Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics

Kathrin Leppek,1,11, Gun Woo Byeon1,11, Wipapat Kladwang2,11, Hannah K. Wayment-Steele3,11, Craig H. Kerr1,11, Adele F. Xu1,1, Do Soon Kim2, Ved V. Topkar4, Christian Choe5, Daphna Rothschild1, Gerald C. Tiu1, Roger Wellington-Oguri6, Kotaro Fujii1, Eesha Sharma2, Andrew M. Watkins2, John J. Nicol6, Jonathan Romano6,7, Bojan Tunguz2,8, Fernando Diaz9, Hui Cai9, Pengbo Guo9, Jiewei Wu9, Fanyu Meng9, Shuai Shi9, Eterna Participants6, Philip R. Dormitzer9,10, Alicia Solórzano9, Maria Barna1 & Rhiju Das2,4,6✉

Therapeutic mRNAs and vaccines are being developed for a broad range of human diseases, including COVID-19. However, their optimization is hindered by mRNA instability and inefficient protein expression. Here, we describe design principles that overcome these barriers. We develop an RNA sequencing-based platform called PERSIST-seq to systematically delineate in-cell mRNA stability, ribosome load, as well as in-solution stability of a library of diverse mRNAs. We find that, surprisingly, in-cell stability is a greater driver of protein output than high ribosome load. We further introduce a method called In-line-seq, applied to thousands of diverse RNAs, that reveals sequence and structure-based rules for mitigating hydrolytic degradation. Our findings show that highly structured “superfolder” mRNAs can be designed to improve both stability and expression with further enhancement through pseudouridine nucleoside modification. Together, our study demonstrates simultaneous improvement of mRNA stability and protein expression and provides a computational-experimental platform for the enhancement of mRNA medicines.
Messenger RNA (mRNA) therapeutics hold the potential to transform modern medicine by providing a gene therapy platform with the capacity for rapid development and wide-scale deployment. Compared to recombinant proteins expressed in mammalian cell lines, manufacturing of mRNA is faster and more flexible because mRNA can be easily produced by in vitro transcription. Over the past decade, technological discoveries in the areas of mRNA modifications and delivery systems have rapidly advanced basic and clinical research in mRNA vaccines1–3. However, technical obstacles facing mRNA therapeutics are also apparent. For example, mRNA vaccines still suffer from decreased efficacy due to poor RNA stability in solution and in vivo and to limited expression of the payload mRNA; these are all pivotal issues that need to be carefully optimized for preclinical and clinical applications3–6.

The development of mRNAs redesigned for increased stability and expression could maximize the impact of producing, delivering, and administering therapeutic mRNAs. However, achieving such designs is hindered by a poor understanding of how the sequence and structure of an mRNA influence its expression and stability, both in solution and in cells. For example, it has typically been assumed that mRNAs with more stable secondary structure might have increased in-solution stability but would have lower in-cell protein output due to the increased difficulty of the cellular translation machinery to process through RNA structure7, but this has not been tested and some recent results suggest that there might not be such a tradeoff8,9. Thus, testing the expression efficiency of mRNAs that are predicted to have highly structured coding regions, which we term “superfolder” mRNAs7, would be advantageous for the overall performance of therapeutic RNA. Our poor understanding of these design rules is due in part to the historical difficulty of rapidly synthesizing full-length mRNAs with different untranslated regions (UTRs) and coding sequences (CDSs) which would enable high-throughput experimental approaches comparing their stability and expression.

Here, we overcome these technical hurdles and characterize hundreds of full-length reporter constructs that encode mRNA sequences with a variety of UTRs and CDSs. We present a massively parallel reporter assay termed Pooled Evaluation of mRNA in-solution Stability, and In-cell Stability and Translation RNA-seq (PERSIST-seq), which enables systematic determination of the effects of UTR, codon sequence and RNA structure on mRNA translation in human cells and on mRNA stability, both in cells and in solution. This represents the first screen of hundreds of mRNAs redesigned across their entire length. We further leverage the unique ability of the Eterna10 community, an online citizen science platform that enables participants to collectively solve RNA design puzzles, to devise solutions with high diversity in sequence and predicted structure. We integrate our datasets to develop a model that accurately predicts protein output for a given mRNA based on its ribosome load and in-cell stability. This model enables the identification of the best mRNA designs in our screen without the need to individually test all mRNAs for protein output. With the further aim of understanding the impact of mRNA structure on in-solution stability, we developed a high-throughput method termed In-line-seq to measure RNA degradation patterns arising from intrinsic in-line hydrolysis11. Nucleotide-resolution data on thousands of small, structured model RNAs from the Eterna platform revealed design parameters for reducing solution hydrolysis and enabled development of a regression model termed DegScore that enables in silico RNA sequence optimization for enhanced in-solution stability. Finally, we compare the effect of nucleoside modifications on mRNA performance, including results of pseudouridine (ψ) and its derivatives on in-solution stability. Ultimately, our findings culminate in the fully automated design of highly structured “superfolder” mRNAs containing 5′ and 3′ UTR elements and structure-optimized CDS regions that simultaneously confer high stability in solution and high protein expression in cells. Together, the combination of select UTRs for expression, DegScore-optimized CDS structure, and ψ modification provides a general technology that can be applied to stabilize and increase protein expression of candidate mRNA therapeutics. We envision that these design rules and our combinatorial mRNA optimization platform will be widely applicable to rapidly engineer future mRNA therapeutics that simultaneously optimize stability and potency.

Results
A combinatorial library for systematic discovery of mRNA design rules. In search of design rules for stable and high-expressing mRNAs, we aimed to characterize a large number of mRNA sequence designs with extensive variations in 5′ UTR, CDS, and 3′ UTR regions. We took advantage of recent accelerations in commercial gene synthesis and developed the massively parallel assay termed PERSIST-seq (Fig. 1a, Supplementary Fig. 1). In this method, mRNA variants can be assayed in parallel for translation efficiency in cells, for stability in cells, and for stability in solution. Full-length in vitro transcription (IVT) DNA templates were obtained through commercial gene synthesis services (Twist, Genscript, Codex). Each template incorporated three additional features: (1) a shared T7 promoter sequence for performing IVT, (2) barcodes in the 3′ UTR to enable multiplexing via inexpensive, short-read sequencing, and (3) a constant region at the 3′ end that enabled pooled PCR and reverse transcription (RT) reactions (Supplementary Fig. 1a–c). This design allowed one-pot amplification and analysis of the library using common flanking sequences. The whole library was in vitro transcribed and modified (3′ poly-A-tailing and 5′ m7G-capping) together. We have no indication for any bias in the efficiency of such RNA modifications for certain sequences or structures in the different mRNA when performed in a pool (Supplementary Fig. 1c). The mRNA library was then transfected into cells and quantified by barcode sequencing in a pool (Fig. 1b), enabling straightforward measurements of translation by polysome profiling and of mRNA degradation over time in cells or in solution.

Our mRNA library contained 233 different mRNA sequences in total (Supplementary Data 1). Hundred and twelve of these mRNAs contained varied 5′ and/or 3′ UTRs (Fig. 1a). While there have been many efforts to increase protein expression by attaching UTRs found in highly translated or stable mRNAs12–14, previous work tested and characterized only a handful of candidates at a time, either with a focus on individual functional UTR elements15 or on screening randomized short UTRs16–19. We decided to harness full-length naturally occurring UTRs to test for increased mRNA expression, including 5′ and 3′ UTR sequences from cellular and viral genomes, in our systematic analysis due to their potential to enhance or fine-tune mRNA translation or stability.

As examples of cellular sequences, we included short 5′ UTRs of cellular mRNAs corresponding to highly abundant proteins that have a high translation rate such as ribosomal proteins (RPs) (RPS25, RPL31, RPL38), or structural components such as tubulin beta-2 chain (Tubb2b) and actin (ActB), as well as human collagen, type I, alpha 2 (hCOL1A2)20. We further included regulatory 5′ UTR elements such as the 5′ terminal oligopyrimidine (TOP) motif from RPL18, which promotes translational activation downstream of mTOR21,22, as well as the Hoxa9 P4 RNA stem-loop which functions as a translation enhancer23. 5′ UTRs previously identified in translation efficiency screens such
as complement factor 3 (C3), cytochrome P450 2E1 (CYP2E1), and Apolipoprotein A-II (APOA2)\textsuperscript{13}, as well as plant rubisco components (RBCS3B, RBCS1A)\textsuperscript{24}, were also included. For 3′ UTR regions (total 22; size range 60–597 nt), we employed known stabilizing RNA structures such as the MALAT1 non-coding RNA 3′-stem-loop structure that resembles an expression and nuclear retention element (ENE) and engages a downstream A-rich tract in a triple helix structure\textsuperscript{25,26}, as well as known expression enhancing 3′ UTRs such as those from human hemoglobin subunit alpha 1 (HBA1)\textsuperscript{27} and cytochrome B-245 alpha chain (CYBA)\textsuperscript{28,29}.

