p. 22719, 2016.

[61] J.-D. Beaudoin and J.-P. Perreault, "5′-utr g-quadruplex structures acting as translational repressors," *Nucleic acids research*, vol. 38, no. 20, pp. 7022–7036,

2010.

[62] H. B. Cagirici, H. Budak, and T. Z. Sen, "G4boost: a machine learning-based tool for quadruplex identification and stability prediction," *BMC bioinformatics*, vol. 23, no. 1, p. 240, 2022.

[63] J. Yang, Y. Cui, D. Yu, G. Zhang, R. Cao, Z. Gu, G. Dai, X. Wu, Y. Ling, C. Yi, *et al.*, "A noncoding a-to-u kozak site change related to the high transmissibility of alpha, delta, and omicron vocs," *Molecular Biology and Evolution*, vol. 40, no. 6, p. msad142, 2023.

[64] J. Xie, Z. Zhuang, S. Gou, Q. Zhang, X. Wang, T. Lan, M. Lian, N. Li, Y. Liang, Z. Ouyang, *et al.*, "Precise genome editing of the kozak sequence enables bidirectional and quantitative modulation of protein translation to anticipated levels without affecting transcription," *Nucleic Acids Research*, vol. 51, no. 18, pp. 10075–10093, 2023.

[65] C. Xu and J. Zhang, "Mammalian alternative translation initiation is mostly nonadaptive," *Molecular Biology and Evolution*, vol. 37, no. 7, pp. 2015–2028, 2020.

[66] S. Nakagawa, Y. Niimura, T. Gojobori, H. Tanaka, and K.-i. Miura, "Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes," *Nucleic acids research*, vol. 36, no. 3, pp. 861–871, 2008.

[67] T. L. Bailey, C. Elkan, *et al.*, "Fitting a mixture model by expectation maximization to discover motifs in bipolymers," 1994.

[68] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The meme suite," *Nucleic acids research*, vol. 43, no. W1, pp. W39–W49, 2015.

[69] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," *Genome biology*, vol. 8, pp. 1–9, 2007.

[70] E. O. Melo, O. P. de Melo Neto, and C. M. de Sá, "Adenosine-rich elements present in the 5′-untranslated region of pabp mrna can selectively reduce the abundance and translation of cat mrnas in vivo," *FEBS letters*, vol. 546, no. 2-3, pp. 329–334, 2003.

[71] R. Weber, U. Ghoshdastider, D. Spies, C. Duré, F. Valdivia-Francia, M. Forny, M. Ormiston, P. F. Renz, D. Taborsky, M. Yigit, *et al.*, "Monitoring the 5′ utr landscape reveals isoform switches to drive translational efficiencies in cancer," *Oncogene*, vol. 42, no. 9, pp. 638–650, 2023.

[72] A. M. Wang, A. A. Creasey, M. B. Ladner, L. S. Lin, J. Strickler, J. N. V. Arsdell, R. Yamamoto, and D. F. Mark, "Molecular cloning of the complementary dna for human tumor necrosis factor," *Science*, vol. 228, no. 4696, pp. 149–154, 1985.

[73] F. Pinci, M. M. Gaidt, C. Jung, G. Kuut, M. A. Jackson, S. Bauernfried, and V. Hornung, "C-tag tnf: A reporter system to study tnf shedding.," *Journal of Biological Chemistry*, vol. 295, no. 52, pp. 18065–18075, 2020.

[74] M. R. Green and J. Sambrook, "Precipitation of rna with ethanol," *Cold Spring Harbor Protocols*, 2020.

[75] Z. Wang, H. Kishimoto, P. Bhat-Nakshatri, C. Crean, and H. Nakshatri, "Tnf resistance in mcf-7 breast cancer cells is associated with altered subcellular localization of p21cip1 and p27kip1," *Cell Death Differentiation*, vol. 12, pp. 98–100, 2005.

[76] M. Ghandadi, J. Behravan, K. Abnous, M. E. Gharaee, and F. Mosaffa, "Tnf-α exerts cytotoxic effects on multidrug resistant breast cancer mcf-7/mx cells via

a non-apoptotic death pathway," *Cytokine*, vol. 97, pp. 167–174, 2017.

