A generalized and efficient approach for complete mRNA design improves translation, stability and specificity

Aidan T. Riley[1,2], McKayla Vlasity[1,2], Joey Zhuoying Huang[1], Wyatt M. Becicka[1,2], Wilson W. Wong[1,2], Mark W. Grinstaff[1,4], and Alexander A. Green[1,2,3]*

[1]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

[2]Biological Design Center, Boston University, Boston, MA 02215, USA

[3]Molecular Biology, Cell Biology & Biochemistry Program, Graduate School of Arts and Sciences, Boston University, Boston, MA 02215, USA

[4]Department of Chemistry, Boston University, Boston, MA 02215, USA

*Corresponding author: aagreen@bu.edu

ABSTRACT

End-to-end, machine-learning based design of mRNA molecules offers a powerful means to tailor their properties for specific tasks. mRNA expression level, immunogenicity, tissue specificity, stability, and localization, among others strongly depend on sequence, thus providing a rich set of properties amenable to optimization. Despite this potential, the various components of mRNA are governed by distinct grammatical and functional rules that hinder the development of a unified algorithmic approach for complete mRNA design. While machine learning and generative AI techniques demonstrate substantial benefits for individual sequence design tasks, adapting these tools to new domains or out-of-distribution tasks remains challenging. In this work, we describe a simple and powerful alteration to integrated gradients (**D**esign by **I**ntegrated **G**radients- **DIGs**) that serves as the foundation for several mRNA design tasks. Using this technique, we demonstrate complete model-informed mRNA design and reveal the underexplored rules governing the assembly of mRNA components into complete, high-performance transcripts. This framework closes the loop on mRNA design by unifying the strengths of predictive neural networks with the vast quantities of diverse RNA data that has been accumulating for decades across transcriptomic profiling, CLIP-seq, and individual experimental findings. By linking the benefits of neural network-based design with diverse alternative datasets, we demonstrate the design of complete mRNA sequences in out-of-distribution settings that exhibit dramatically improved translational capacity, enhanced cell-type-specific expression, and improved stability.

INTRODUCTION

Sequence-to-function machine learning techniques have shown remarkable performance across a range of RNA design tasks[1-9]. From precise engineering of messenger RNA (mRNA) stability[8] and ribosome load[2] to the design of novel RNA sensing mechanisms[1,3,4,9], an increasing degree of control over RNA systems is available through in silico design. Despite this benefit, a core outstanding challenge for this discipline arises when trying to translate the benefits of model-informed design to experimental settings that were not encountered during training[10]. In particular, approaches that design mRNA 5' untranslated regions (UTRs), coding sequences (CDS), or 3' UTRs perform extremely well within their original screening contexts, but can struggle when generalizing to new cargos, cell lines, or when they are ported from in vitro to in vivo models[4,10]. This presents a challenge for therapeutic mRNA design, as screening techniques surveying the span of cell lines, reporters, and animal models necessary to achieve generalization along these axes are unlikely to be available in the near future. In contrast, DNA sequencing technologies have led to the accumulation of a vast body of transcriptomic data profiling countless biological conditions of therapeutic or diagnostic interest[11-13]; however, there are no available techniques for efficiently integrating these datasets with existing sequence-function design tools. Addressing this compatibility gap to mobilize heterogenous sources of data could generalize mRNA design and unlock unprecedented improvements in mRNA expression, specificity, and immunogenicity in a number of therapeutic contexts.

This incompatibility is driven by a misalignment between machine-learning-based design techniques and available data. Predictive models can typically be applied and fine-tuned in a straightforward and general manner[4], but their accuracy sharply reduces with limited data or radically changing experimental context. Generative design approaches built around predictive oracles face an additional, and more challenging, set of constraints. Foundational large language models (LLMs) offer a potential solution to generalizing design given their striking degree of success across human language tasks; however, the power-law scaling governing LLM performance with respect to both available data and computing[14] is fundamentally misaligned to the field of biological design where datasets, if available, are often sparse and foundational developments are continuously emerging. Alternative generative techniques such as generative adversarial networks, variational autoencoders, and activation maximization offer a more direct

ability to design with respect to a predictive oracle but are intrinsically constrained by the degree of generalization defined by the predictor[4,7,9]. Furthermore, these techniques introduce a differentiability requirement throughout their implementation and thus adapting generative model outputs to new grammatical constraints or to extend their design to efficiently account for multiple predictive sources of information is extremely challenging.

While these challenges of generalization affect each portion of the mRNA individually, they are further complicated by the interdependencies of each mRNA component when assembled into a complete transcript. Previous studies outlined the synergistic relationship between naturally occurring UTR elements[16], while combination screening techniques have revealed the large dynamic range available through UTR selection[17]. De novo, rational design under this combinatorial space of considerations is however severely underexplored due to the lack of machine-guided tools that can efficiently design each portion of the mRNA with respect to arbitrary predictive quantities.

Unlike generative techniques, various model interpretation methods exist that return a static encoding of a predictive model's input preferences. While these tools provide a rich understanding of how a black box model constructs its decisions, to the best of our knowledge, these encodings have yet to be explored in the context of designing new inputs. In this work, we investigate a subtle modification to integrated gradients[18] that creates an interpolation between the highest- and lowest-performing members in a dataset used to train our previously published predictive RNA models[4] (Fig. 1a). This technique returns an array encoding the model's preferences for specific nucleotides, which can be traversed in linear time using different algorithmic approaches to design new sequence examples. By evading any requirement of differentiability during the design process, an unprecedented degree of control over the content of new constructs is possible while still exhibiting extreme predicted values. This degree of control translates the benefits of predictor-supervised design to unseen reporter genes, regulatory mechanisms, cell lines, and *in vivo* conditions. Using this toolset to redesign each portion of an mRNA, we apply the concept of coherence to the integration of functional components into a complete vector, identifying elements that constructively or destructively affect mRNA function. We further demonstrate that use of these tools and principles increases translational capacity, extends vector stability, and improves the

dynamic range of RNA regulatory mechanisms across a range of contexts not observed during training, offering a promising solution to the challenge of generalizing machine learning-based mRNA design.

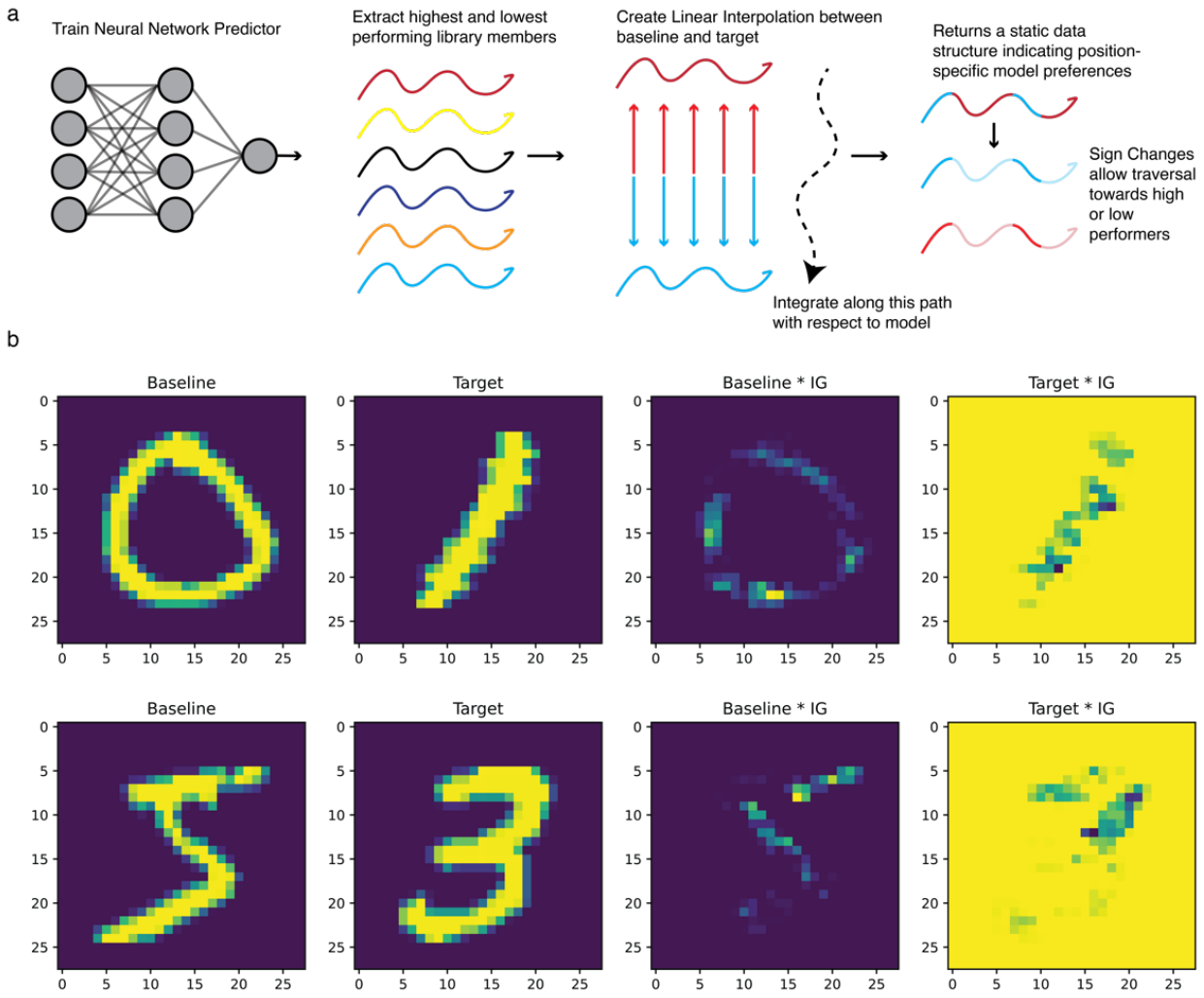

Figure 1 | Overview of Design by Integrated Gradients (DIGs). a, DIGs utilizes a trained neural network predictor to create a static data structure encoding model preferences. A linear interpolation between the highest and lowest examples within the training dataset is constructed, and the model gradients are integrated along this path. The result of this process is an array in which the sign of individual positions indicates preference/aversion towards the model output. **b,** Application of the DIGs technique to the MNIST dataset of handwritten digits. Element-wise multiplication of the DIGs array by the input reveals specifically which pixels were involved in the class assignment and could be utilized as a template for designing new examples of that class.