Viruses have evolved a suite of compact regulatory elements for hijacking the host translation machinery to effectively promote translation and stability of their own mRNAs. For example,
internal ribosome entry sites (IRESs) can recruit ribosomes to initiate translation without the need for the full repertoire of eukaryotic initiation factors, whereas other elements in the 5′ and 3′ UTRs of viruses encode structures that facilitate long-range RNA-RNA interactions to enhance protein expression or mRNA stability. Therefore, several UTRs originating from viral genomes were included: the 5′ and 3′ UTR elements of the SARS-CoV-2 RNA genome\(^{30,31}\), along with variants described below; the Dengue virus (DEN2) 5′ and 3′ UTRs, which are thought to enhance viral protein expression\(^{32,33}\); 5′ and 3′ UTR elements from various tombusviruses (e.g., turnip crinkle virus (TCV)) which encode 3′ cap-independent translational enhancer RNA structures that recruit the translational machineries\(^{34,35}\); tobacco mosaic virus\(^{36}\) (TMV) and tobacco etch virus\(^{37}\) (TEV) 5′ leader sequences; a poxvirus poly(A) leader sequence that is proposed to facilitate translation\(^{38}\); and the 3′ UTRs of Sindbis virus (SINV) and the rabies virus glycoprotein, which increases viral RNA stability through recruitment of host proteins\(^{39,41}\).

As our main reference, we chose the 5′ and 3′ UTRs from \textit{human hemoglobin subunit beta} (\textit{hHBB}), which is one of the most efficiently expressed mammalian mRNAs and is commonly used in investigations of mRNA translation and stability\(^{42,43}\). Non-\textit{hHBB} UTRs are referred to here as “UTR variants.” To test these UTR variants, all reporter mRNAs encode the Nanoluc luciferase (Nluc) open reading frame (ORF) as its CDS region\(^{44,45}\). We decided to use Nluc because its short ORF of 621 nt allowed for synthesis and pooled amplification of full-length DNA templates with UTRs of up to 600 nt attached at each end. Employing Nluc further enabled precise quantitative readout for comparing translation efficiencies in follow-up experiments on individual mRNAs through ratiometric measurements including a firefly luciferase (Fluc) mRNA spike-in control in transfection experiments integrated with luciferase luminescence assays\(^{23}\). Additional mRNAs encoding for enhanced green fluorescent protein (eGFP) and a shorter candidate multi-epitope vaccine (MEV)\(^{7}\) were included as controls in some experiments, further discussed below.

To test the impact of CDS sequence and predicted CDS structure on mRNA stability and translation, we sought to maximize the diversity of CDS sequences and structures for model protein targets. Consequently, we asked participants from the Eterna massive open laboratory\(^{10}\) to design CDSs encoding a variety of model mRNAs, without specific optimization metrics, in a series of challenges (Fig. 1a). These puzzles included design challenges for eGFP, MEV, and Nluc (‘OpenVaccine: Design of eGFP and epitope mRNA molecules’ and ‘OpenVaccine: Lighting Round Design + Vote of Nanolucerase’). Later rounds of design challenges included degradation-specific metrics (AUP and DegScore, discussed later) within the game interface to guide optimization. We also included CDSs using several algorithmic approaches. First, we included sequences designed using commercially available algorithms to optimize codon adaptation index (CAI)\(^{16}\). Second, we designed sequences using a “GC-rich” approach, in which each codon is stochastically sampled from codons highest in GC content, based on a strategy developed by CureVac\(^{9}\). Third, we included CDSs designed using the LinearDesign algorithm\(^{47}\), which returns a deterministic minimal free energy solution that is weighted by codon optimality. Finally, we used the Ribotree Monte Carlo tree search method to optimize AUP for eGFP and to compare them to eGFP designs developed by Moderna\(^{7,8}\). These design methods yielded a total of 121 CDS variants in the library (Fig. 1a, Supplementary Data 1). To ensure useful cross comparison, \textit{hHBB} UTRs were used for each of these CDS variants. Together, all candidates from these diverse sources of UTR and CDS sequences were combined into a single library of 233 mRNA constructs.

**High dynamic range of translation driven by UTRs.** To assess translation efficiencies, PERSIST-seq uses transfection of mRNA pools into human cells (here HEK293T). The cell lysate then undergoes sucrose gradient fractionation, which separates mRNAs into actively translating and non-translating fractions that are analyzed by RT-PCR of barcode regions and Illumina sequencing. Actively translating mRNAs have a higher number of ribosomes associated with them and are found in polysomal fractions whereas non-translating or poorly translating mRNAs are present in the free mRNA fraction or are associated with 40S ribosomal subunits (Fig. 1c). After initial studies confirming differences in polysome loading of a highly translated endogenous mRNA, \textit{human ActB}, with that of a transfected control mRNA that has scrambled short UTR sequences\(^{44}\) (Supplementary Fig. 1d), we carried out PERSIST-seq to examine the polysome profiles of diverse constructs in the 233x-mRNA library (Supplementary Fig. 1f). In parallel, we perform PERSIST-seq to compare the in-cell and in-solution stability of the mRNA designs in the same library over time (Supplementary Fig. 1e). Importantly, we are quantifying relative differences across polysome fractions or time points, from which we calculate ribosome load (Supplementary Fig. 1f), or fit decay curves for half-lives (Supplementary Fig. 1g), respectively. Thus, any construct-specific biases in RT efficiencies are regressed out (i.e., only the relative ratios of the counts between time points, not the absolute read counts, are relevant for interpreting stability). We observed a wide variation in mRNA distribution across the fractions (here expressed as ribosome load, defined as the weighted sum of mRNA proportions multiplied by the ribosome number in a fraction) (equation in Fig. 1d). The largest variation in ribosome load was observed for the 5′ UTR variants group (Fig. 1d). These data suggested a strong potential for using different 5′ UTRs to tune the translational efficiency of target mRNAs—more than for any other region.

We categorized the origins of UTR sequences described above as “cellular”, “viral”, and “chimera” (modular UTR combinations). Overall, the mRNA designs with highest ribosome load...
were observed across 5′ UTRs of cellular as well as viral origins (Fig. 1e, Supplementary Data 1). These 5′ UTRs included: mouse COL1A2, Hoxa9 P4, Rpl18a TOP, plant RBCS1A; the poxvirus poly(A) leader sequence fused to a scrambled 5′ UTR sequence and also the 5′ UTRs of plant viruses TEV and TMV. The dengue virus 5′ and 3′ UTRs both individually increase ribosome loading, and combining them into one mRNA resulted in an additive effect (Fig. 1e, Supplementary Data 1). All of these sequences had a higher ribosome load (1.7–2.3) than the hHBB 5′ UTR (1.57), thereby identifying potential UTR design strategies to boost mRNA translational efficiency. Moreover, the chimeric fusion of the hHBB 5′ UTR with elements such as the TEV or 5′ TOP sequence in the same 5′ UTR increased polysome loading. We also found that human beta-actin (hACTB) does not perform well in terms of ribosome loading but the mouse beta-actin (mACTb) performs in the same range as the scrambled short 5′ UTR (scrUTR) or hHBB. Overall, our polysome screen successfully identified a wide range of 5′ UTR sequences that can be deployed to successfully promote translation within cells. Most surprising, in contrast to previous reports48–50 we found 5′ UTRs that are highly structured, such as the dengue virus (DEN2), can support efficient translation initiation. Our present studies do not permit us to more directly compare the effect of structured 5′ UTRs on translation beyond individual candidates.

We next asked if the highly structured 5′ UTRs used in our screen could be augmented to further improve ribosome loading. While the combined DEN2 3′ and 5′ UTRs are well loaded with ribosomes (Fig. 1e), the DEN2 5′ UTR alone did not perform as one of the top candidates 5′ UTRs. To this end, we chose the SARS-CoV-2 5′ UTR to further optimize as it is much more highly structured than DEN2 5′ UTR while still sustaining sufficient translation levels. We performed a detailed mutagenesis analysis of the structured 5′ UTR51–53 from SARS-CoV-2 genomic RNA50,51 as an example (Fig. 1f). We first observed that mutation of the uORF in the 5′ UTR (AUG mutated to UUG; CoV2-UUG) resulted in a higher ribosome load than the wild-type 5′ UTR (Fig. 1f, g). Then, to systematically determine the impact of each stem-loop on ribosome load, we introduced partial and full truncations of stem-loops (SL) 1–5 (SL1–5) into CoV2-UUG (Fig. 1f, Supplementary Data 2). In addition, we included larger truncations of combined deletions of adjacent stem-loops and introduced the Hoxa9 P4 stem-loop, a 35 nt element that recruits 40S ribosomal subunits to enhance translation23, in lieu of the uORF (Fig. 1f, g, Supplementary Data 2). Intriguingly, polysome profiling of the 5′ UTR SL mutant constructs revealed a wide range of overall improved ribosome load (Fig. 1g). Deleting the full 5′ half (dSL1-3) or the half (dSL4-5) of the 5′ UTR overall increased ribosome load. In particular, deletion of SL1 from the SARS-CoV-2 5′ UTR increased ribosome load from 1.23 (CoV-2) to 1.5 (CoV-2-UUG-dSL-1) which almost reaches the level of ribosome load of hHBB (1.57) (Fig. 1g). These results on the SARS-CoV-2 5′ UTR indicate that ribosome load can be fine-tuned through the modulation of distinct elements in structured viral 5′ UTRs, and that these effects can be read out through PERSIST-seq.