Appendix A

Suppelemntary Material

A.0.1 Supplementary Figures

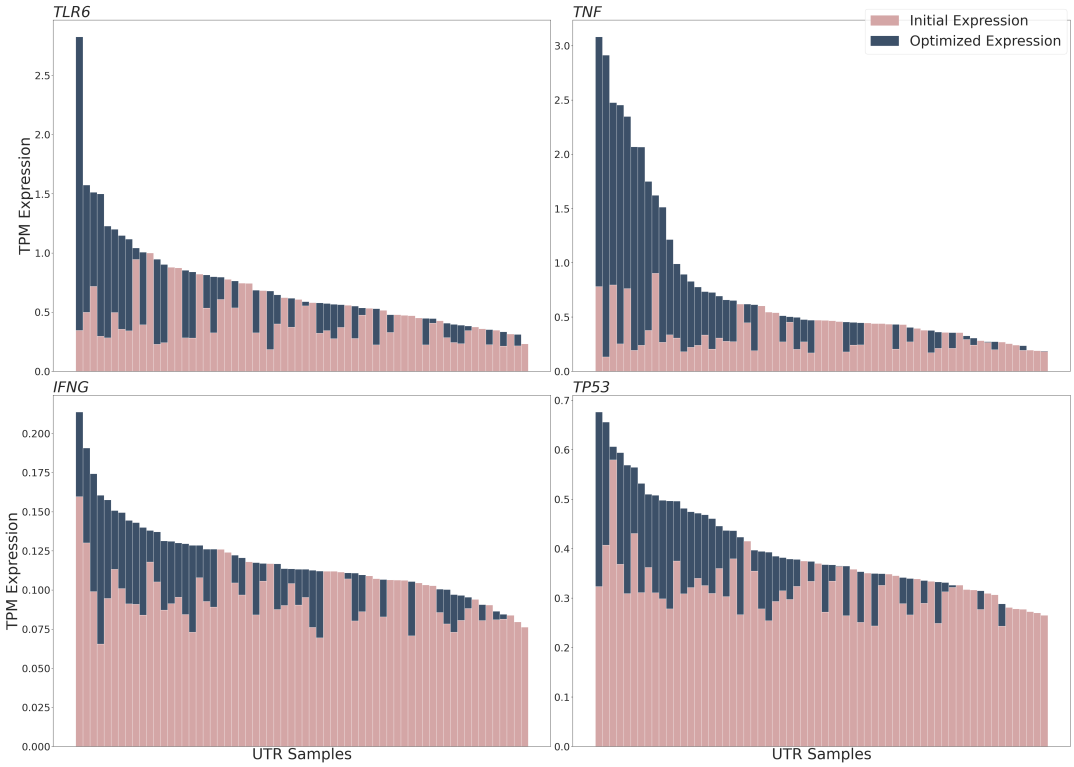

Fig. A.1: **Optimization with maximum GC content limitation.** The optimized sequences with a maximum GC content limitation of 65%. We performed the optimization and selected the best sequence with the condition of their GC content being below the selected value. Although this method does not perform as well as the uncontrolled optimization in increasing the gene expression, it does result in considerable increases in all four genes (*TLR6*, *IFNG*, *TNF*, *TP53*) while maintaining the GC content. The optimized expression values are behind the initial values in all panels, and for the few sequences where the blue bar is not shown, the optimized value is smaller than the initial value.