RESULTS

Design by Integrated Gradients (DIGs)

Integrated gradients is a widely utilized approach for unpacking a predictive neural network's decision making process[18]. This technique involves the identification of a 'baseline' input to a model, typically a zero or noise vector, and the creation of a linear interpolation between this baseline and a 'target' sample from the dataset. The model's gradients are accumulated along this interpolated path, returning an explanatory feature attribution that relates the target's input features to the model's prediction. While integrated gradients provide domain-specific insights across any number of predictive modeling use cases, to the best of our knowledge, this technique has not been explored as an engine for designing new data examples with desirable attributes. Thus, we conceptualized an approach whereby we deploy integrated gradients to create a feature map between two samples demonstrating opposite and extreme functions that is carried forward for downstream design tasks. As a demonstration, we first implemented this approach on the MNIST dataset of handwritten images popularly used as a machine learning benchmark. By creating an interpolation between two distinct input examples (Fig. 1b), integrated gradients returned a feature map which characterized the differences between two samples, rather than unpacking an individual element. Further decomposition of this feature map into the regions containing strictly positive or negative values creates a template for designing a new example resembling either the target or baseline (Fig. 1b). Pairing these templates with any number of programming techniques or alternative sources of information guides the extremely efficient, model-informed creation of new samples without the need to retain differentiability. We refer to this technique of pairing an integrated gradients array with downstream design algorithms as design by integrated gradients (DIGs). DIGs offers an extremely simple and efficient approach for designing each section of an mRNA molecule with respect nearly any predictive attribute.

DIGs designs high-performance 5' and 3' untranslated sequences.

We first deployed the DIGs approach for mRNA design by creating new 5' and 3' UTR sequences designed to impact either the ribosome load (RL) or stability of a reporter protein (Fig. 2). We utilized a SANDSTORM predictive neural network to first map the sequence-function relationship between 5' UTRs and RL, 3' UTRs and RL, and 3' UTRs with stability using previously published datasets[2,19]. After model training, we calculated integrated gradients between the highest- and

lowest-performing members of the dataset, with the feature attribution map carried forward for design. To create a new sequence example using this mapping, we conducted a basic traversal in a 5' to 3' direction, selecting the most favored nucleotide at each individual position for maximal samples and the least favored nucleotide for minimal samples (see Methods).

To explore the effect of the DIGs approach to experimentally impact the stability of a transcript, we first analyzed the sequential and structural compositions of the resulting designs. In line with a large body of previous literature, the DIGs UTR designed to exhibit minimal stability is nearly completely unpaired in structure while exhibiting a large degree of AU-rich elements known to hinder stability[20,21] (Fig. 2c). In contrast, the maximal stability DIGs UTR exhibits a highly folded structural profile and a GC-rich sequence composition which confers degradation resistance[17,22] (Fig. 2c). Adhering to known determinants of function along both sequential and structural axes highlights the ability of a simple DIGs traversal to account for features learned by the dual-input SANDSTORM predictive model, an important consideration for designing realistic sequences. We next inserted the DIGs 3' UTRs into a codon-optimized mCherry expression vector harboring three variations of a 5' UTR (predicted to have moderate RL). Both the mCherry coding sequence and the 5' UTRs were designed by the vector provider (Twist Biosciences). We measured fluorescence generated by translation of the constructs in HEK 293T cells at 24 and 48 hours to assess the impact of stability on total protein expression (Fig. 2d). In line with our objectives, incorporation of the DIGs maximal stability 3' UTR increases expression by 3.8x at 24 hours compared to the minimal stability variant. This improvement persists at 48 hours, conferring a sustained 2.9x improvement. These results highlight that even a simple computational design approach built around a model-informed encoding provides impactful alterations to mRNA performance.

Several studies report improved protein expression through model-based design of 5' UTR segments with respect to ribosome load, including our previous work, GARDN and SANDSTORM. To the best of our knowledge however, the synergistic effects of combining a 5' and 3' element both engineered to exhibit maximal RL is unexplored. Thus, we utilized DIGs to design a high RL 5' and 3' element using independent experimental datasets of ribosome profiling (Fig. 2e). We again characterized the sequential and structural attributes of these elements, recreating our previous findings of a preference for A-rich sequences with minimal secondary

structure incorporated into the 5' UTR to maximize ribosome load[4]. Despite available datasets, model-based design to reengineer the 3' UTR for maximal RL is lacking, leaving the sequential preferences tied to this attribute open-ended. Interestingly, a strong preference for U-rich elements exists within our DIGs maximal RL 3' UTRs, and a nearly perfectly paired secondary structure between the 5' and 3' segments corresponding to maximal RL (Fig. 2e). To assess the degree to which synergistic engineering of RL may impact resulting expression, we evaluated a set of mCherry sequences harboring a DIGs 3' UTR designed to exhibit minimal RL paired with variations to a standard twist 5' UTR (predicted to have moderate RL) and a second set containing a DIGs optimized 5' UTR and DIGs optimized 3' UTR. The combined RL designs display a 4.6-fold improvement in mCherry expression in HEK 293T cells at 24 hours (Fig. 2f), indicating the benefit afforded through selection of UTR elements engineered along the same axis. Interestingly, the assembly of a DIGs-stability element with a DIGs-RL element did not demonstrate synergistic effects, but antagonistic decreases in expression. This finding exemplifies one of the many limitations in model-based mRNA component design when attempting to generalize to out-of-distribution reporter contexts, motivating a subsequent investigation of how different design objectives contribute to overall protein expression.

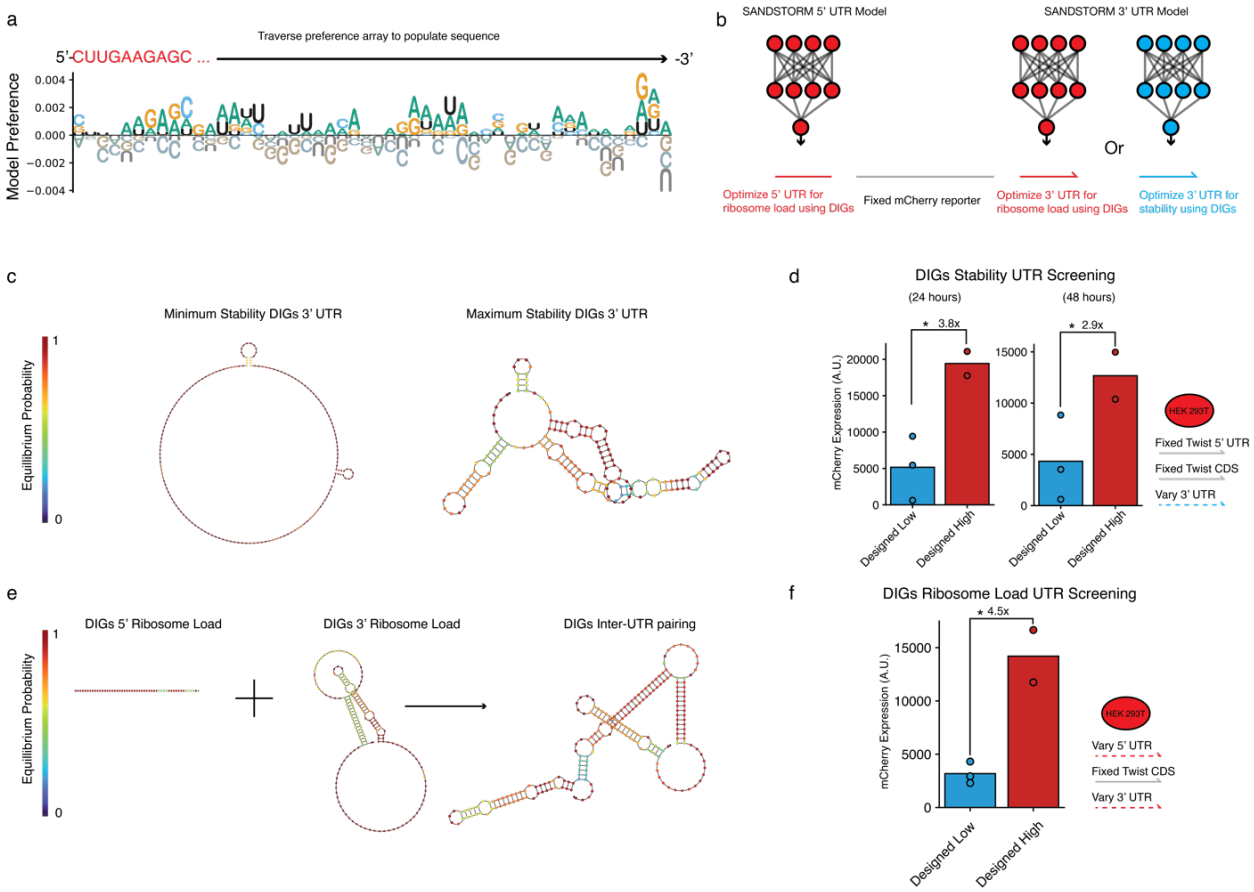

Figure 2 | DIGs UTRs enhance mRNA ribosome load (RL) and stability. a, DIGs preference arrays can be utilized as a template for designing new UTR sequences with targeted experimental attributes. **b,** We designed 5' and 3' UTRs using pre-trained SANDSTORM models and DIGs to tune the stability and RL of a reporter transcript. **c,** DIGs stability UTRs recreate the structural expectations of high- or low-stability sequences. **d,** Engineering 3' UTR stability through DIGs offers improved expression at 24 and 48 hours for an mCherry reporter protein. **e,** DIGs was tasked with designing both the 5' and 3' UTR to enhance ribosome loading, revealing a large degree of expected binding between elements designed from different datasets. **f,** DIGs RL UTRs show enhanced expression of an mCherry reporter protein. Data points indicate the mean for an individual design across three biological replicates, while bars indicate the mean for the design group. Comparisons denoted * indicate p value < 0.05 comparing each design across groups using independent t-test.