In-cell mRNA stability is a major predictor of total protein output. The total protein output from an mRNA depends not only on its in-cell translational rate but also on how long it remains intact inside cells. To assess the in-cell mRNA stability of the library of constructs in a pooled fashion, PERSIST-seq quantifies the fractions of mRNAs remaining at multiple time-points following transfection of the library into cells. To ensure recovery of intact full-length mRNAs rather than their degraded fragments, PERSIST-seq uses a two-step protocol, first generating amplicons covering the entire CDS regions of mRNAs through reverse transcription-PCR (RT-PCR) and then using primers flanking just the barcode region for a second PCR before short-read Illumina sequencing to count intact mRNAs at time points (Supplementary Fig. 1e).

For our library of 233 mRNAs, PERSIST-seq gives a dynamic range of in-cell half-lives, ranging from <5 h to over 15 h (Fig. 2a). We had originally expected that the sub-library comprised of varying 3′ UTR sequences would give the most variation in in-cell...
stability, since those mRNAs included diverse cis-regulatory elements that are known to recruit cytoplasmic factors to aid or prevent mRNA decay55–58. However, we instead observed the widest in-cell stability variation in the CDS and 5' UTR variant groups (Fig. 2a). We also noted a non-linear relationship between ribosome load and mRNA stability, where there is a positive association between a moderate increase in ribosome load and mRNA stability while the opposite is true for the highest ribosome load values (Fig. 2b, Supplementary Fig. 3). Examining this more closely, we find that metrics that separately look at polysome versus monosome loading better capture this trend: high polysome/monosome ratios predict lower stability ($r_s = -0.516$) while monosome/pre-80S ratios predict increased stability ($r_s = 0.467$). These findings identify an unexpected rule for the design of mRNA: increasing translation efficiency may counterintuitively decrease protein expression, when excessive ribosome load values. The figure illustrates the relationship between in-cell mRNA half-life and ribosome load, with distinct groups and metrics showing the trend.
polysome loading leads to decreased mRNA stability such that the total amount of protein produced over time actually becomes lower. To explore the implications of this apparent tradeoff, we sought to integrate both translation (ribosome load) and in-cell stability (Fig. 2c) into a simple quantitative model to understand and predict their relative impact on integrated protein expression (Fig. 2d, e)\(^59\). To this end, we used differential equations to describe biochemical kinetics of mRNA translation and mRNA/protein decay assuming first-order rate of translation and exponential decay for mRNAs and for proteins (Fig. 2d).

This expression enables the ranking of expected protein levels across the UTR variants in our mRNA library, as \(k_t\) and \(k_{m}\) are related to PERSIST-seq measurements of ribosome load and in-cell half-lives, respectively, and the other parameters are known (Fig. 2e). Overall, predicted protein expression per mole of transfected mRNA is modeled to be mostly driven by the in-cell mRNA half-life (Fig. 2c, top panel), which more closely trends with predicted protein expression (Fig. 2e) than ribosome load (Fig. 2c, lower panel). To test this model, we predicted and measured, using the luciferase readout, the protein levels observed at a given time (here 12 h post-transfection) when an equal mass of mRNA is transfected for each construct. The predicted and measured luciferase activities showed a correlation of \(r = 0.71\) (Fig. 2f), supporting the accuracy of our model. However, for short expression (early time points (6 h) after mRNA transfection) or for “above-threshold” protein half-lives, translation can become the dominant predictor. Indeed, the analysis of the correlations of ribosome load and in-cell mRNA half-life with protein expression at 6 and 12 h post transfection indicates that only at later time points (12 h), mRNA half-life is a driver and an approximate predictor of protein expression (Supplementary Fig. 4b, \(r_s = 0.41\)). Therefore, depending on the desired parameters (early burst of high protein expression or total protein output integrated over a longer time), the most optimal set of UTRs can vary. Taken together, our large-scale unbiased pooled measurement of translation and mRNA stability provides a platform to test large numbers of different mRNA sequence designs for desired protein expression and has allowed us to infer that in-cell mRNA stability is the dominant predictor of total protein output (Fig. 2c, e).

The degradation of RNA in solution is a major obstacle in the distribution of mRNA therapeutics to patients\(^6\). Thus, our final use of PERSIST-seq was to evaluate structure-based RNA design strategies to yield more stable RNAs in solution. Just as with the in-cell stability measurements, we measured the fraction of mRNA CDS regions that remain intact after degradation, taking advantage of the same RT-PCR to select for intact mRNAs, followed by a PCR amplifying the short barcode regions in the 3’ UTR (Fig. 1a, Supplementary Fig. 1a, e). To mimic the high effective pH and positively charged environment that can arise in lipid nanoparticles, protamine, and other formulations for mRNA therapeutics\(^8\), we used a high pH buffer containing magnesium (\(\text{Mg}^{2+}\)) to accelerate degradation (10 mM \(\text{MgCl}_2\), 50 mM Na-CHES, pH 10.0). Box hinges: 25% quantile, median, 75% quantile, respectively, from left to right. Whiskers: lower or upper hinge ±1.5 x interquartile range. Bar: 1.0 x range. Points: median

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-022-28776-w | www.nature.com/naturecommunications

---

**Fig. 2 In-cell RNA stability drives downstream protein expression levels.** **a** In-cell half-life of each mRNA design in HEK293T cells. Box hinges: 25% quantile, median, 75% quantile, respectively, from left to right. Whiskers: lower or upper hinge ±1.5 x interquartile range. **b** Higher polysome load correlates with decreased in-cell half-life. Correlation between in-cell half-life and mean ribosome load across the entire profile (left), monosome-to-free subunit ratio (center), or polysome-to-monosome ratio (right). **c** In-cell half-life and mean ribosome load for individual mRNA designs with varying UTRs. **d** Kinetic model for predicting protein expression from mRNA half-life and ribosome load. **P**(c) is protein quantity at time t; \(k_t\) is translation rate; and \(k_{m}\) and \(k_p\) are rates of mRNA and protein decay, respectively. **e** Protein expression predicted using the kinetic model in (d) on the basis of mRNA half-life and ribosome load. Predicted protein expression of each UTR variant; note closer similarity to in-cell half-life data than to ribosome load in (c). **f** Correlation of predicted protein expression and Nluc/Fluc activity at 12 h in HEK293T cells. Predicted protein expression is normalized by mRNA length (corresponding to mRNA load). Predicted protein expression of each UTR variant; note closer similarity to in-cell half-life data than to ribosome load in (c). **g** mRNA lifetimes are strongly dependent on mRNA length and designed structures, revealed by time courses of mRNA degradation under accelerated aging conditions (10 mM \(\text{MgCl}_2\), 50 mM Na-CHES, pH 10.0). Box hinges: 25% quantile, median, 75% quantile, respectively, from left to right. Whiskers: lower or upper hinge ±1.5 x interquartile range. Whiskers: lower or upper hinge ±1.5 x interquartile range. Points: median and 25th and 75th percentiles—interquartile range: IQR—and whiskers extend to maximum and minimum values. **i** Nucleotide DMS accessibility mapped onto structures from DMS-directed structure analysis. **ii** In-solution half-life of various mRNA design variants. mRNA lifetimes are strongly dependent on mRNA length and designed structures, revealed by time courses of mRNA degradation under accelerated aging conditions (10 mM \(\text{MgCl}_2\), 50 mM Na-CHES, pH 10.0). Box hinges: 25% quantile, median, 75% quantile, respectively, from left to right. Whiskers: lower or upper hinge ±1.5 x interquartile range. Whiskers: lower or upper hinge ±1.5 x interquartile range. Points: median and 25th and 75th percentiles—interquartile range: IQR—and whiskers extend to maximum and minimum values. **i** Nucleotide DMS accessibility mapped onto structures from DMS-directed structure prediction.
Eterna-guided In-line-seq yields additional design principles and DegScore predictor. The mRNA designs above varied in-solution stability based mainly on computational predictions of mRNA structure and the assumption that nucleotides that are not base paired in structure would be uniformly prone to hydrolytic degradation. We hypothesized that we might further improve in-solution stability through a deeper understanding of any specific sequence and structure features that lead to enhanced or suppressed hydrolysis in such unpaired regions. For example, base identities and local structural features such as the size and symmetry of apical loops and internal loops may play roles in determining in-solution mRNA degradation.