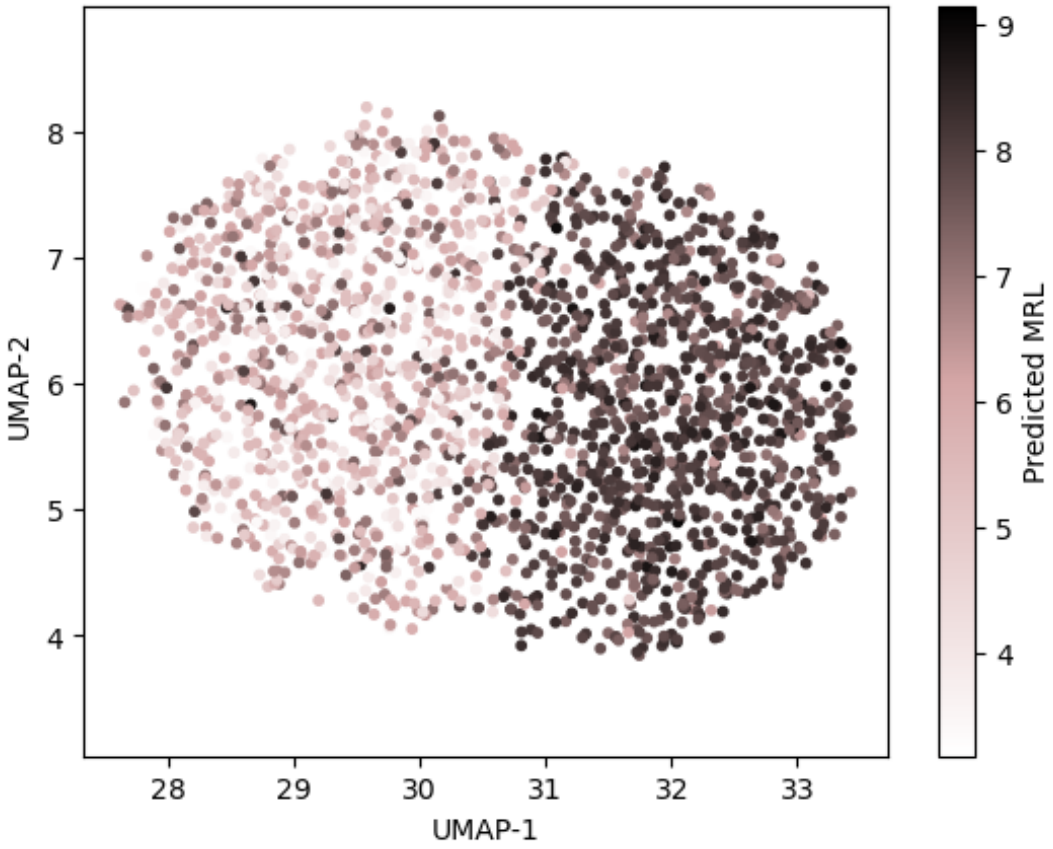

Fig. A.2: **The latent space of UTRGAN for generated and optimized UTRs.** Here, we show the latent space of the model for 5' UTRs generated and optimized using UTRGAN. The UMAP is performed on the latent vector of 2048 sequences, including 1024 initially generated sequences and 1024 sequences resulting from optimization of the first set of sequences. The reduced vectors are colored by their predicted MRL values. The plot shows that the model has learned a meaningful latent space, and sequences before and after optimization are well-separated in the latent space.

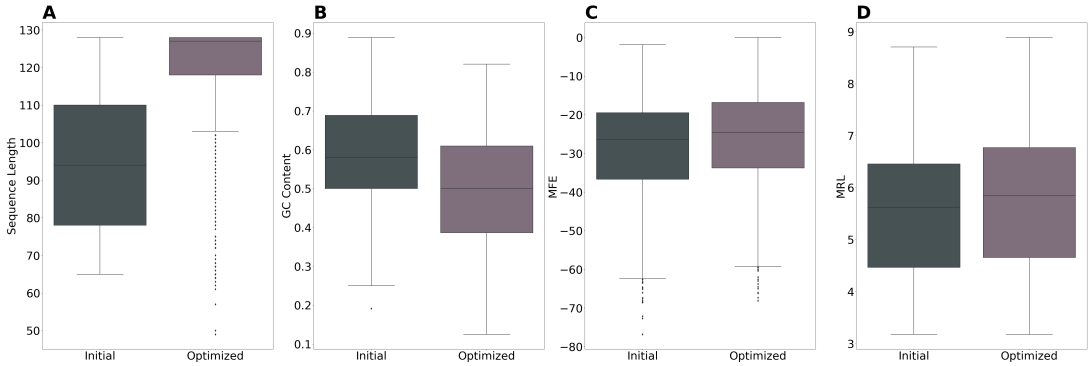

Fig. A.3: **Effect of TE optimization with respect to length, GC content, MFE, and MRL of the sequences.** The box plots characterize the values using the 25th, 50th, and 75th, also known as quartiles (Q1, median, Q3) and the interquartile range (IQR = Q3 - Q1), with whiskers extending to a maximum of 1.5 times the IQR. Outliers beyond the whiskers are plotted separately.