Multi-functional DIGs UTRs enable out-of-distribution performance.

The lack of available design tools that can integrate diverse sources of information currently restricts the creation of high performance 3' UTR sequences and is a major bottleneck for generalizing mRNA design. These sequences profoundly alter translation, stability, subcellar localization, and more through an ensemble of RNA binding protein elements, microRNA (miRNA) interactions, and sequence modifications. However, the datasets profiling these interactions are largely siloed. Designing a maximally suited sequence for a single objective is likely to come at the expense of others without explicit accounting. While an immense multi-domain dataset and sufficiently parameterized foundation model may be able to incorporate these diverse sources of information in an unsupervised, 'top-down' manner, DIGs offers the direct ability to design with respect to both predictive neural networks and various sources of domain-specific information, in a 'bottom-up' approach.

As a demonstration of this potential, we explored the integration of the machine-guided DIGs framework with the smallest scale of experimental information available – a functional investigation of a single UTR element[23]. We redesigned a viral 3' UTR sequence to maximize RL or stability while retaining the endogenous regulatory elements within this sequence. This UTR, which contains a manually identified quartet of miRNA-142-3p binding sites, confers cellular tropism to this viral strain, silencing expression within certain immune cells[23,24]. We traversed the DIGs stability and RL arrays as described earlier while bypassing the miRNA annotations, thus leaving them fully intact. We reasoned that the resulting constructs would improve the dynamic range of these elements through either increasing mRNA stability or RL in cellular contexts that lack miRNA-142-3p while retaining silencing in immune cell lines. These designs notably represent a significant departure from the original distribution of sequences utilized to train the predictive models; the viral UTR template is more than double the length of the training samples and previous DIGs designs. In line with our design objectives, the redesigned UTRs returned by DIGs increase the selectivity of an mCherry mRNA vector characterized in HEK 293T (miRNA-142-3p -) and THP-1 (miRNA-142-3p +) cells compared to the endogenous sequence (Fig. 3b). The redesign constructs improve selectivity to 5-fold, 8.29-fold, and 9.92-fold, respectively, compared to the 4.75-fold of the original. These results highlight that the DIGs approach provides

an avenue for incorporating bespoke experimental results with the benefits of model-based design without introducing complex fine-tuning techniques.

To explore how descriptive omics datasets could enable the generalizability of 3' UTR design using DIGs, we first explored a single-cell RNA sequencing dataset profiling murine transcriptomic responses to mRNA vaccination[25]. We hypothesized that within this dataset, upregulated RNA binding proteins are identifiable that provide a stabilizing or enhancing effect to endogenous transcripts in this specific context, and that incorporating such a binding motif into an exogenous vector optimized for other attributes will confer this protective effect. In line with this hypothesis, we identified Cold-inducible RNA Binding Protein (CIRBP) as highly differentially expressed within the various immune cells of the murine spleen following mRNA vaccination (Fig. 3c). CIRBP enhances the stability and expression of different transcripts in various stress response and circadian rhythm pathways[26-28], with putative binding sites identified through CLIP-seq screening techniques[29]. We carried the probabilistic reconstructions of the CIRBP binding motif forward for DIGs-based design while additionally incorporating novel miRNA-122 binding sites, a miRNA that is highly expressed within liver tissue[30]. The optimal placement of these elements within a 3' UTR designed by DIGs we hypothesized will provide a stabilizing effect to the mRNA vector upon cellular exposure to mRNA components, while conferring miRNA-induced degradation in liver cell lines expressing miRNA-122. The combination of these effects while optimizing for stability and RL provides an ideal expression profile for many RNA-based therapeutic applications but is not directly achievable using available sequence-function screens or model-guided design tools.

To incorporate these elements into a DIGs design, we employed an iterative greedy algorithm to locate the ideal insertion point of each regulatory motif. This entailed scanning each regulatory element against the DIGs stability or RL array from a 5' to 3' direction and calculating the implied score of inserting the motif at each position, eventually choosing the highest score for final placement. As shown in Figure 3d, DIGs-guided insertion of the CIRBP and miRNA-122 binding sites recovers the translational capacity of the minimal stability element by 3.21-fold at 24 hours and 4.27-fold at 48 hours, confirming that this approach bridges diverse datasets with the benefits of model-informed design. Taken together, these demonstrations position DIGs as a promising approach to convert the benefits of model-informed mRNA component design into new domains.

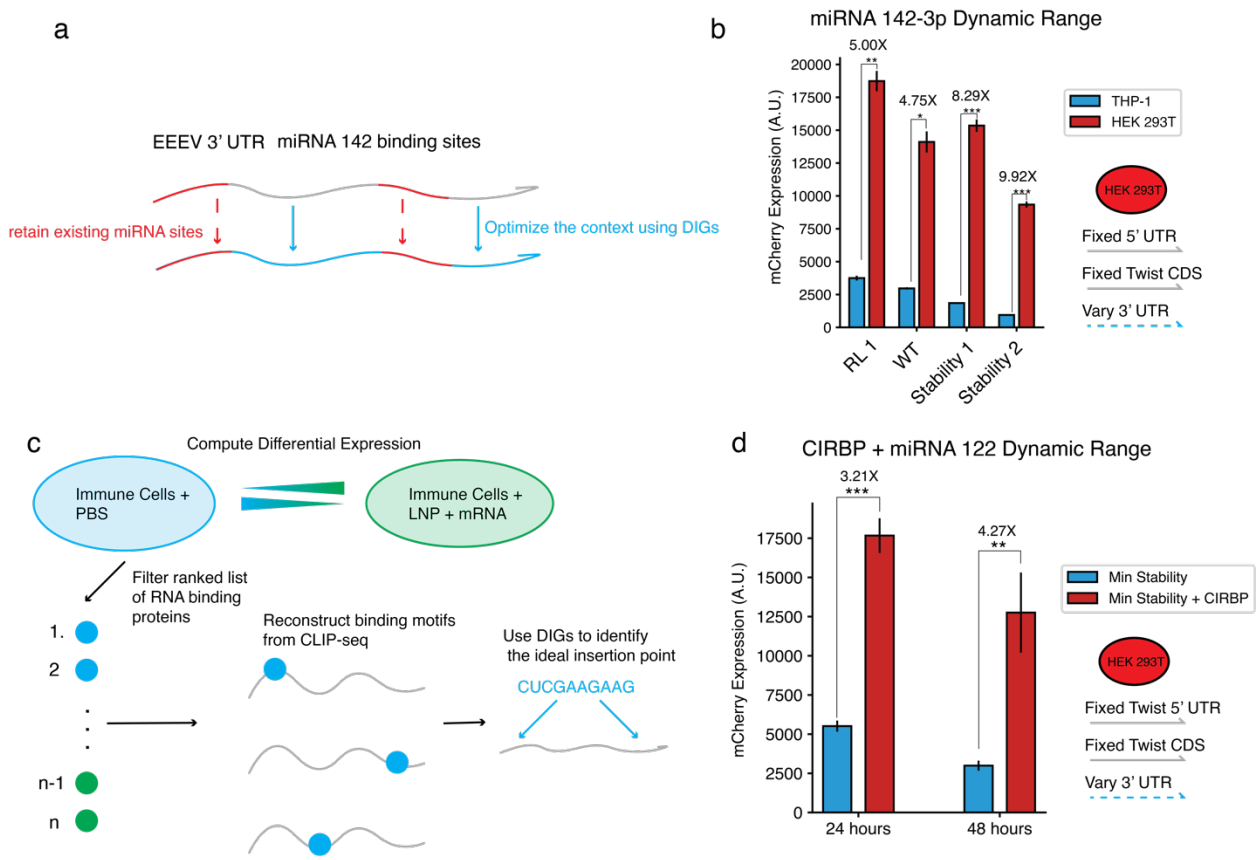

Figure 3 | DIGs incorporates diverse data sources to enable out-of-distribution UTR design. **a,** A naturally occuring viral 3' UTR was redesigned with respect to either ribosome load or stability while retaining 4 endogenous miRNA-142-3p binding sties. **b,** The DIGs redesigned UTRs improve the dynamic range of miRNA-enabled cell type specific expression between HEK 293T and THP-1 cells (* $p<10^{-5}$, ** $p<10^{-6}$, ***$p<10^{-7}$). **c,** A single-cell RNA sequencing dataset was mined for upregulated RNA binding proteins upon exposure to mRNA components. CIRBP was identified through this process as highly upregulated and a CLIP-seq dataset describing binding preferences was utilized to create a probabilistic reconstruction of the binding motif. The DIGs stability array was utilized to identify the optimal insertion point for both CIRBP binding motifs and miRNA-122 binding sites. **d,** The DIGs-guided insertion of these binding elements into our minimum stability UTR element demonstrated a 3.21-fold improvement in expression at 24 hours following transfection and sustained a 4.27-fold improvement at 48 hours post-transfection in HEK 293T cells (*** $p<10^{-4}$, **$p<10^{-2}$). Bars represent the mean expression measurement of an individual design averaged across three biological replicates. Error bars indicate the standard deviation.

Combinatorial screening of DIGs UTRs reveals functional rules for selecting 5' and 3' UTR pairs.