To test such effects and to potentially discover unknown ones, we challenged Eterna participants to generate a large and diverse set of RNA molecules featuring designed secondary structure motifs in a special challenge (‘OpenVaccine: Roll-your-own-structure’, RYOS). Limiting the lengths of these molecules to 68 nucleotides and soliciting unique 3′ barcode hairpins enabled massively parallel synthesis and characterization of thousands of RNA molecules for their structure and degradation profiles (Fig. 3a). In particular, we obtained single-nucleotide-resolution measurements of 3030 RNA fragments using In-line-seq, a uridine were particularly prone to degradation (Fig. 3c, Supplementary Data 4) and structure models guided by these data (Fig. 2i). Overall, the global assessment of in-solution RNA degradation using PERSIST-Seq reveals the effects of RNA length and structure on RNA half-life in solution.

Thus, independent of the nucleotide identity 5′ of the U, this bond is a hotspot for in-line nucleophilic attack. In addition, we noted rules for hydrolytic degradation that depended on the type of RNA structural loop in which a nucleotide appears. We visualized sequence dependence within triloops by aggregating all triloops by sequence position at each position in the triloop (n > 5). Each visualization in Fig. 3c represents the median degradation measure (n > 5). A particularly salient characteristic was suppressed hydrolysis in symmetric internal loops compared to asymmetric internal loops (Fig. 3c). To distill these observations into a predictive model, we trained a windowed ridge regression model called ‘DegScore’ based on these In-line-seq data (Fig. 3d; see Methods) which quantitatively captured features like the increased hydrolysis rates at linkages leading to 3′ U (Fig. 3d). The DegScore regression coefficients with the largest magnitude corresponded to the identity of the nucleotide 3′ of a linkage (Fig. 3d). Of these coefficients, G and C were the most favorable (least hydrolysis) to have 3′ of the linkage, followed by A, and 3′ U was most detrimental to degradation, matching our prior observations.

To test the accuracy of the DegScore metric derived from In-line-seq data, we made predictions of in-solution half-lives for the mRNAs measured in the PERSIST-seq experiments, which were carried out independently (Fig. 1b, Supplementary Fig. 2g). For the Nluc CDS variants that showed the widest variance in in-solution half-lives, we observed a strong correlation of DegScore predictions to the in-solution half-lives (Spearman R = −0.66, p < 0.0001). Strikingly, the accuracy of DegScore outperformed the accuracy of two other metrics described before to parametrize RNA structure but which do not take into account sequence or structure-motif dependences of RNA hydrolysis: the free energy of the predicted MFE secondary structure, the metric used in several design algorithms including LinearDesign (dG_{MFE}; R = −0.50), and the predicted summed unpaired probability of the RNA structure ensemble (SUP; R = −0.62) (Fig. 3e). The difference in correlation coefficients were not significant with p < 0.05 as evaluated by a two-sided significance test on dependent overlapping correlations (see Methods); more experimental studies are likely needed to achieve significance for these and future metrics. Beyond the Nluc CDS variants, we confirmed that DegScore gave the highest accuracy in predicting in-solution stability when evaluated over all the measured mRNAs, including low and high structure eGFP mRNAs from Moderna researchers, Eterna, and Ribotree (Supplementary Table 1 and Supplementary Fig. 8a, b).

Pseudouridine stabilizes RNA in solution. Given that the linkages with 3′ U were particularly sensitive to degradation, we hypothesized that the base chemistry of U may be directly linked to the degradative capacity of this nucleoside and sought to test whether chemical alternatives to U might alleviate degradation. In particular, we focused on ψ and m1ψ, since these substitutions for U have been widely adopted for mRNA therapeutics and vaccines due to improved in-cell translation and to better control of innate immune responses through avoidance of recognition by cellular toll-like receptors (TLR7 and TLR8)70, RIG-I, and PKR71–73. While ψ and derivatives have been reported to stabilize mRNAs against decay in cells through avoidance of recognition by cellular toll-like receptors (TLR7 and TLR8)70, RIG-I, and PKR71–73. While ψ and derivatives have been reported to stabilize mRNAs against decay in cells, the effects of these modifications on RNA stability in solution have not been reported. We selected RNA sequences from the Eterna RYOS challenge (Fig. 3a) that were designed to contain U-rich loops or U-rich unstructured regions, resynthesized these RNAs with standard nucleotides or with ψ or m1ψ substituted for U, and measured their in-line degradation over time via capillary electrophoresis. We observed that substitution of U with either ψ or m1ψ led to
Fig. 3 High-throughput in-line hydrolysis uncovers principles of in-solution RNA degradation. a Eterna participants were asked to design 68-nucleotide RNA fragments maximizing sequence and structure diversity. In total, 3030 constructs were characterized and probed using high-throughput in-line degradation (In-line-seq). b Nucleotide-resolution degradation of 2165 68-nt RNA sequences (filtered for signal quality), probed by In-line-seq, sorted by hierarchical clustering on degradation profile. c Sequences span a diverse set of secondary structure motifs, revealing patterns in degradation based on both sequence (i.e., linkages ending at 3’ uridine are particularly reactive) and structure (symmetric internal loops, circled, have suppressed hydrolytic degradation compared to asymmetric internal loops). d The ridge regression model “DegScore” was trained to predict per-nucleotide degradation from sequence and loop assignment information. Coefficients with the largest magnitude corresponded to sequence identity immediately after the link, with U being most disfavored. e DegScore showed improved predictive power on mRNAs over two other metrics previously posited to predict RNA stability. Half-life in-solution mRNA half-life, calculated from degradation coefficients of the exponential decay fit on time course data in PERSIST-seq. Errors are standard deviations estimated by exponential fits to bootstrapped data. dG(MFE): Free energy of minimum free energy structure, calculated in RNAfold v2.4.14. Sum p(unpaired): Sum of unpaired probability, calculated in RNAfold v2.4.14. N1-methylpseudouridine (m1Ψ) modifications stabilize selected short RNAs at U nucleotides in both loop motifs and fully unstructured RNAs. f Introduction of pseudouridine (Ψ) and N1-methylpseudouridine (m1Ψ) modifications stabilizes selected short RNAs at U nucleotides in both loop motifs and fully unstructured RNAs. g Capillary electrophoresis characterization of fragmentation time courses of Nluc mRNA molecules designed with extensive structure (LinearDesign-1) and relatively less structure (Yellowstone), synthesized with standard nucleotides and with Ψ modifications. The full-length mRNA band is indicated with a red asterisk. The full-length mRNA band is indicated with a red asterisk. The full-length mRNA band is indicated with a red asterisk. h Exponential fits of capillary electrophoresis measurements of intact RNA over ten time points confirm significant differences in between-in-solution lifetimes of LinearDesign-1 and Yellowstone Nluc mRNAs. Inset: Calculated half-lives, mRNA half-life data are presented as mean values ± SD, as estimated from one biological replicate via bootstrapped exponential. Asterisks correspond to two-sided significance tests with ****p < 0.0001, ***p < 0.001, **p < 0.01.
suppression of in-line hydrolysis at the substituted residues, presumably through changed nucleophility at the site of substitution (example constructs in Fig. 3f, statistics over all constructs probed by capillary electrophoresis in Supplementary Fig. 5). We also observed suppression of in-line hydrolysis at nucleotides 1 to 2 positions 5′ of the substitution (Fig. 3f), possibly due to locally enhanced base stacking.28 Structure-mapping data by SHAPE and DMS profiling confirmed that ψ and m1ψ substitutions did not change the chemical reactivity of the RNAs outside the substituted positions, consistent with no change in global secondary structure; the suppression of in-line hydrolysis appears to be due to local chemical or structural effects at the site of substitution. However, more detailed measurements are needed to more holistically understand the effect of ψ on secondary structure.27,28

As a further test of the stabilizing effect of nucleoside modification, we prepared six constructs from the 233x-mRNA library with U or ψ (Fig. 1), including the LinearDesign-1 and Yellowstone RNAs (Fig. 2). In-solution mRNA half-lives were measured using capillary electrophoresis to evaluate the fraction of intact mRNA over time (Fig. 3g, h, pairwise significance comparisons between constructs in Fig. 3h inset). Consistent with our in-line hydrolysis data on small RYOS RNAs, we observed a 1.2–2.7-fold stabilization for these longer Nluc-encoding mRNAs when U was substituted with ψ (Fig. 3h, Supplementary Fig. 6c). This finding indicated that beyond redesigning RNA sequences to adopt stable structure, in-solution RNA stability can be further improved by incorporating modified U nucleosides.

Additive effects of UTR, CDS, and ψ improvements. Thus far we found that highly structured 5′ and 3′ UTRs can support efficient protein synthesis and based on our RNA degradation predictor DegScore and Eterna-derived designs, we saw that the highly structured CDSs can strongly affect in-solution mRNA stability and protein synthesis (Figs. 2, 3). We next investigated if selected UTRs and CDSs in combination with nucleoside modifications may achieve stable and highly translated mRNAs that profit from additive effects of individual improvements. (Fig. 4, Supplementary Fig. 9).