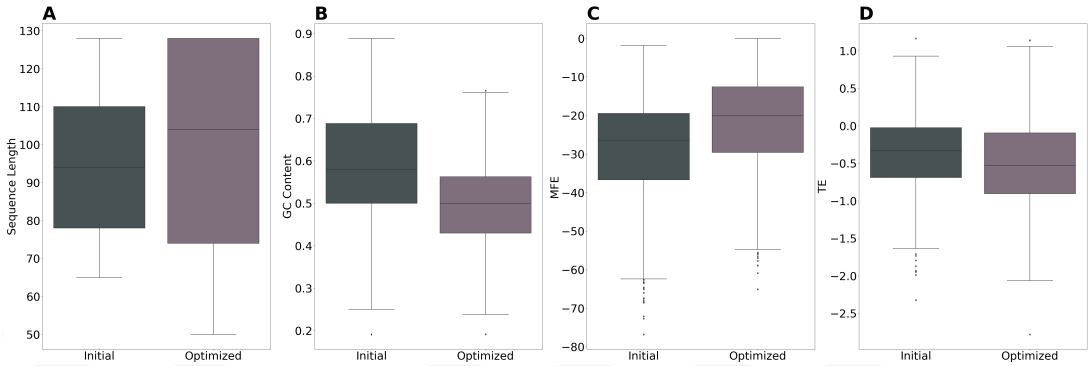

Fig. A.4: **Effect of MRL optimization with respect to length, GC content, MFE, and TE of the sequences.** The box plots characterize the values using the 25th, 50th, and 75th, also known as quartiles (Q1, median, Q3) and the interquartile range (IQR = Q3 - Q1), with whiskers extending to a maximum of 1.5 times the IQR. Outliers beyond the whiskers are plotted separately.

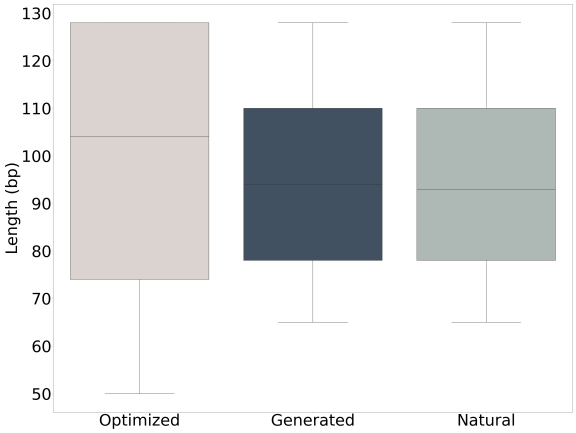

Fig. A.5: **Distribution of length of the generated, MRL-optimized, and natural 5' UTRs**