While these results prove useful for translating individual UTR elements to out-of-distribution tasks, the rules that govern component assembly affect and complicate RNA performance. A substantial body of previous literature concluded that engineering mRNA molecules to exhibit maximal stability is the ideal approach when optimizing for overall protein expression[5,8,31-38]. Specifically, maximal RL leads to poor expression through translation-dependent degradation[31], while moderate or minimal RL of a highly stable mRNA mitigates this effect[8,17]. These highly stable sequences predominantly recruit monosomes, evading this degradation pathway while prolonging the lifetime of the transcript and increasing cumulative expression over time. In contrast, high RL leads to a bursting expression pattern at early time points ultimately leading to less total protein over time[17]. This perspective is the basis for many sequences that are actively being explored within the clinic or are already approved for treatment[37].

Given the lack of tools that effectively design each component of the mRNA simultaneously, we hypothesized that the synergistic effects of extreme RL and stability-based UTRs may be incompletely characterized. Understanding this interaction term in greater detail will provide insight on why the insertion of individual optimized UTR elements into alternative vectors can limit performance. We screened a small library of DIGs UTR elements designed to effectively survey the extremes of these attributes with a fixed reporter sequence. This library consists of 3' UTR elements designed to have maximum RL, maximum stability, minimum RL, or minimum stability as well as 5' UTRs designed to exhibit maximum RL, and 5' UTRs predicted to have moderate RL commonly used in Twist Biosciences vectors. Importantly, the sequence-function mappings guiding the creation of each of these components originate from different experiments, functional assays, and laboratories, leaving the principles guiding their rational selection completely open-ended. We assessed this library at multiple time points to potentially identify distinct temporal profiles offered through each attribute. The expression profiles of the sequences demonstrate a wide dynamic range at 6, 24, and 48 hours post-transfection respectively, between the highest and lowest performers (Fig. 4a), highlighting the importance of the rational assembly of components to design complete, high-performance mRNAs.

Profiling this library of elements revealed a surprising result – combining a 5' and 3' element designed for maximum RL creates a robust expression profile over time rather than demonstrating clear signs of translation-dependent degradation. The same improvement in expression is also seen when selecting a 5' element for moderate RL and a 3' element for maximum stability. In contrast, the combinations of orthogonal elements designed to exhibit extreme traits (such as high RL 5' and high stability 3') nearly universally decrease expression (Fig. 4a). This result may explain the stability-focused findings of previous studies, where the placement of a high RL 5' UTR within an otherwise unmodified transcript counterintuitively reduced total protein expression[17,31], while highly structured 5' UTRs that would be expected to limit ribosomal binding increased total protein output. For the remainder of this paper, we will refer to combinations of mRNA elements designed to emphasize the same attribute (e.g. RL or stability) using the term coherent selection, while combinations of RNA elements where there is a mismatch in attributes will be termed incoherent selection.

A time-course analysis of various UTR elements at 6, 24, and 48 hours post-transfection reveals an additional surprising finding: a simple coherent selection process effectively optimizes total protein over time, regardless of the axis for design. While we identified a slight shift in expression favoring the stability designs at 48 hours, no significant differences are present between the orthogonal design approaches in terms of fluorescence AUC calculated over all time points, while most combinations of incoherent UTR elements show minimal expression. Individual UTR combinations additionally validate that designs created through stability engineering still demonstrate high early expression, while those designed through RL persist despite translation-dependent degradation effects (Fig. 4a). This benefit of coherent UTR selection additionally held when considering our synthetic miRNA-142 redesign constructs (Fig. 4b), which fall greatly outside the length distributions of the original dataset used for model training, indicating the generalizability of this principle. To investigate the ubiquity of this principle, we further screened a subset of the coherent and incoherent design UTR pairs in THP-1 cells, an immortalized human macrophage line (Fig. 4c). Interestingly, while the coherent pairs still outperformed their incoherent counterparts along both axes, we found a strong preference for the stability-based vectors and a relative reduction in the RL-based designs. This finding suggests that individual cell types may exhibit preferences for tunable design objectives; a detailed understanding of this

principle will enhance model-based mRNA design in out-of-distribution experimental settings and potentially serve as a lever for engineering cell type specific expression.

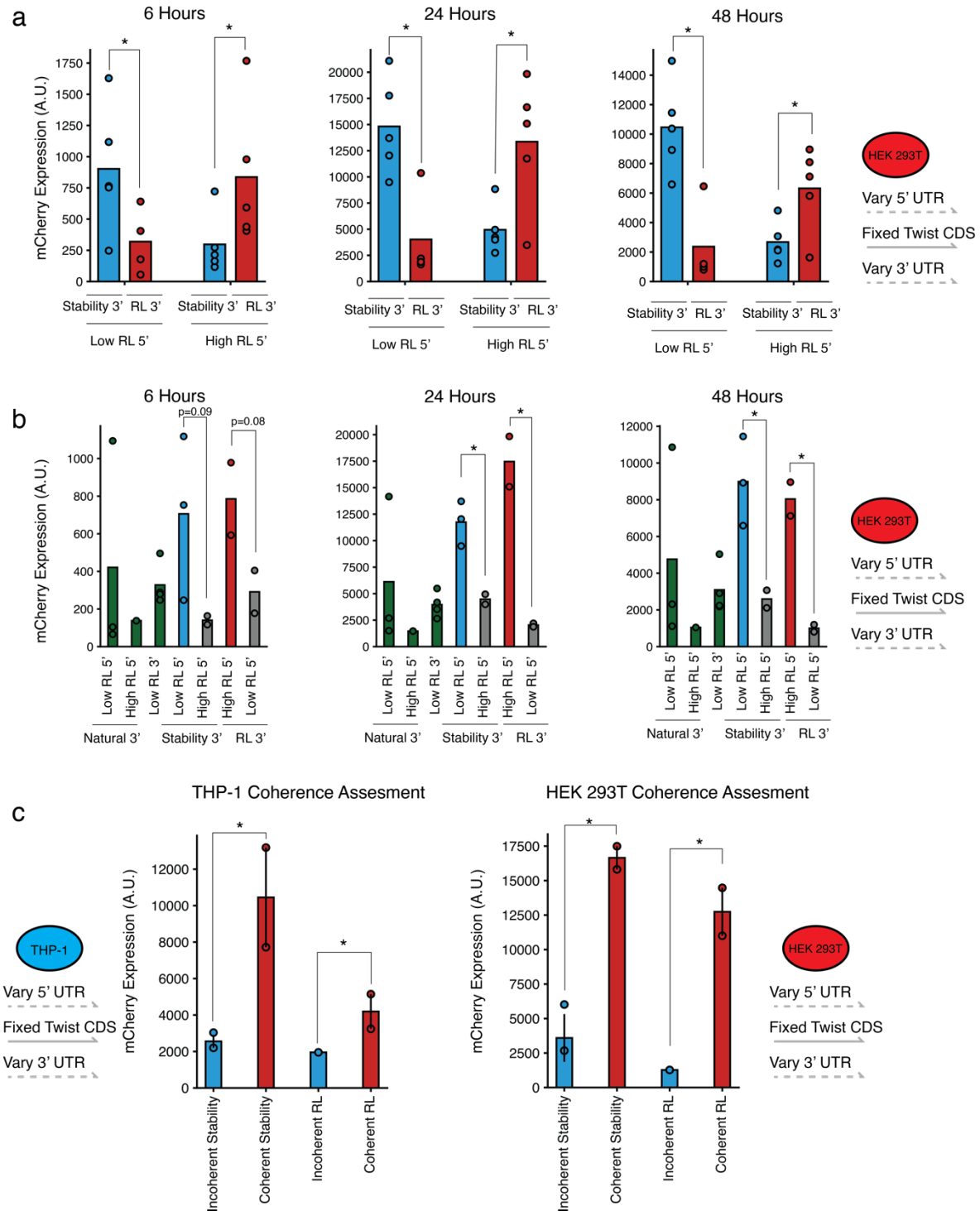

Figure 4 | Combination screening reveals coherent UTR selection criteria. a, A library of engineered DIGs 3' UTRs was assessed in combination with DIGs and Twist 5' UTRs in HEK

293T cells. mCherry expression at 6, 24, and 48 hours is reported. Comparisons denoted * indicate $p<0.05$ measured using Brunner-Munzel non-parametric statistical test. **b,** The coherent selection criteria remains informative when considering miRNA-142 redesigns which harbor four fixed miRNA binding sites and are double the length of original DIGs designs. Comparisons denoted '*' display $p<0.05$ measured using Brunner-Munzel non-parametric statistical test. **c,** Expression measurements from a subset of the coherent and incoherent design combinations reveal that coherent selection remains informative across cell types, and also reveals that different cells may exhibit preferences for specific axes of design. Comparisons denoted '*' display $p<0.05$ measured using independent t-test for each pairwise sample comparison within the groups across three biological replicates. Data points in all plots represent the mean expression for an individual design across three biological replicates. Bars represent the mean measurements from a specific group of designs. Error bars indicate the standard deviation of the design group.

Transcriptomic analyses challenge the assumptions of sequence design paradigms.

Building off of our initial UTR findings, we sought to uncover if various transcriptomic datasets could be mined for informative signals to enhance model-based mRNA design in diverse settings. A robust body of literature details the benefit of ubiquitously designing coding sequences and 3' UTRs to exhibit the most extensive secondary structure possible[8], while incorporating the most abundant codons possible. The motivation for this design objective has been informed by various in vitro experiments which demonstrate that maximally folded mRNA molecules are more stable against hydrolysis and exhibit longer half lives in immortalized cell lines[8,17,37]. This paradigm does not however consider the dynamic impact of innate immune sensing present within in vivo models, tissue-specific differences in codon abundance, or the potential synergistic effects of coherent mRNA component selection on overall expression. Given the dynamic variance of these attributes across the settings in which mRNA could be useful, refining our understanding of how these factors should be incorporated into design remains a critical outstanding challenge.