We first determined the combined contribution of different CDS designs and UTRs on mRNA stability and protein expression. It is important to distinguish ribosome load, the weighted sum of mRNA proportions multiplied by the ribosome number in a fraction, from protein output of an mRNA. Protein expression of an mRNA is not directly inferable from its ribosome load, thus luciferase-based expression analysis needs to confirm a positive effect on protein production based on improved mRNA design. We chose six CDS designs and three UTR combinations from our screen (Figs. 1a, 4a, Supplementary Fig. 9). The selected CDS designs were from diverse origins that support a range of in-solution half-lives (from 0.69- to 1.8-fold relative to Nluc start; Fig. 2g). The combined 5′ and 3′ UTRs were individually predicted and/or confirmed (Fig. 2f) to facilitate the highest protein expression in our library (Fig. 2e, f); our standard hHBB 5′ and 3′ UTRs; a SARS-CoV-2 5′ UTR dsl-3 variant paired with the dengue virus 3′ UTR, predicted to have high translational efficiency but shorter in-cell half-life due to increased length; and C3 5′ UTR paired with a SINV U-rich element 3′ UTR predicted to have good translational efficiency with a short length (Fig. 2e, f, Supplementary Data 1). In terms of in-solution stability, we expected the mRNAs with the longer UTRs to have reduced half-life across all CDSs, which was confirmed experimentally (Supplementary Fig. 10).

To evaluate protein expression, the 18 mRNAs were individually transfected and luciferase activity was measured after 6 h (translation rate before significant mRNA decay) and 24 h (total protein output after mRNA decay) (Fig. 4a). After 6 h, two CDS variants (LinearDesign-1 and —423.7) achieved similarly high protein levels compared to Nluc start, when combined with hHBB 5′ and 3′ UTRs. The LinearDesign-1 result was striking given its monosaccharide-concentrated polysome profile in PERSIST-seq (Fig. 1e) and the expectation that high mRNA structure should adversely impact the cellular translational apparatus. The result was nevertheless consistent with our model that enhanced in-cell half-life of a structured mRNA may compensate for low translation efficiency (Fig. 2c–e). Indeed, by 24 h, LinearDesign-1 CDS mRNA displayed a two-fold increase in luciferase yield compared to Nluc start (Fig. 4a). LinearDesign-1 CDS mRNAs also exhibited a particularly long in-solution half-life (Supplementary Fig. 10). Among the UTR combinations and CDS designs tested, most demonstrated lower overall luciferase activity than Nluc start with hHBB UTRs (Fig. 4a). One exception was the CoV-2-UG-dsl-3/DEN2 UTR combination, chosen based on its high ribosome load (Fig. 2d, e), which was able to support levels of protein synthesis at 6 h nearly as high as the hHBB UTR for the LinearDesign-1 CDS; however, this expression was lowered by 24 h (Fig. 4a). This finding is consistent with faster mRNA decay compared to hHBB UTRs and our results supporting that in-cell mRNA stability is a primary driver of protein output (Fig. 2e, e).

Given the improvements in in-cell stability we observed for ψ-modified mRNAs, we further tested the effect of ψ on in-cell stability and protein levels achieved from Nluc start vs. LinearDesign-1 CDS (Fig. 4a). As expected, preparation of mRNAs with ψ led to increased in-solution stability across both these CDSs, independent of UTRs (Supplementary Fig. 10). In terms of protein expression, we observed variable effects on overall luciferase activity by different UTR combinations with fixed CDSs compared to unmodified mRNAs after both 6 and 24 h of expression. Both Nluc start and LinearDesign-1 CDSs with hHBB UTRs maintained high protein expression at 6 and 24 h (Fig. 4a) indicating that, despite ψ modification and highly structured CDSs, translation is sustained. Overall, these results demonstrated the importance of mRNA stability to protein output, and suggested that the hHBB UTRs, highly structured CDSs, and use of ψ is preferred for increasing in-solution stability and protein output.

Integration of all design rules leads to mRNAs with both high in-solution stability and high protein output. For our final experiments, we sought to test whether further optimization, especially of the CDS, might allow both enhanced in-solution mRNA stability and in-cell protein expression. A simple calculation of the number of synonymous mRNAs for the Nluc CDS (621 nt) using a human codon table reveals that there are 1.6 × 1010 potential sequence combinations. The hypothesis that mRNA stability may be increased by changing the CDS sequence has been explored theoretically through algorithms designed to optimize the predicted MFE of a CDS, as well a biophysical model for hydrolysis that predicted a minimum two-fold decrease in hydrolysis could be achievable; but these predictions for increases in stability have not been experimentally tested. To test these algorithms and aforementioned models in the context of our model mRNAs, we collected a variety of CDS designs to compare to the Nluc start and Yellowstone mRNAs, including (1) the default output of available mRNA design algorithms, including those provided by Genewiz, Twist, and IDT websites and others that may enhance mRNA structure (LinearDesign and use of GC-rich codons), (2) highly structured constructs that were rationally designed through an Eterna competition
('OpenVaccine: Focus on the NanoLuciferase mRNA'), and (3) output of the automated mRNA structure design tool Ribotree, a stochastic optimization algorithm that can start from different seed sequences (random or LinearDesign) and improve in-solution mRNA half-life as guided by different predictors (AUP, DegScore) (Fig. 4b). From these approaches we generated 24 different CDS designs for Nluc mRNA and, based on our data above, we appended the hHBB 5' and 3' UTRs as constant regions to mediate high rates of translation initiation and used ψ due to its enhancement of in-solution mRNA stability (Fig. 3f, g). These 24 different CDS designs displayed a wide variety of predicted structural diversity even for a short ORF such as Nluc (see Supplementary Fig. 11).

Each mRNA design was subjected to accelerated degradation and capillary electrophoresis to measure the in-solution half-life of each individual CDS. Compared to the Nluc start, Eterna-originated designs exhibited up to 2.9-fold higher in-solution half-life (see, e.g., RLT-10) while designs from commercial methods showed significantly lower enhancement. The in-solution half-life for each design is shown in Supplementary Table 1.

Supplementary Fig. 11 provides a comparison of the predicted and measured in-solution half-life for each design. The in-solution half-life was measured using capillary electrophoresis and is expressed as the percentage of the initial mRNA concentration remaining after 6 hours. The data is shown in a box plot, with a higher percentage indicating a longer in-solution half-life.

The in-solution degradation & functional assays (Fig. 4b) illustrate the impact of different predictors (AUP, DegScore) and the use of different algorithms (Ribotree, LinearDesign) on the in-solution half-life of mRNA. The figure shows the normalized Nluc/Fluc (A.U.) for each design, with a higher value indicating better performance. The data is plotted as a scatter plot, with each point representing a different CDS design.

In conclusion, the use of different design approaches and predictors can significantly impact the in-solution half-life of mRNA. The results demonstrate the potential for optimizing mRNA design to enhance stability, which is crucial for the successful delivery and expression of therapeutic genes.
algorithms or optimized GC-content exhibited similar in-solution half-lives (Fig. 4c, significance values in Supplementary Data 5). Designs from the LinearDesign server and modifications to these designs from both Eterna participants (e.g., AB_rhiju_mod3) and the RiboTree algorithm produced constructs with increased half-life up to 2.3-fold over Nluc (triangles in Fig. 4c). When individual mRNA designs were assayed for protein expression, despite overall longer in-solution half-lives, at 6 h, both Eterna- and RiboTree-derived designs had similar or lower luciferase activities than Nluc start, whereas most vendor-derived and GC-rich designs had slightly higher activities (Fig. 4c). At 24 h, the trend of lower luciferase activity remained for 6 of 8 RiboTree-derived designs tested (Fig. 4c, green). However, in contrast, 6 of 8 RiboTree-optimized mRNA designs demonstrated higher luciferase activities than Nluc start at 24 h (Fig. 4c, blue). The sequence output by RiboTree starting from a LinearDesign CDS solution, with optimization guided by DegScore and flanking hHBB UTR sequences, yielded an mRNA that was both highly stable in solution \( t_{1/2} = 2.3 \) relative to Nluc start and exhibited high levels of protein expression (1.7-fold increase, relative to Nluc start) (RiboTree_LinearDesign_degscoreall_1; Fig. 4c). This simultaneous increase in in-solution stability with improved and sustained in-cell protein expression provided a strong demonstration of the impact of rational design for mRNA.

To gain insight into what led to success for this RiboTree_LinearDesign_degscoreall_1 sequence, we examined the specific regions RiboTree modified from the starting sequence as it computationally minimized the DegScore metric from a starting LinearDesign server solution (Fig. 4d, with further comparison in Supplementary Fig. 11). These computational modifications are characterized by reducing the presence of Us in loops, and shifting local base-pairing to minimize the overall size of loops, even if such shifts result in additional smaller loops. These modifications are consistent with mitigating hydrolysis as modeled in the DegScore predictor (Fig. 3d). Taken together, these data suggest that by reducing the overall presence of Us in loops and reducing the number of hairpins to generate a “linear” highly double-stranded mRNA can lead to enhanced mRNA stability and protein expression.