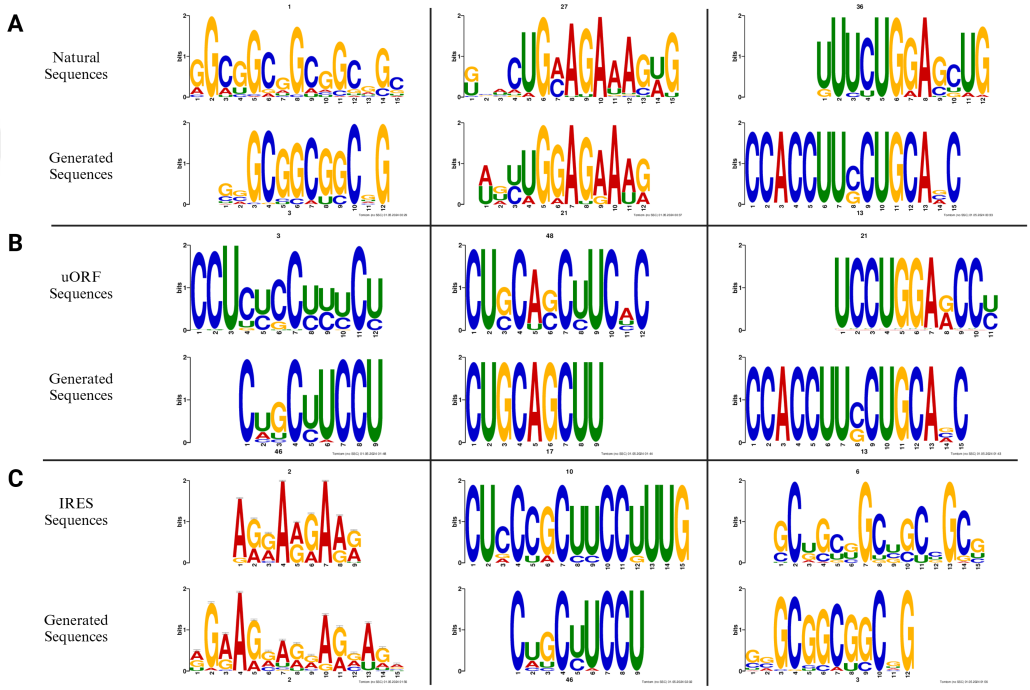

Fig. A.6: **Top Matching Motives between UTRGAN-generated and various natural sequences.** The top-3 motif matches between the sequences optimized using UTRGAN and the natural 5' UTR, uORF, and IRES sequences are shown on panels **A**,**B**, and **C**, respectively. The leftmost motif is the top match for each category.

A.1 Supplementary Notes

A.1.1 Experimental Setup

The model is implemented using the Tensorflow 2 library and in Python. We train and optimize the model on a Super-Micro Super-Server 4029GP-TRT with 2 Intel Xeon Gold 6140 Processors (2.3 GHz, 24.75 M cache), 256 GB RAM, 6 NVIDIA GeForce RTX 2080 Ti GPUs (11 GB, 352 bit), and 2 NVIDIA TITAN RTX GPUs (24 GB, 384 bit). We only use a single TITAN RTX GPU. The training takes around 14 hours for 3,000 epochs, and we select this checkpoint based on the validation loss as well as the p-values obtained from the statistical tests. The TE, MRL, and gene expression optimization take 3, 4, and 15 minutes, respectively. Optimizing UTR sequences for higher average expression in 8 genes takes 90 minutes.

A.1.2 Hyperparameter Optimization

The main hyperparameters of our model include the size of the latent dimension, the number of generator layers, and the number of discriminator layers. To find the best combination of parameters, we run a grid search on a reasonable range of parameters. The model is trained up to 200,000 generator iterations. Many of the instances of the model fail to learn and overfit, resulting in the validation loss going below the training loss, and the training loss begins to increase in some cases. In the cases that the mentioned scenario doesn't happen, we select the parameters where the p-values for the statistical tests are closer to desirable values. The latent dimension in UTRGAN is 40, and we present the number of layers in Supplementary Tables 6 and 7.

A.1.3 *In vitro* Experiment Setup

As discussed in the results, we compare the generated 5' UTRs and the human β-globin 5' UTR in the translation of the TNF-α protein. The methodology of the

conducted experiment is explained in detail in the sections below.

A.1.3.1 Assembling of the constructs

For the in vitro transcription (IVT) constructs, we utilized the pUC57-T7 promoter-Human -globin 5' UTR-sfGFP- Human -globin 3' UTR-118 poly(A) construct previously obtained from Genewiz for our laboratory. Five UTRs were selected randomly from a pool of top ten *de novo* UTRGAN-generated and UTRGAN-optimized UTRs based on mean ribosomal load (MRL). This pool of UTRs was designated as UTR1 to UTR10 (see Supplementary Table 4) to simplify labeling. The selected ones UTR2, UTR5, UTR7, UTR9, UTR10, and the human β-globin 5' UTR (used as a control) were cloned into the pUC57 vector as described earlier, which includes a coding region for human TNF$-\alpha$. The human TNF$-\alpha$ coding region was sourced from the pLI$-$TNF plasmid [73] (deposited to Addgene by Veit Hornung, Addgene #171179). The selected 5' UTR regions were incorporated into the coding region through polymerase chain reaction (PCR) and then ligated into the digested pUC57 vector instead of sfGFP. Following cloning, sequence verification of the constructs was conducted using Sanger sequencing (Genewiz). Throughout the cloning process, the DH5-α strain of *Escherichia coli* was employed.