While increased immune stimulation may be advantageous for vaccine development, there are many therapeutic applications for mRNA in which immune activation must be avoided[39]. Furthermore, the interaction between total expression of an RNA payload and dsRNA-activated immune degradation has yet to be explored thoroughly in the context of design. We conducted an

in-depth analysis of transcriptomic changes following exposures to different types of RNA to assess whether a global preference for different structural profiles or codon usage contributes to differential expression (Fig. 5). Beginning with a dataset mapping the time-course transcriptomic response of human macrophages to infection by the chikungunya virus[40], a self-replicating RNA alphavirus[41], we found that the length-normalized minimum free energy (MFE) of upregulated transcripts is significantly greater than down-regulated transcripts across both 8 and 24 hours (Fig. 5a). We continued this analysis using a dataset characterizing the tissue-specific transcriptomic response to chikungunya infection in both macaque and mouse models[42]. In these in vivo settings, a similar association of minimally structured upregulated genes is seen in both the lungs and spleens of these animals, while the liver demonstrates the opposite trend in both animals (Fig. 5c). Of the tissues analyzed, the heart is the only example that demonstrates a preference for less-structured transcripts in macaques, but no significant difference in mice. This finding could suggest a tissue-specific preference for differing degrees of secondary structure or codon usage in response to exogenous RNA exposure.

To isolate this effect from potential impacts of viral proteins on immune response, we next sought to focus on experiments that describe the response to an mRNA vaccine enclosed within a lipid nanoparticle at single-cell resolution in mice[25]. We computed the differential expression of cells within the excised muscle of mice exposed to a COVID-19 vaccine mRNA enclosed within an LNP and those exposed to an empty LNP, revealing a striking increase in length-normalized MFE for upregulated genes across nearly every cell type (Fig. 5b). Interestingly, the structural differences between up- and downregulated genes does not vary as a function of spike mRNA count, indicating a global effect at the site of vaccination. Furthermore, when isolating the upregulated genes of specific cells, a clear difference can be seen between immune and muscle cells, indicating that the optimal composition of codons may differ depending on the cell type and condition of interest (Fig. 5d). Understanding the exact role of mRNA codon usage and secondary structure in a context-specific manner could greatly enhance our ability to execute rational mRNA design in previously unobserved settings, but a detailed explanation of these contributing factors remains unobservable with existing tools.

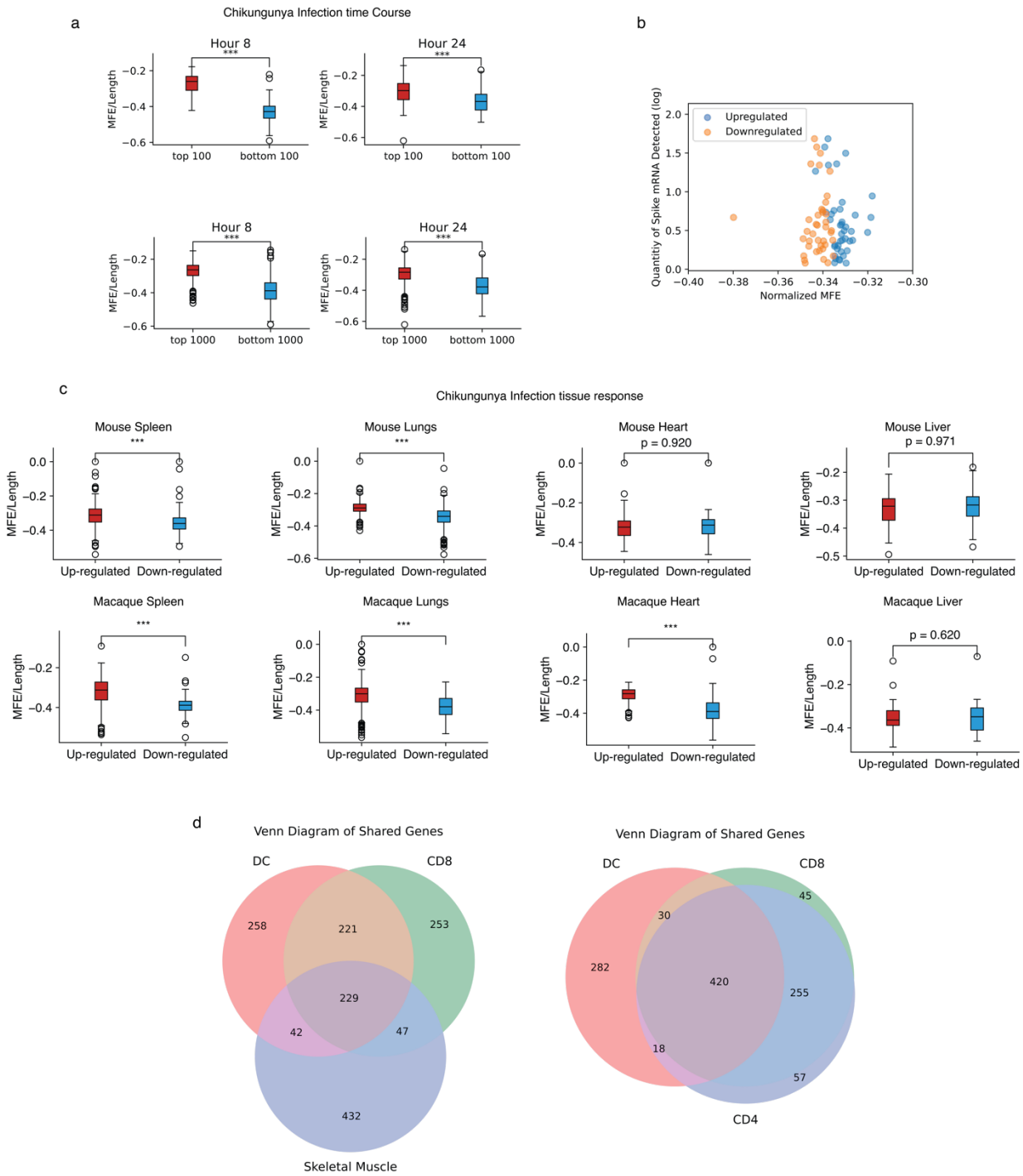

Figure 5 | Transcriptomic analyses uncover tissue-specific structural and codon preferences. a, Differential expression analysis of human macrophages following chikungunya infection reveals that upregulated transcripts exhibit reduced secondary structure compared to downregulated transcripts at both 8 and 24 hours after infection. **b,** Following COVID-19

vaccination, nearly every cell type in the murine vaccination site demonstrates a reduction in secondary structure for upregulated genes compared to down-regulated genes. **c,** A tissue-specific preference for different degrees of secondary structure can be seen in both mice and macaque. In both animals, the lungs and spleen demonstrate a preference for reduced structure, while the liver demonstrates a preference for increased structure. **d,** The specific genes that are upregulated following COVID-19 vaccination show heterogeneity when comparing the muscle and spleen, motivating an exploration of tissue-specific codon preferences.

DIGs CDS designs uncover the contributions of secondary structure and codon usage on mRNA performance.

While these analyses provide a motivation to explore context-specific forward design of coding sequences, converting these findings into an efficient CDS design approach is currently unsatisfiable with available techniques. Adherence to an arbitrary amino acid chain poses a challenge for gradient-based design techniques, the findings of these transcriptomic analyses are not easily amenable to train sequence-function predictors, and thermodynamic tools explicitly optimize for extensive secondary structure while using species-wide codon adaptation indices. In contrast, DIGs is compatible with arbitrary predictive oracles, can account for grammar in a flexible manner, and can be utilized for bi-directional design to parse out the explicit contributions of secondary structure and codon usage on mRNA performance while exploring unprecedentedly unstructured CDS designs in terms of minimum free energy (MFE) (Fig. 6).

We aimed to investigate this problem by creating a library of synonymous coding sequences that varied based on their degree of secondary structure and their use of context-specific codon preferences. To create MFE-varying CDS designs using DIGs, we first utilized a suite of SANDSTORM predictors to estimate the MFE of various CDS datasets (Fig. 6a). These datasets were created by sampling hundreds of thousands of examples of the same protein coding sequence in silico (e.g. mCherry, Green Fluorescent Protein), using uniform codon biases and were labelled according to their length-normalized MFE calculated by thermodynamic models (see Methods). Each predictive model was trained on an individual synthetic protein dataset and queried using integrated gradients between the highest and lowest MFE example. This array therefore contains

globally informed, position-specific information related to the secondary structure of possible CDSs (Fig. 6b) that can be traversed in linear time with respect to the target amino acid chain.

Unexpectedly, a simple greedy algorithm selecting the highest scoring valid codon suggested by this array efficiently generalizes beyond the initial training dataset of 500,000 sequences when tasked with designing either extremely highly or lowly structured sequences (Fig. 6c,d). DIGs additionally generalizes far beyond an analogous greedy approach that directly selected codons using a thermodynamic model informed over the entire sequence. This emergent property reflects the advantage of utilizing globally informed information encoded within the position-specific array returned by the DIGs approach. When compared to LinearDesign set to ignore codon biases, mCherry sequences returned by a simple greedy selection using DIGs do not reach the same extreme minimum, however, the traversal of the DIGs array could in principle be extended to the more sophisticated lattice parsing approach that empowers LinearDesign. The runtime of DIGs does however outperform either LinearDesign or greedy design based on thermodynamic models by several orders of magnitude while allowing bidirectional optimization with a simple pythonic implementation (Fig. 6c). Furthermore, while we explore secondary-structure-based design of the CDS using DIGs, this approach is the only available tool that can account for CDS-specific grammar while designing with respect to arbitrary differentiable predictive models. This feature will be increasingly essential as DNA screening technology continues to be compatible with extended library lengths and direct sequence-function readouts of CDS elements become possible.