Correlating data from these final 12 Nluc RNAs provided additional insight into the biophysical features impacting mRNA performance (Fig. 4, Supplementary Fig. 12). Most notably, DegScore (adjusted to account for reduced PSU reactivity, see Methods) correlated strongly with measured in-solution half-life \( (R = -0.64, p < 0.001) \), and moderately with 24-h protein expression \( (R = -0.33, p = 0.12) \), yet was uncorrelated with 6-h protein expression (Fig. 4e). However, 6-h protein expression correlated strongly with the predicted number of hairpins \( (R = 0.59, p < 0.001) \) and the “Maximum Ladder Distance \( 78^\circ \)” or the maximum helix path length \( (R = -0.80, p < 0.001) \) (Supplementary Fig. 12c, Supplementary Data 6). The observations suggest that resistance to RNA hydrolysis, as predicted by DegScore and quantified by in-solution half-life, is important for longer protein expression, but that other RNA sequence and structural features govern protein expression at shorter timescales. For instance, longer or more branched double-stranded RNA stems may potentially hinder the ability of the ribosome to unwind RNA secondary structure.
Fig. 5 Stability and cellular expression of selected highly structured RNA designs in solution and formulated with polyplex. a Schematic for testing the synergy between RNA modifications and mRNA design rules on downstream stability and protein output. mRNAs were in vitro synthesized with or without ψ and subjected to degradation conditions. Samples were collected overtime and the RNA was purified before being transfected into HEK293T cells. Luciferase activity was measured 24 h after transfection. b Luciferase activity of the reference Nluc sequence and DegScore-optimized CDS with or without ψ after being subjected to in-solution degradation. mRNA half-lives (t1/2), per construct are given in hours (hrs). Plotted on y-axis are the geometric mean of Nluc/Fluc reporter activity ratios normalized to time zero. Error bars indicate geometric standard deviation. n = 4 biologically independent samples. c Schematic for testing the effect of RNA formulation on downstream stability and protein output from selected RNA designs. mRNAs were in vitro synthesized, formulated with polyplex (PLX), and subjected to degradation conditions and/or expression analysis. Samples were collected over time and the formulated RNA was added to HEK293T cells. d In vitro stability of RNAs formulated with polyplex over 14 days at 5 °C. RNA half-lives were calculated based on the degradation slopes: Nluc start (reference) (14 days), Geneviz_1 (30 days), BugacMan’s_Lost_LD+finetuning_mod_Deg-2-ed (58 days), RLT-10 (69 days) and Ribotree_LinearDesign_degscoreall_1 (46 days). Results correspond to technical duplicates. e Expression of Nluc from HEK293T cells transfected with selected RNA designs formulated with polyplex. Expression was measured by fluorescence after the RNAs were formulated with polyplex, incubated at 5 °C in degradation conditions for 0 and 14 days, and then added to the medium of the cultured cells. Results correspond to technical replicates; normalized Nluc/Fluc activity ± SD. n = 3; ns not significant. *p ≤ 0.05 was considered significant (two-tailed unpaired Student’s t-test; ns: p > 0.05; *p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001; ****p ≤ 0.0001).

Discussion

mRNA-based therapeutics are transformative in the way in which human disease is treated, such as infectious disease (e.g., the timely COVID-19 mRNA vaccines)79−82. However, there still remain several major hurdles for mRNA-based therapeutics to be effective, many of which are directly linked to the intrinsic features of mRNA molecules. RNA is inherently unstable due to its 2‘-hydroxyl group, and degradation of the mRNA in solution by in-line nucleophilic attack, potentially exacerbated by lipid formulations, poses a major challenge. Moreover, once delivered into patient cells, the candidate mRNA must outcompete other cytoplasmic mRNAs for the translation machinery and avoid the cellular mRNA degradation machinery to express maximum amounts of the desired encoded protein. Our study presents advances in tackling this challenge, building on two RNA-seq-based technologies.

Our first set of experimental findings derive from the integrated PERSIST-seq technology, which enables parallel evaluation of the effects of UTR and CDS sequence and structure on in-cell mRNA translation efficiency, in-cell mRNA stability, and in-solution stability. This technology identifies previously unknown tradeoffs and additive effects between different tunable aspects of mRNA performance and their interdependencies, we find that in-cell mRNA stability may be a greater driver of protein output than high ribosome load. This might particularly be important when proteins need to be expressed from a single dose of mRNAs for long periods of time. Specifically, mRNAs associated with the heaviest polyosomes, thought to be linked to the highest translation efficiency, were found to be less stable: there may be a ’sweet spot’ of ribosome loading and translation efficiency to achieve maximal total protein output. This effect may derive from overcrowding of ribosomes on mRNA coding regions due to efficient and rapid initiation, which can induce ribosome queuing and steric collisions that have recently been found to lead to translation-dependent mRNA decay83,84. In turn, mRNAs may be
more stable with an optimal number of ribosomes loaded for efficient protein synthesis. A potential explanation for a highly structured CDS still being well translated could be the strong intrinsic helicase activity of the ribosome while translocating along mRNAs, which can unwind stable CDS structure in a two-step mechanism inside the elongating ribosome, without detriment to protein output. We also cannot rule out the potential contribution of cellular RNA helicases that may play a role in restructuring or unwinding structured CDS. Overall, we observed a wide range of UTR-dependent translation efficiencies in PERSIST-seq which could be further fine-tuned with specific UTR sequence or CDS structure alterations. We note that for applications with very short mRNAs (as used for MEVs, and represented in our MEV sequences) or much longer mRNAs (as are needed for antigens like the SARS-CoV-2 Spike protein), the best UTRs and the sweet spot for optimal translation efficiency may be different. In this respect, certain UTR combinations outperform hHBB in translation efficiency and include multiple cellular 5' UTRs as well as, unexpectedly, the denegue virus 5' and 3' UTRs, which both individually increase ribosome loading and combining them in one mRNA resulted in an additive effect. As illustrated with the SARS-CoV-2 5' UTR, selective translational enhancers can be further identified through careful mutagenesis and deletion strategies aimed at narrowing selective regulatory regions within viral leader sequences. We also note that our experiments were mostly limited to HEK293T as model human cells; the best UTRs for high protein output are likely to be cell-type dependent and therefore will depend on the application. Our library of UTRs and the PERSIST-seq technology should be well-suited for discovering and leveraging enhanced UTRs for future applications with different protein targets and cell types.

Our second set of experimental improvements relate to mRNA stability in aqueous solution. PERSIST-seq confirms that extremely highly structured mRNAs can exhibit more than double the in-solution half-lives of conventionally designed mRNAs, with strong implications for improved storage in solution. Further design insights came from a newly developed method called In-line-seq for high-throughput in-line probing, applied to thousands of diverse E5RNA-derived short RNAs. The In-line-seq results reveal a number of structural and simple sequence rules for mitigating in-solution RNA hydrolysis: a key determinant of in-solution RNA degradation is the presence of uridine and that RNA linkages 5' of a uridine residue are particularly susceptible to degradation, which can be alleviated through the inclusion of ψ or mψ. Synthesis of mRNAs with modifications—already in wide use for crucially mitigating innate immune response and translational shutdown by mRNA therapeutics—is therefore a simple additional improvement to achieve greater in-solution stability while sustaining protein expression.

Further leveraging our in-line hydrolysis data, we developed DegScore, a model for hydrolytic degradation that was independently validated on PERSIST-seq data and which enables in silico optimization of any RNA sequence. By combining optimal UTRs, DegScore optimization with RiboTree, and ψ modification, we are able to achieve high mRNA stability and improved protein expression. We note that DegScore was trained on degradation from unmodified nucleotides that does not account for our observed stabilization via ψ. Thus, future studies training similar models on degradation data from ψ and other nucleoside modifications may result in improvements to CDS stabilization via algorithmic design.

Ultimately, by applying DegScore, pseudouridine, and our insights from PERSIST- and In-line-seq we designed an mRNA that demonstrated a ~1.7-fold increase in protein output and ~2.5-fold increase in in-solution half-life. While it is difficult to directly assess the exact contribution of each mRNA design element given the confounding variable that each element presents, our results suggest that a given 5' or 3' UTR has a greater influence on ribosome load, while the inclusion of a structured CDS greatly influences the stability of the mRNA both in-cell and in-solution. Thus, it is likely the combination of both of these that allowed for the design of a highly translated and stable mRNA.

We also note that our study has focused on characterizing degradation of non-formulated mRNAs in order to understand limitations on stability imposed by the fundamental biophysical and biochemical properties of RNA; such non-formulated mRNAs also appear optimal for certain applications including personalized cancer mRNA vaccines. For applications that benefit from mRNA formulated in lipid nanoparticles or other carriers, we expect future studies applying PERSIST-seq and In-line-seq to formulated mRNA libraries to reveal additional insights. Lastly, it has been proposed that highly structured mRNAs may retain their structure and in-solution stability under temperature shifts, mutations, and changes in UTRs, motivating the term “superfolder” mRNAs; it will be interesting to test these predictions through further PERSIST-seq studies. These results demonstrate a modular and flexible platform that is applicable to potentially any protein target of choice and can accelerate the design of overall improved mRNA therapeutics solutions.