A.1.3.2 *In vitro* transcription and removal of the inorganic contaminants

The cloned plasmids were isolated using GeneJET Plasmid midiprep kit (Thermo Scientific) from overnight-grown bacterial cultures. Linearization was achieved using a single restriction enzyme located at the end of the poly(A) tail. The linearized plasmids were purified using the Nucleic Acid Purification Kit (Monarch). To generate the IVT mRNAs, we employed the HiScribe® T7 ARCA mRNA kit (NEB). The resulting mRNAs were purified using the Monarch RNA cleanup kit (NEB). Following nanodrop analysis of the mRNAs, isopropanol precipitation [74] was performed

if any inorganic contaminants affecting the A230/260 ratio were detected. This involved mixing the mRNA samples with an appropriate amount of 3 M sodium acetate solution (pH 5.2) to achieve a final concentration of 0.3 M, followed by adding an equal volume of isopropanol. The mixture was then incubated at -20°C overnight. After incubation, the mRNAs were recovered by centrifugation at 14,000 g for 10 minutes. The resulting pellet was washed with 500 µl of ice-cold 70% ethanol and centrifuged under the same conditions. Any remaining ethanol was evaporated, and nuclease-free water was used to dissolve the mRNA samples. These samples were then stored at -80°C until the transfection experiments were conducted.

A.1.3.3 Cell line maintenance

We utilized the HEK293T and MCF7 mammalian cell lines for subsequent experiments. To prepare a complete growth medium for both cell lines, 440 ml of high glucose Dulbecco's Modified Eagle's Medium (DMEM) was mixed with 50 ml of heat-inactivated Fetal Bovine Serum (FBS), 5 ml of 100x L-Glutamine (200 mM), and 5 ml of 100x Penicillin/Streptomycin. This mixture was then filtered using a Corning Disposable Filter Unit with a pore size of 0.45 µm. The resulting medium was stored at +4°C. Cell media were changed every other day, and subculturing of both cell lines was performed when they reached 90% confluency.

A.1.3.4 mRNA transfection and sample collection

For the production of human TNF-α under different UTRs, mRNA transfection was performed on HEK293T cells. These cells were seeded into 6-well plates at a density of 300,000 cells per well, with each well containing 1 ml of growth medium. Following a 24-hour incubation period in a 37°C humidified incubator with 5% CO2, the media were aspirated, and 1 ml of fresh medium was added to each well. Subsequently, 5 µg of IVT mRNAs containing the 5' UTRs (including human β-globin, UTR2, UTR5, UTR7, UTR9, and UTR10) were transfected into the cells using Lipofectamine 3000 (Invitrogen). One well served as a negative control without mRNA to assess the

impact of the transfection reagent. The transfected cells were then returned to the incubator for an additional 24 hours. After the incubation, 1.2 ml of supernatant from each well was collected into ice-chilled tubes and kept on ice. To eliminate cellular debris, all samples were centrifuged at 5800 rpm for 5 minutes. The supernatants were concentrated to 400 µl using 0.5 ml 3 kDa centrifugal filters (Amicon) and stored at +4°C for subsequent cytotoxicity experiments, which were planned to be carried out on the same day to prevent degradation.

A.1.3.5 TNF-α cytotoxicity assay

The cytotoxic effect of TNF-α on cancer cells [75] was evaluated using the previously described methodology [76]. MCF7 cells were seeded into a 96-well plate at a density of 6,000 cells per well, with each well initially containing 100 µl of medium. After a 24-hour incubation period in a 37°C humidified incubator with 5% CO2, the old media were removed, and 65 µl of fresh medium was added to each well. This fresh medium was then combined with 35 µl of the concentrated TNF-α samples obtained from mRNA transfection. Top of Form All samples, including the negative control and those with different 5' UTRs (such as human β-globin, UTR2, UTR5, UTR7, UTR9, and UTR10), were studied in triplicate. After 48 hours of incubation, the supernatants were removed from the wells, and 90 µl of fresh medium was added to each well. Subsequently, 10 µl of a 5 mg/ml MTT solution was added to the wells, and the plate was incubated for 4 hours at 37°C. Following this incubation, the supernatants were discarded, and 100 µl of DMSO was added to each well to dissolve the formazan crystals. The plate, wrapped in aluminum foil, was then shaken for 15 minutes to ensure complete dissolution of the crystals. The absorbance at 570 nm was measured for each well, and the viability of each treatment was calculated by comparing it with the negative control sample.

A.2 Supplementary Tables

Tab. A.1: The top generated 5' UTR sequences for high gene expression in the target genes, that are *TLR6*, *IFNG*, *TNF*, and *TP53*. The table shows the top 5' UTR for each gene. These sequences result in higher predicted expression than the other generated ones when attached to these genes as 5' UTRs.