Unlike other methods that could potentially design a sequence with respect to a predictive neural network, DIGs traversals can be trivially supplemented with arbitrary, non-differentiable sources of information. We leveraged this feature to incorporate context-specific codon usage metrics into our designs that varied based on structure. This entailed calculating the most upregulated genes in a tissue and condition of interest, collating the codon usage of these genes into a unified codon adaptation index (CAI) that could be queried, and supplying these suggestions in parallel to the structural considerations to the DIGs traversal during runtime (see Methods). Our final library therefore contained seven CDS variants of mCherry. Three DIGs designs used the codon preferences of the murine spleen (encompassing AT-rich codon preferences) following chikungunya exposure and demonstrated varied structures of -64.4 kcal/mol, -87.5 kcal/mol, and

-215.8 kcal/mol. These minimal structure elements represent a class of coding sequence that are unachievable using other available techniques. An additional two DIGs designs utilized the codon preferences of the liver in this same experiment (encompassing GC-rich codon preferences), demonstrating MFEs of -312.7 kcal/mol and -323.5 kcal/mol. The final two library members were a standard, codon-optimized mCherry sequence from Twist and a construct returned by LinearDesign using, with MFE values of -272.5 kcal/mol and -431.3 kcal/mol respectively.

We assessed the DIGs liver and LinearDesign constructs using the coherent UTR pair consisting of the full-length twist 5' UTR and stability-based DIGs redesigned 3' UTR containing miRNA-142-3p binding sites in HEK 293T cells. We hypothesized that this group of elements constitute the most direct extension of previous CDS design approaches, which emphasized moderate-to-low ribosomal loading and stability maximization, while incorporating our UTR selection principles and out-of-distribution specificity offered through miRNA-142-3p binding. This experiment revealed that the lowest MFE DIGs design created with context-specific codons exhibit the maximal degree of expression at both 24 and 48 hours, followed by LinearDesign, twist, and the final DIGs design (Fig. 6e). Intriguingly, the optimal DIGs construct outperforms LinearDesign despite being extensively less structured, indicating that context-specific codon selection may supersede structure when optimizing for expression. The performance of this vector highlights how individual experimental findings, context-specific descriptive datasets, and machine-guided design can be unified to create high performance mRNA vectors in the absence of direct high-throughput screening data.

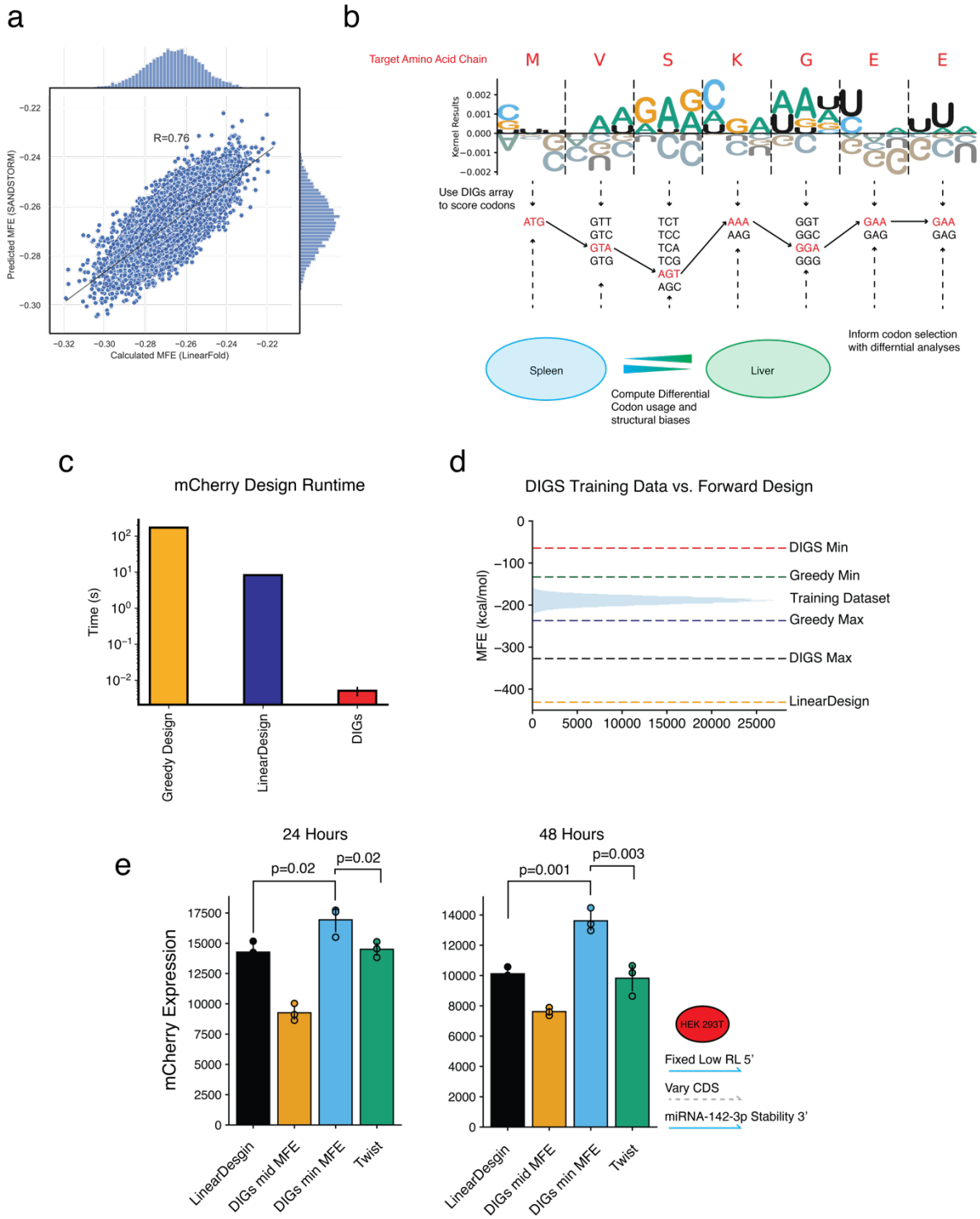

Figure 6 | DIGs efficiently designs model-informed coding sequences. a, A SANDSTORM model trained to predict the minimum free energy (MFE) of different coding sequences can be used to inform structure-based DIGs CDS design. **b,** The DIGs array is traversed in linear time to select the optimal codon at each position based on structural considerations and the tissue-specific

codon index between multiple cell types. **c,** DIGs-based design is extremely efficient, enabling several order-of-magnitude improvement compared to LinearDesign or an analogous greedy approach. **d,** DIGs forward designs extend beyond the structural attributes of the sequences accounted during training, an important emergent property of this approach. **e,** DIGs coding sequences designed to exhibit maximal secondary structure while utilizing the codon preferences of the liver following mRNA exposure shows improved expression at 24 and 48 hours compared to either LinearDesign or twist. Bars indicate the mean expression measurement across three biological replicates.

DISCUSSION

In the most basic deployment settings, we have shown that simple design approaches built on top of DIGs can modulate the RL, stability, and secondary structure of an mRNA sequence under the guidance of a predictive model and generalize beyond the examples encountered during training. By evading the requirement of differentiability, DIGs further provides a generalized way to unify diverse sources of experimental information with the benefits of model-guided sequence design. This creates an important advantage on out-of-distribution design tasks, where forward-engineered sequences can be created without exhaustive additional screening or the computational overhead of a foundation model. DIGs addresses many of the hurdles that affect the deployment of machine-guided RNA design methods outside of basic research settings, offering a promising avenue to enhance their practical utility.

By designing each section of the mRNA, we have further uncovered the concepts governing the coherent assembly of de novo components into complete, high-performance transcripts. These findings have extended our understanding of the rules guiding rational mRNA design and could be used to guide future generative techniques and sequence-function screening experiments. We have further refined the consensus of the role of stability on mRNA performance, a critical finding that should enhance rational mRNA design across several clinically relevant indications.

Despite the promise of the DIGs technique across each of the described design tasks, some computational and biological limitations persist. In our current implementation, the objective of DIGs is to provide a single optimal sequence for the given attribute of a predictor, limiting the

diversity of generated outputs. While we have shown that this approach can generalize beyond observed training samples into unexplored regions of the sequence space, creating diverse examples with similarly extreme attributes is a fundamental objective of generative design. We expect that far more sophisticated programming techniques built on top of DIGs arrays will enable this task. As DIGs returns a data structure amenable with dynamic programming, it is possible that through careful algorithmic design, novel, globally optimal sequences for a given model could be returned, a guarantee that is unsatisfiable through other means. Furthermore, while coherent selection applies across many of the out-of-distribution demonstrations that we present throughout this work, we do not expect this principle to apply exactly across the range of naturally occurring UTR sequences. The principles of orthogonal, synergistic design pathways do however provide a strong foundation through which a comprehensive model of UTR interaction could be constructed.

In conclusion, DIGs is an extremely efficient method for closed-loop rational mRNA design. RNA regulatory mechanisms, out-of-distribution reporters, varying cell lines, and context-specific descriptive data can each be incorporated with model-based optimization through DIGs in a straightforward manner. These attributes will greatly improve the utility of rational mRNA design across diverse therapeutic modalities and could immediately enhance many RNA vectors that are being explored in clinical applications. More generally, DIGs provides a tight feedback loop for converting DNA sequencing descriptions of a biological condition of interest into potent effectors of de novo biological outputs.

METHODS

SANDSTORM Datasets and Predictive Model Training

5' UTR

The SANDSTORM 5' UTR RL model was trained as described in our previous work[4]. Briefly, the data consists of 50nt UTR elements that were experimentally screened using polysome profiling techniques. Input data points are transformed into a paired one-hot encoding and structural contact map, at which point they are passed to the model for prediction. The model was trained for 20 epochs on this data using a train-test split of 80-20 using the Adam optimizer with a learning rate of 0.001.