Overall, we report an mRNA design methodology that can enhance mRNA stability in solution while sustaining or even increasing protein expression inside cells. There does not have to be a tradeoff between mRNA structure, stability, and protein output. Looking ahead, our computational-experimental platform can be rapidly developed to customize highly structured mRNAs for target proteins. As mRNA-based medicines are explored for a wide range of human diseases including cancer therapies, we hope that these insights and methods can help these medicines become more effective, manufactured at a lower cost per patient, and more accessible and widely distributed to alleviate disease.

**Methods**

**In vitro transcription of reporter mRNAs.** The preparation of mRNAs was based on in vitro transcription from DNA templates. DNA templates were amplified by PCR using AccuPrime Pfx (Life Technologies, 12344024) and purified using the QIAQuick PCR Purification Kit (Qiagen). For the hHBB control mRNA, the DNA template was amplified with forward (5' - GATGAGTCGAGAGAAATTC-3') and reverse (3' - ATTAAGGTAGAGCAAATGAC-5') primers. The forward primer binds to the T7 RNA polymerase promoter common in DNA template for all mRNA designs; the reverse primer is complementary to a common "const3" region at the end of all tested mRNA 5' UTRs. For the IVT template pool, individual DNA templates were pooled for a pool of hundreds of constructs at an equimolar concentration and are amplified with outer primers in a pooled format. For the pooled template, 1 µL of each construct (~20 ng/µL stock concentration) was pooled to be used as the PCR template. The Pfx PCR contained the following: 2.5 µL 10x Pfx buffer, 0.25 µL forward primer (100 µM), 0.25 µL reverse primer (100 µM), 0.75 µL DMSO (NEB), 0.25 µL Pfx Polymerase (Thermo), 20.5 water, and 0.5 µL template DNA (Monarch PCR & DNA Cleanup Kit (NEB, T1030L). The source of the 3′UTR was the hHBB- Fluc starting CDS ("ncut start") is derived from the pCDS31.5-UTR-3xHA-Nluc plasmid encoding the HA-tagged Nanoluc CDS. Individual template DNA or the 233-mRNA library was amplified from linear DNA synthesized on a BioXP 3200 system (Codox DNA) or by Twist Bioscience, using the fixed forward "T7_F_28nt" and reverse (const3) primer. The forward primer binds to the T7 RNA polymerase promoter common in DNA template for all mRNA designs; the reverse primer is complementary to a common "const3" region at the end of all tested mRNA 5' UTRs. For the IVT template pool, individual DNA templates were pooled for a pool of hundreds of constructs at an equimolar concentration and are amplified with outer primers in a pooled format. For the pooled template, 1 µL of each construct (~20 ng/µL stock concentration) was pooled to be used as the PCR template. The Pfx PCR contained the following: 2.5 µL 10x Pfx buffer, 0.25 µL forward primer (100 µM), 0.25 µL reverse primer (100 µM), 0.75 µL DMSO (NEB), 0.25 µL Pfx Polymerase (Thermo), 20.5 water, and 0.5 µL template DNA (Monarch PCR & DNA Cleanup Kit (NEB, T1030L). For the hHBB-Fluc control mRNA, the DNA template was amplified from the pG3-HB plasmid using the primers KLS88/KLS89 which were A 20 µL transcription reaction contained maximum 5 µg linear DNA template, 4 mM of each NTP (Ambion), 2 µL/200 U MEGAscript T7 RNA polymerase (Ambion) and 1x T7 MEGAscript Transcription Buffer (Ambion). After a total incubation for 3 h at 37 °C, the DNA was digested by addition of 1 µL of Turbo DNase (Ambion, AM2218) for 15 min at 37 °C. For pseudouridylated mRNAs, pseudouridine triphosphate (Trilink Biotechnologies, N1019-5) was substituted for uridine triphosphate at an
equivalent concentration. mRNA was purified using MegaClear columns (Thermo Scientific, Ambion, AM1908). A 20 µl reaction usually yielded 100–150 µg of RNA. For the preparation of HEK293T cells, m7G-capped and polyadenylated mRNAs were generated as follows. In vitro transcribed mRNA was then mG-capped and polyadenylated using the ScriptCap m7G Capping System (CellScript, C-SCC6025) and A-Plus Poly(A) Polymerase Tailing Kit (CellScript, C-PAF5104H), respectively, according to the manufacturer’s instruction with the following modifications. Aliquots of 30 ng of each RNA were processed in parallel, diluted to 34.25 µL in water and heated for 5 min at 65 °C to denature and placed on ice. The 50 µL capping reaction containing 5 µL 10x ScriptCap buffer (CellScript), 5 µL 10 mM GTP (CellScript), 2.5 µL 2 mM S-adenosyl-methionine (SAM, 200 µM stock, CellScript), 1.25 µL ScriptInib RNase Inhibitor (CellScript), and 2 µL Capping enzyme (20 U, CellScript, 10 µL/5 µL). For the capping step, the 37 °C incubation was performed for 1 h and the capped RNA was placed on ice. Polyadenylation was performed from the resulting RNAs without purification in between. The poly(A) reaction contained 30 µg of capped mRNA in 50 µL 6.6 µL 10x A-Plus poly(A) tailing buffer (CellScript), 6.6 µL 10 mM ATP (CellScript), 0.3 µL ScriptTreated RNAS Inhibitor (CellScript), and 2.5 µL A-Plus Poly(A) Polymerase (10 U, 4 U/5 µL, CellScript) in a total reaction volume of 66 µL. We aimed to add a 150 nt-long poly(A)-tail for which we incubated the capped mRNA for 30 min at 37 °C with 10 U of poly(A), after which the reaction was placed on ice. The mRNA was again purified using MegaClear columns and mRNA concentration was determined on a Nanodrop 2000 (Thermo Fisher). This usually yields 30–40 µg of capped and polyadenylated mRNA. mRNA quality was determined by 4° urea-PAGE, 1% formaldehyde agarose gel or capillary electrophoresis with an Agilent 2100 Bioanalyzer (Agilent Technologies). A list of all primer sequences used are provided in Supplementary Data 7. Raw data images of gels presented are given in the Source Data.

**Formulation of RNAs in Polyplex (PLX).** In vivo-jetPEI® was purchased from Polyplus Transfection®. All other chemicals were multi-compendial grade and procured from Sigma-Aldrich and Fisher. Polyplex was prepared by mixing RNA and in vivo-jetPEI® in optimized buffer matrix. Briefly, in a 2 cc sterile glass vial, RNA stock solution was mixed with HEPES buffer and Glucose. PEI solution was prepared with histidine buffer and glucose in another sterile glass vial. Two parts were mixed in 1:1 ratio by volume, followed by 10-time inversion by hand. The final matrix was 5 mM histidine, 5 mM HEPES, and 5% w/v glucose, with N:P ratio of 12 and final pH 6.0. Samples were then aliquoted and transferred to 5 °C for stability study.

**Cell culture and transfections.** HEK293T (ATCC: CRL-3216) cells were cultured in Dulbecco’s Modified Eagle’s Medium (DMEM, Gibco, 11965–118) containing 2 mM L-glutamine, supplemented with 10% fetal bovine serum (EMD Millipore, 10079-14-0), 100 U/ml penicillin, and 0.1 mg/ml streptomycin (EmbryoMax ES Medium (DMEM, Gibco, 11965-118) containing 2 mM L-glutamine, supplemented with 10% fetal bovine serum (EMD Millipore, 10079-14-0). HEK293T cells were seeded in a 10 cm plate 24 h before transfection. 10 µg of pooled RNAs were transfected with complete DMEM supplemented with 10% FBS and Pen/Strep. For transfec-

**NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-022-28776-w | www.nature.com/naturecommunications**

---

**Measurement of RNA Stability by capillary electrophoresis for RNAs formulated in Polyplex (PLX).** All chemicals and reagents other than Fragment Analyzer kits (Agilent) were purchased from Sigma Aldrich and Fisher Scientific. The PLX-RNA samples were aliquoted and incubated at 5 °C. After each time point (0 h, 6 h, 1 day, 2 days, 3 days, and 14 days), sample aliquots were stored at −70 °C. Aliquots were thawed at the same time, and RNA integrity values were determined using an Agilent 2100 Bioanalyzer (Agilent, CA) in duplicate. Briefly, 45 µL of PLX-RNA samples were mixed with 5 µL of heparin solution (20 µL/L heparin, 10 mM HEPES, 1 mM EDTA, pH 7.4), and the mixture...
was incubated at 30 °C for 20 min. These processed samples were further diluted 200-fold using Fragment Analyzer high sensitivity RNA diluent maker (DNF-300-0004, Agilent Technologies). Truncated samples were loaded onto the Fragment Analyzer (FA). The FA analysis was performed using an RNA high sensitivity kit (DNF-472-0500, Agilent, CA), following the vendor-recommended testing protocol. The key experimental parameters are: Injection voltage — 7 kV; injection time — 100 s; separation voltage — 6 kV; separate time — 30 min. The FA electropherograms were processed with Prosize software (v3.0.1.6, Agilent, CA). The RNA degradation rate was modeled with first-order reaction kinetics (RNA integrity vs time) as shown in the following equation:

\[ \text{Integrity} = \text{Integrity}_0 e^{-kt} \]  

(1)

where \( \text{Integrity}_0 \) is the integrity of the sample at time zero, \( k \) is the half-life of RNA (50% degradation of RNA as from initial time point).