Gene	5' UTR
TLR6	CGGUGUCUGGAAGCCUUCUUUCGCGGUGCUGGACUCCGGGCC CGAGGCCCCGAAUAGUCGCGGAAGUUCUGGGACGAAUCCGG GGCCCGGUCGAACCGGCGGGGGCGCCGGCCGGCCGUCCAG
IFNG	AGACGUGGACCGCUUCGCGGCGGCUGUCGGGUGGAGUGGGAG GCGAGCCGGAAGUGGCGGUCCUGGCUGCUCCGGGUCGUCCCUG CCGGCGUCGGAGCGUCCCUCGGAACCAGGAGUGCCGGCGACG
TNF	GAGUUGCGGCCGGGCGGCGUCGCGGCGGUCGGGGCCGAGAGUC GUGGCCGGAAGCCGCCGUUGGAUGGGGCCGGGUCGGUUGG AGACGCCGGCGGCCCCGCCGCCCCCTTTGGUGGGCCCGA
TP53	CGGCUUGCUGCCGCCCGCGACAGCGGAGACAAGCAGGGAGGUC CUGCCGCCUCCCAGAAAUAUUGGGGGGGCUGGGGAAGCCGCAG GCCGACUGAGUCGGGAGGCGGACGGCGGAAGGACGCGGCGGA

Tab. A.2: The top generated 5' UTR sequences for joint optimization of translation efficiency and high gene expression in the target genes, that are *TLR6*, *IFNG*, *TNF*, and *TP53*. The table shows the top 5' UTR for each gene. These sequences result in higher predicted expression as well as higher predicted translation efficiency than the other generated ones when attached to these genes as 5' UTRs.

Gene	5' UTR
TLR6	UGGCUGUGGCGGCCGGAAAUCUGAGCCUCCGGGCCAGUGGGC CGGGAGCGGUCGGGUCUGGAGCGGCGGAGGGAGAACGGAAGCU GCGGAGGCCCCGCACGCGGGGGGCUGAGGUCGGCCUCGUCAGU
IFNG	UGGCUGUGGCGGCAGGAAUUCUGUGCCUCCGGGCCUGUGGGU CGGGAGCGGCCGGGUCUGGAGCGGCGGUGGGAGAAAGGAAGCG GCGGUGGCCCCGCAAGCGGGGGGCUGAGGGCGGCCGGGGCGG
TNF	AGUCCGCUGGCGUGGCCCAGCCCCGGGGCCGGUUCUCGGGAA CGAGAAGCUGGUGAUCUCCGUUGGGGCGCGGUCGGAUGCGGUU GAGGGAGGCGAGGACGCGAAAGCGGGCCCGCAGCGGAGCCGGC
TP53	GGGAGGUGGUUCCGAAAUCGCCGUGAAGCCAAGCCCUCUGCU CCGUUUCCUCUUUUCCACGGACUGGGAGGUUUUCGGGACCACG CCAGGAGAACCACGAGGGCUAGGCAGAGGACUGAAGCGUCGGA

Tab. A.3: The top generated 5' UTR sequences for GC content controlled high gene expression in the target genes, that are *TLR6*, *IFNG*, *TNF*, and *TP53*. The table shows the top 5' UTR for each gene. These sequences result in higher predicted expression compared to other generated 5' UTRs with GC content lower than 65%.

Gene	5' UTR
TLR6	GUGAAAGGCGUGCCGCCGCCCUAGAGUCCUGGGAAGACAGUC AAGCGGAGAGAGGGAAGUUUCCCUGUACUUCCUCCCUCCCGGC CGCGAGUCCUGGCAUCCCGCUGAGGCUGG
IFNG	UGGCUGUGGCGGCAGGAAUUCUGUGCCUCCGGGCCUGUGGGU CGGGAGCGGCCGGGUCUGGAGCGGCGGUGGGAGAAAGGAAGCG GCGGUGGCCCCGCAAGCGGGGGGCUGAGGGCGGCCGGGGCGG
TNF	CUUCGCCCUUCUUCAUAAUCAGUCGACAAGCGGUUUCUGGAA GGCGGCGUCUCGCUGUUGCUGUGGCCUCUGGAAGCAGGAGAAG AGGCGGCUUGGAAACGCUGGAAGCGGCCGCGGCGUCCUCAG
TP53	AAGACGAGGCCAGGCUGCGCAAAUCCGGCUGGGGCCGAGUUC UUGGUCCGUUUUCCCAAUAAGAGUGGUCCCGGGAAUGGACGAG ACGCCGCCGGAAGCGUCUCUGGAGGUUGAUGGCAAG

Tab. A.4: The top 10 generated 5' UTR sequences for optimized MRL.