3' UTR

The same SANDSTORM architecture was again applied on a previously published dataset of 3' UTR performance[19]. In this experimental setup, both polysome profiling and mRNA abundance (used as a proxy for stability) were reported for a library of viral genomic tiles inserted into the 3' UTR of a reporter. We reprocessed the raw read data from this study using the bowtie 2[44] alignment program under *sensitive local* settings and filtered aligned reads for strict barcode matches using a custom python framework. For mRNA abundance, the MPRA R[45] package was utilized to calculate RNA reads normalized to DNA reads as the target variable for model training, with normalized reads averaged across replicates. For RL, the aligned reads for each polysome fraction were again filtered for exact barcode matches and a single RL quantity was calculated using the formula from Seo et al.[19]. Input sequences were processed for prediction tasks as described above and independent model training was conducted for mRNA abundance and RL respectively using the same SANDSTORM architecture and training settings[4].

CDS

SANDSTORM was utilized to estimate the MFE of CDS elements by labelling a synthetic dataset of synonymous sequences with thermodynamic models. Briefly, 500,000 random mCherry coding sequences (711nts) were created in silico by uniformly sampling from all valid codons. The secondary structure and MFE of each element was calculated using LinearFold[46]. The length-normalized MFE was calculated and carried forward as the target variable for model predictions, while the one-hot encoded sequences and structural contact maps for each CDS member were used as inputs. The model was trained with the same hyperparameter settings for 20 epochs.

5' UTR-CDS Interaction Term

Previous studies have highlighted improved RL by reducing the structural interactions of the 5' UTR and 5' portion of the CDS[37]; in attempt to leverage this during DIGs design we conducted a new training routine to predict the MFE between a specific CDS and library of synthetic UTR elements. To create labelled training data for this task, an in-silico library of 10,000 generated UTR sequences returned by the GARDN-UTR model was created spanning a range of GC content and predicted RL. We utilized NUPACK[47] to predict the MFE between each of these UTR elements and an individual CDS element of interest, repeating this process 3 times for each of the DIGs

spleen CDS elements. A SANDSTORM predictive model was then trained on the MFE value using each UTR as an input while utilizing the same training settings described previously.

Integrated Gradients Calculations

Integrated gradients[18] was calculated throughout this work on the one-hot encoded sequence channel of the SANDSTORM model architectures. One-hot encoded sequence inputs transform strings of nucleotides into a vector of shape $4 \times N$, where N is the length of the sequence and each channel encodes a 0 or 1 to indicate the presence of a specific nucleotide at each position from 0 to N. Integrated gradients is defined as the path integral computed over the straight line path between a baseline, x', and a specific input sample, x. Assuming a differentiable predictive model $P_\delta(x)$ accepting i total input dimensions, integrated gradients along the ith dimension is defined as follows:

(1) $$IntegratedGradients_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial P_\delta(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

In the original integrated gradients work, x' is conceptualized as a zero- or noise-vector. The resulting data structure therefore offers an explanatory map for the portions of the input, x, which are contributing to the output score, $P_\delta(x)$. In this work, we reframe the technique to assign both x and x' to be opposite and extreme examples of the feature which P_δ predicts. Without altering the computation, this shifts the interpretation of the output to an explanatory map describing the features which contribute to the opposite and extreme values of x and x'. The output of this process is therefore a static data structure encoding model preferences across the maximal observed outputs of P_δ, which can be traversed as a template for designing new example data inputs with extreme features. These traversals can take the form of nearly any programming technique without the requirement of retaining differentiability. In line with this interpretation of the resulting data structure, we found that assigning x and x' as opposite and extreme values enabled a more robust output design than assignment of x'as a noise or zero vector.

For the integrated gradients calculation, we utilized between 100-250 interpolated points between x and x'. Additionally, in cases where the output design from our DIGs traversals returned examples with more extreme attributes than any labelled example, we repeated the process using

this new example as either x or x'. We found that this process typically plateaued between 1-5 repetitions.

Transcriptomic Analysis

Precomputed differential expression measurements were utilized in our investigations of chikungunya time course and tissue-specific expression reported by the original authors[40,42]. For time course infections, the top N genes were ranked based on log fold change. For the tissue-specific chikungunya infection in mice and macaques, we filtered for significantly differentially expressed genes in terms of p values ($p<0.05$) and then utilized all upregulated and downregulated transcripts for further analysis. Endogenous CDSs for human, mouse, and macaques were exported from BIOMART[48], with entries shorter than 100nt pruned from further analysis. MFE values for CDS elements were calculated using LinearFold[46]. For BIOMART entries which returned multiple transcript variants, we utilized the mean MFE across reported variants while utilizing random isoform selection for codon preference calculations. In both analyses the unexposed samples were utilized as the baseline for differential expression.

scRNA analysis of the mouse vaccination study[25] was conducted using the scanpy package[49] and the count matrix reported by the original authors. The original paper presented separate scRNA experiments form both the site of vaccination (muscle) and the lymph nodes. Differential MFE analysis was conducted using the skeletal muscle cell group following first LNP exposure as the baseline sample to isolate effects from dsRNA exposure. Upregulated or downregulated genes were plotted after filtering for adjusted $p<0.05$ and $|\log_2(\text{fold-change})| >1$. For RNA binding protein analysis, CD4+ T cells upon exposure to PBS were utilized as the baseline for differential analysis. In this case PBS was used as RNA binding proteins upregulated in response to either the LNP or mRNA would be useful for therapeutic applications. Manual processing of upregulated genes (adjusted $p<0.05$, $|\log_2(\text{fold-change})| >1$) was conducted to identify potential RNA binding proteins known to have a positive effect on mRNA stability or translation from the literature.

DIGs design algorithms

5' UTR

DIGs design of 5' UTR elements was conducted using two setups. In the first basic example, a 50-nt 5' UTR was created by traversing the integrated gradients array returned by the SANDSTORM 5' UTR RL predictor in a 5' to 3' direction, selecting the most favored nucleotide at each position. The target and baseline samples were selected as the highest and lowest-labelled examples from the original dataset. In the second setup, we conducted joint design based on the SANDSTORM model which predicted 5' UTR RL and the SANDSTORM model which predicted the MFE between a library of 5' UTRs and individual coding sequences. Using both of these DIGs arrays, we traversed from a 5' to 3' direction, applying a weight (alpha and beta) to the suggested nucleotide at each position from each channel, allowing us to tune the relative contributions of each in the resulting design.

3' UTR

The basic stability and RL 3' UTRs were 170 nt in length and designed using the same 5' to 3' traversal that enabled 5' UTR design, selecting the most favored nucleotide at each position. For the miRNA-142-3p UTR redesign experiments, we utilized the binding site annotations from Trobaugh et al.[23,24] and conducted a slightly modified traversal. In this setting we populated the sequence in the 5' to 3' direction taking the most favored nucleotide at each position, but left the annotated binding sites unaltered. The template UTR was 353 nt long; over 2x the length of those within the training dataset used to create the DIGs template. To achieve this length in our synthetic constructs, we stitched together multiple copies of the DIGs array while retaining the binding sites in their absolute position relative to the template.

To create the synthetic CIRBP and miRNA-122 element, we first examined different probabilistic reconstructions of potential CIRBP binding motifs using several independent methods. First was the MEME[50] motif automatically generated from hits listed within the CLIP-seq database entry[29]. The second was identified from the following publication[26]. Third was utilizing standard integrated gradients on a SANDSTORM model trained to predict CLIP-seq peaks. Fourth was using the pyhmmer[51] toolkit to construct a profile HMM from CLIP seq binding sites. Finally, we employed a variation of the RaptGen Variational Autoencoder architecture[52] to generate novel instances of

the CIRBP binding site reconstructed from CLIP-seq data. While these methods varied greatly in their respective outputs, the motif CUCGAAGAA was consistently identified across them and used as the basis for forward design. miRNA-122 binding sites were defined as the complete reverse complement of the sequence identified from the human miRNA tissue atlas[53]. The insertion of these elements was guided by traversing the maximum stability template from a 5' to 3' direction and calculating the total score implied by the insertion of the motif at each position, eventually taking the highest score. The motif was inserted into the minimum stability sequence at the corresponding location. For successive motif insertions into the UTR, the previously selected loci was omitted from future passes to prevent the inclusion of overlapping elements. We inserted 2 copies of the CIRBP motif and 2 copies of the miRNA-122 binding site per UTR.

CDS

mCherry CDS designs were created using the integrated gradients array of the SANDSTORM-MFE predictive model. When designing with respect to purely structure, we traversed the array from 5' to 3' in incremental steps of size 3. Each possible synonymous codon for a given amino acid ($c \in C(a_i)$) was indexed into this array (φ) and assigned a score based on the sum of each individual nucleotide ($\varphi_{i,c}$). We selected the highest possible sum for inclusion into the sequence. When incorporating context-specific codon information, we supplemented this process with a custom CAI calculated from the upregulated genes identified in our transcriptomic analyses (ψ). Each synonymous codon was now jointly considered for its impact on secondary structure by referencing the DIGs array and the favorability of the codon in the specific context through the CAI. Weight parameters were multiplied by each value to modulate the relative contribution of each information channel in the resulting design. The selection of a specific codon c at position i of the growing amino acid chain can be expressed as follows:

(2) $$c_i = \frac{argmax}{c \in C(a_i)} [\alpha * \ \varphi_{i,c} + \ \beta * \ \psi(c)]$$

Plasmid and Template Construction

All 5' UTRs and 3' UTRs were purchased as single stranded DNA ultramers from Integrated DNA Technologies. Coding sequences (CDS) for mCherry were purchased as gene fragments from Twist Bioscience. All 5' UTR, 3' UTR, and CDS combinations were cloned together with Gibson

assembly purchased from New England Biolabs into vector backbone from Twist Biosciences. T7 promoter (GCGCTAATACGACTCACTATAGGG) was upstream of each 5' UTR. Plasmids were cloned in DH5α *Escherichia coli* and purified using a QIAprep Spin Miniprep Kit (Qiagen). Plasmids were linearized with PmlI and SapI for 15 h at 37 °C and purified using Monarch Spin PCR & DNA Cleanup Kit (New England Biolabs).

mRNA Preparation

mRNA was synthesized using mMESSAGE mMACHINE T7 Transcription Kit (Thermo Fisher Scientific) with 500 ng template. IVT reactions were performed at 37 °C for 2 h then treated with DNase for 15 min and post-transcriptional polyadenylation using a Poly(A) Tailing Kit (Thermo Fisher Scientific) for 45 min. mRNA was purified using Monarch Spin RNA Cleanup Kit (New England Biolabs) and eluted in RNase/DNase-free water. Quantity was determined with Nanodrop One Spectrophotometer (Thermo Fisher Scientific) and concentrations were adjusted to 500 ng/µl before storing at −80 °C.

Cell Culture and Transfection

HEK 293T cells were grown in DMEM containing 10% FBS and 1% penicillin–streptomycin, plated at 10,000 cells per well in 96-well plates and allowed to adhere overnight. mRNA was transfected using Lipofectamine MessengerMax Transfection Reagent, according to the manufacturer's protocol (Thermo Fisher Scientific). Six, twenty-four, and forty-eight hours later, media was removed, and cells were treated with 0.05% trypsin for 3 min to unadhered the cells from the plate. DMEM was added to stop trypsin. Cells were analyzed for mCherry expression with Attune NxT Flow Cytometer (Thermo Fisher Scientific) and analyzed using FlowJo (v10.8.1).

THP1 cells were grown in RPMI containing 10% FBS and 1% penicillin–streptomycin in a 96-well plate at 100,000 cells per well in 200 µl media. On the same day mRNA was transfected using Lipofectamine MessengerMax Transfection Reagent, according to the manufacturer's protocol (Thermo Fisher Scientific). Twenty-four hours later cells were analyzed for mCherry expression with Attune NxT Flow Cytometer (Thermo Fisher Scientific) and analyzed using FlowJo (v10.8.1).

Acknowledgments

This work was supported by startup funds from Boston University; and a National Institutes of Health (NIH) Director's Transformative Research Award (R01EB037112) to A.A.G. A.T.R. was supported by the NIH Training Program in Quantitative Biology and Physiology (5T32GM008764). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing Interests

A.A.G. and A.T.R. have filed patent applications related to this work. A.A.G. holds equity in Gardn Biosciences and En Carta Diagnostics. A.T.R. holds equity in Gardn Biosciences. W.W.W. holds equity in Senti Biosciences, 4Immune Therapeutics and Keylicon Biosciences. M.W.G. holds equity in Keylicon Biosciences. The remaining authors declare no competing interests.

REFERENCES

1. Sumi, S., Hamada, M. & Saito, H. Deep generative design of RNA family sequences. *Nat Methods* **21**, 435–443 (2024).
2. Sample, P. J. *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* **37**, 803–809 (2019).
3. Wong, F. *et al.* Deep generative design of RNA aptamers using structural predictions. *Nat Comput Sci* **4**, 829–839 (2024).
4. Riley, A. T., Robson, J. M., Ulanova, A. & Green, A. A. Generative and predictive neural networks for the design of functional RNA molecules. *Nat Commun* **16**, 4155 (2025).
5. Morrow, A. K. *et al.* ML-driven design of 3' UTRs for mRNA stability. 2024.10.07.616676 Preprint at https://doi.org/10.1101/2024.10.07.616676 (2024).
6. Iwano, N., Adachi, T., Aoki, K., Nakamura, Y. & Hamada, M. Generative aptamer discovery using RaptGen. *Nat Comput Sci* **2**, 378–386 (2022).
7. Castillo-Hair, S. M. & Seelig, G. Machine Learning for Designing Next-Generation mRNA Therapeutics. *Acc. Chem. Res.* **55**, 24–34 (2022).
8. Zhang, H. *et al.* Algorithm for optimized mRNA design improves stability and immunogenicity. (2023).

9. Valeri, J. A. *et al.* Sequence-to-function deep learning frameworks for engineered riboregulators. *Nat Commun* **11**, 5058 (2020).
10. Schlusser, N., González, A., Pandey, M. & Zavolan, M. Current limitations in predicting mRNA translation with deep learning models. *Genome Biology* **25**, 227 (2024).
11. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* **2015**, 951–969 (2015).
12. Hong, M. *et al.* RNA sequencing: new technologies and applications in cancer research. *Journal of Hematology & Oncology* **13**, 166 (2020).
13. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
14. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
15. Linder, J. & Seelig, G. Fast activation maximization for molecular sequence design. *BMC Bioinformatics* **22**, 510 (2021).
16. Suga, K., Yamada, K. & Hamada, M. Deciphering the comprehensive relationship between 5′ UTR and 3′ UTR sequences with deep learning. 2025.05.17.654644 Preprint at https://doi.org/10.1101/2025.05.17.654644 (2025).
17. Leppek, K. *et al.* Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat Commun* **13**, 1536 (2022).
18. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. Preprint at https://doi.org/10.48550/arXiv.1703.01365 (2017).
19. Seo, J. J., Jung, S.-J., Yang, J., Choi, D.-E. & Kim, V. N. Functional viromic screens uncover regulatory RNA elements. *Cell* **186**, 3291-3306.e21 (2023).
20. Barreau, C., Paillard, L. & Osborne, H. B. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res* **33**, 7138–7150 (2005).
21. Tebo, J. *et al.* Heterogeneity in Control of mRNA Stability by AU-rich Elements *. *Journal of Biological Chemistry* **278**, 12085–12093 (2003).
22. Kudla, G., Lipinski, L., Caffin, F., Helwak, A. & Zylicz, M. High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *PLOS Biology* **4**, e180 (2006).

23. Trobaugh, D. W. *et al.* Cooperativity between the 3' untranslated region microRNA binding sites is critical for the virulence of eastern equine encephalitis virus. *PLoS Pathog* **15**, e1007867 (2019).
24. Trobaugh, D. W. *et al.* RNA viruses can hijack vertebrate microRNAs to suppress innate immunity. *Nature* **506**, 245–248 (2014).
25. Kim, S. *et al.* Innate immune responses against mRNA vaccine promote cellular immunity through IFN-β at the injection site. *Nat Commun* **15**, 7226 (2024).
26. Morf, J. *et al.* Cold-Inducible RNA-Binding Protein Modulates Circadian Gene Expression Posttranscriptionally. *Science* **338**, 379–383 (2012).
27. Liu, Y. *et al.* Cold-induced RNA-binding proteins regulate circadian gene expression by controlling alternative polyadenylation. *Sci Rep* **3**, 2054 (2013).
28. Corre, M. & Lebreton, A. Regulation of cold-inducible RNA-binding protein (CIRBP) in response to cellular stresses. *Biochimie* **217**, 3–9 (2024).
29. Zhao, W. *et al.* POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Research* **50**, D287–D294 (2022).
30. Faramin Lashkarian, M. *et al.* MicroRNA-122 in human cancers: from mechanistic to clinical perspectives. *Cancer Cell International* **23**, 29 (2023).
31. Bicknell, A. A. *et al.* Attenuating ribosome load improves protein output from mRNA by limiting translation-dependent mRNA decay. *Cell Rep* **43**, 114098 (2024).
32. Narula, A., Ellis, J., Taliaferro, J. M. & Rissland, O. S. Coding regions affect mRNA stability in human cells. *RNA* **25**, 1751–1764 (2019).
33. Forrest, M. E. *et al.* Codon and amino acid content are associated with mRNA stability in mammalian cells. *PLOS ONE* **15**, e0228730 (2020).
34. Presnyak, V. *et al.* Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **160**, 1111–1124 (2015).
35. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol* **23**, 245 (2022).
36. Wu, Q. *et al.* Translation affects mRNA stability in a codon-dependent manner in human cells. *eLife* **8**, e45396 (2019).

37. Mauger, D. M. *et al.* mRNA structure regulates protein expression through changes in functional half-life. *Proceedings of the National Academy of Sciences* **116**, 24075–24083 (2019).
38. Wu X, Bartel DP. Widespread influence of 3'-end structures on mammalian mRNA processing and stability. Cell. 2017;169:905–17.e11. 33. Van Etten J, Schagat TL, Hr
39. Zhu, Y., Zhu, L., Wang, X. & Jin, H. RNA-based therapeutics: an overview and prospectus. *Cell Death Dis* **13**, 1–15 (2022).
40. Gray, M. *et al.* Chikungunya virus time course infection of human macrophages reveals intracellular signaling pathways relevant to repurposed therapeutics. *PeerJ* **10**, e13090 (2022).
41. Bartholomeeusen, K. *et al.* Chikungunya fever. *Nat Rev Dis Primers* **9**, 1–21 (2023).
42. Lin, D. *et al.* Multiorgan proteomic analysis of infected animal models predict potential host factors for chikungunya virus. *MedComm (2020)* **6**, e70013 (2025).
43. Lee, A. J. & Ashkar, A. A. The Dual Nature of Type I and Type II Interferons. *Front. Immunol.* **9**, (2018).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
45. Myint, L., Avramopoulos, D. G., Goff, L. A. & Hansen, K. D. Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* **20**, 209 (2019).
46. Huang, L. *et al.* LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics* **35**, i295–i304 (2019).
47. Fornace, M. E. *et al.* NUPACK: Analysis and Design of Nucleic Acid Structures, Devices, and Systems. Preprint at https://doi.org/10.26434/chemrxiv-2022-xv98l (2022).
48. Smedley, D. *et al.* BioMart – biological queries made easy. *BMC Genomics* **10**, 22 (2009).
49. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
50. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Research* **43**, W39–W49 (2015).
51. Larralde, M. & Zeller, G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* **39**, btad214 (2023).

52. Iwano, N., Adachi, T., Aoki, K., Nakamura, Y. & Hamada, M. Generative aptamer discovery using RaptGen. *Nat Comput Sci* **2**, 378–386 (2022).
53. Rishik, S., Hirsch, P., Grandke, F., Fehlmann, T. & Keller, A. miRNATissueAtlas 2025: an update to the uniformly processed and annotated human and mouse non-coding RNA tissue atlas. *Nucleic Acids Research* **53**, D129–D137 (2025)