Sequence	Predicted MRL	UTR
GAGGAGCAGAAAUUGGCCGCUGUUGCGUUAAUC CACAACGGUGCUGUUGUGAGUUGACAGCAGGU CCAGGGAUAAGGUGUGGUGAAGUUGGUGUCUGU AGCAUAUCCAUGGAGAUGAUUGCAAUAGUA	9.06	UTR1
UUUAACACAUAUCAGAAAAAGAAUCAAGAACUG UUGGAGCCGGGUUAAAAAAAACUUUCCAAGUA AUGUUAAAG	9.02	UTR2
CUUUCCUGAGAAUUCCCUGGAAGCACUUGUUCA ACCACAGCACUAGGUUUCAAAGCAAGGUUCAC ACCCAGCCUCCACAAGGCCUGGGGAUGAUGAAG AGGGGAAUA	8.91	UTR3
GUUUUAUCCAGCAUCCAUUCGAGAACUUGAUCA AGUCCCACAUGCUACCCAGGACUAGGAGAGAA GGAAGCAGAAAGACAGUAAAUUGCCGCGGAUAU UUUCAG	8.81	UTR4
GUUACCAUACUGUAAUCAGAAUUCUGCUUCUAA GGAUAAAUUUACUUCUACUUAUUAUAAAAAAAG	8.80	UTR5
UGGACUCAGUCCCAACCAAAUAGUGGUGUUUCU GCAGCUCCAAACUUGAGAGAAGAACUUCAAG CUCUCCUUUAAACCGGUGUUUGC	8.74	UTR6

GUGUUUCCCAUUUAGUCACUAAGAGGUGACGCCA UACCACAUACAGGUGGAGGAGCAUUAAAGC	8.577	UTR7
GGGUGGUCGGAGCGGCUGCAGGAGGCAGUAGGA AGAAGAGAAAAGGAAGGAGGAGGAUCCGAA CACGUCACAAAACAGUCGG	8.575	UTR8
AGAAACCAGAAGACCUUCUGGAGGAAGACUUC CAUAGCUUUGUCUCAGUGAAGGCAGC	8.56	UTR9
AGAAAUUUCAGUAAUAAUCGUCCUGGAGGAGU GGGUAACAUAUUUGAAGAUUUCCAUUUCAGCUUU	8.49	UTR10

Tab. A.5: The top 3 generated 5' UTR sequences for optimized translation efficiency. The predicted *log*-TE for these sequences is 1.5, 1.3, and 1.2, respectively.

GAAUCUCAUCUUCGCCCUUAACCAUCUCAGAGACCUUC CAACAGAACUUGGUGUAAAAGCCAGGCUUCCCCG CCUCUGCAGUGCUGCGGCUUGGCUGCCUCUCCGCGCGAGG
GACCAAGCACACCUCACCCCGCGCCCCCAGUCCGGGCG CCGGGCUCCUCGGUGGUCUCAGCUGCU
ACCAUCCCUCCAGCAGCUGGGACGUGCCUGCGCCGCAG CCGUGGCCGCCUCGCUGCGGCUGUCUCCCCCCAG

Tab. A.6: The layers of the Generator in the WGAN model we use to generate realistic 5' UTR sequences. In this configuration, the batch size for training the model is 64, as seen in the table.

Input/Layer	Kernel size	Output shape
z_0	-	64 x 40
Dense	-	64 x 5120
Reshape	-	64 x 128 x 40
Convolution (1D)	1	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Convolution (1D)	1	64 x 128 x 5
Softmax	-	64 x 128 x 5

Tab. A.7: The layers of the Discriminator (Critic) in the WGAN model we use to generate realistic 5' UTR sequences.

Input/Layer	Kernel size	Output shape
Convolution (1D)	1	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Residual block (1D)	[5]x2	64 x 128 x 40
Flatten	-	64 x 5120
Dense	-	64 x 1