**Library preparation and amplicon sequencing.** Up to 250 ng RNA in 2.75 µL was mixed with 0.25 µL 2x RT_Consel2_N12_Read1Partial (Supplementary Data 7) and 0.25 µL 10x SYBR, 0.25 µL 10 µM Read1Partial_F, 0.25 µL 10 uM Read2Partial_Const1_R (ThermoFisher Scientific, AM2694), 0.25 µL Superscript IV (Thermo 18091050). The reaction was incubated at 55 °C for 45 min and inactivated at 80 °C for 10 min.

First round PCR was performed under following conditions: 1 µL RT reaction, 10 µL 2x Q5 Hot Start Master Mix (NEB M0494S), 0.2 µL 100x SYBR (Thermo ST5635, 1 µL 10x Read1Partial_F, 1 µL 10 µM 30.50 Hbb_Fwd_Nuc_Fwd mix in 20 µL total volume. Cycling conditions were 98 °C for 60 s, and 15 cycles of 98 °C for 10 s and 72 °C. Second round PCR was performed under the following conditions: 1 µL first round PCR, 10 µL 2x Q5 Hot Start Master Mix, 0.2 µL 100x SYBR, 1 µL 10x Read1Partial_F, 1 µL 10 µM Read2Partial_Const1_R in 20 µL total volume. Cycling conditions were 98 °C for 60 s, and 5 cycles of 98 °C for 10 s, 72 °C for 5 s. Sequencing adaptors were added using the following conditions for final round PCR: 1 µL second round PCR, 10 µL 2x Q5 Hot Start Master Mix, 0.2 µL 100x SYBR, 1 µL 10 µM NEBNext Index Primer (NEB E7355, NEB E7500, NEB E7710, NEB E7370, NEB E6609), 1 µL 10 µM NEBNext Universal PCR Primer in 20 µL total volume. Cycling conditions were 98 °C for 60 s, and 5 cycles of 98 °C for 10 s, 72 °C for 5 s. All barcoded samples were then pooled at equal volumes and purified with 1.1x SPRISelect beads (Beckman Coulter E23317). Sequencing was performed at the Stanford Functional Genomics Facility (SGF) at Stanford University, on an Illumina NextSeq 550 instrument, using a high output kit, 1 x 76 cycles. Primer sequences and the sequencing construct layout are provided in Supplementary Data 7.

**Amplicon sequencing data analysis.** After bcl conversion and demultiplexing with Illumina bcl2fastq, the constant regions were trimmed using cutadapt.91 The trimmed reads were aligned to the indexed reference of barcode sequences using Bowtie2 with the following options: -L 11 -N 0 --no Browns.) The aligned reads were then multiplexed onto the reference sequence by adding 5 µL of 15% dimethyl sulfate (v/v in ethanol) to the reaction mixture, and the mixture was incubated at 25 °C for 1 h. The remaining DNA was then hydrolyzed by adding the total volume of water to the reaction mixture, and the mixture was incubated at 37 °C for 15 min. Hydrolysis was quenched by bringing the reaction to 50 µL with 50 mM NaOH, 2.5 µM EDTA and incubation at 65 °C for 15 min. Hydrolysis was quenched by bringing the reaction volume to 50 µL with water, adding 100 µL of Oligo Binding Buffer (Zymo Research) and proceeding through the Zymo Oligo Clean and Concentrator 5 kit (Zymo Research). The template-primer mix was then used to amplify the full-length cDNA product was verified using agarose gel electrophoresis with SYBRSafe (Invitrogen).

**SHAPE and DMS chemical mapping of full-length mRNAs.** For DMS-based chemical mapping of the LinearDesign-1 and Yellowstone mRNAs, 1 µg of RNA was brought to 10 µL in water, unfolded at 95 °C for 2 min, then snap-cooled on ice for 1 minute. The RNA was then mixed with 25 µL water and 10 µL 5X folding buffer (1.5 M sodium cacodylate pH 7.0, 50 mM MgCl2) and folded at 37 °C for 30 min. The folded RNA was modified by adding 5 µL of 15% dimethyl sulfide (v/v in ethanol) or water (negative control). Both reactions were incubated at 37 °C for 6 min, quenched by the addition of 50 µL beta-mercaptoethanol, purified using the Zymo RNA Clean and Concentrator 5 kit (Zymo Research), and eluted in 12 µL water. For reverse transcription, 10 µL of the modified RNA was mixed with 1 µL of 10 mM oVT555, incubated at 65 °C for 5 min, and snap-cooled on ice. Then, the template-primer mix was converted to cDNA by adding 4 µL of 5X TqRT1 First Strand Synthesis Buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl, 15 mM MgCl2, 2 µL 10 mM dNTPs, 1 µL freshly prepared 100 mM DTT, and 0.5 µL TqRT-III (Invitrogen). The reaction was mixed and incubated at 57 °C for 2 min. The RNA was then hydrolyzed by the addition of 2 µL hydrolysis buffer (0.5 M NaOH, 0.25 M EDTA) and incubation at 65 °C for 15 min. Hydrolysis was quenched by bringing the reaction volume to 50 µL with water, adding 100 µL of Oligo Binding Buffer (Zymo Research), and proceeding through the Zymo Oligo Clean and Concentrator (Zymo Research) purification kit (Zymo Research). The cDNA was then eluted into 10 µL of 2X SuperScript IV (Invitrogen) buffer (1.5 M sodium cacodylate pH 7.0, 50 mM MgCl2) and incubated at 37 °C for 20 min. Nine microliters of the folded cDNA was mixed with 1 µL of either 100 mM 1M7 (1-Methyl-7-nitroisatoic anhydride, freshly mixed in DMSO) or neat DMSO (negative control), mixed thoroughly by pipetting, and incubated at 37 °C for 75 s (roughly 5 MIN) at 57 °C. After chemical treatment, the volume of both reactions was brought to 50 µL with water, cleaned up with the Zymo RNA Clean and Concentrator 5 kit, and eluted in 12 µL of water. Ten microliters of eluted RNA was mixed with 1 µL of 200 ng/µl random nonamer primer (New England Biolabs), incubated at 65 °C for 5 min, then snap-cooled on ice. Then, 8 µL of 2X Mal buffer (125 mM Tris pH 8.0, 187.5 mM KCI, 25 mM DTT, 1.25 mM each dNTPs, 15 mM MnCl2), was added to the primer-template mixture, incubated at room temperature for 2 min, and then mixed with 1 µL of SuperScript II (Invitrogen). The reverse transcription reaction was thoroughly mixed by pipetting, incubated at room temperature for 10 min, then at 42 °C for 3 h. The reverse transcription enzyme was heat inactivated at 70 °C for 15 min, snap-cooled on ice, then immediately mixed with 8 µL NEBNext Second Strand Synthesis Reaction Buffer (New England Biolabs), 4 µL NEBNext Second Strand Synthesis Enzyme Mix (New England Biolabs), and 48 µL water. The second strand synthesis reaction was mixed and incubated at 16 °C for 60 min. The resulting double-stranded cDNA was purified with 1X volumes of Select-a-Size DNA Clean and Concentrator MagBeads (Zymo Research).
Double-stranded cDNA was prepared for Illumina sequencing with the NEBNext Ultra II FS DNA Library Prep Kit for Illumina (New England Biolabs) and T7 primers. Briefly, 100–500 ng of DNA from either 1M7 or DMS conditions in 26 µL of water was fragmented and end-repaired by adding 7 µL NEBNext Ultra II FS Reaction Buffer and 2 µL NEBNext Ultra II FS Enzyme Mix, mixed thoroughly by vortexing, and incubated in a thermocycler with heated lid on at 37 °C for 20 min, then 65 °C for 20 min. Fresh ligation adapter was prepared by denaturation in solution of 15 µM each of T7 oligo and T7RssU1-sticky ends (0.2 M TLE (10 mM Tris pH 8.0, 0.1 mM EDTA, 100 mM NaCl) at 95 °C for 1 min, then annealed via slow cooling at −0.1 °C/25 °C. The fragmented and end-repaired DNA was mixed with 2.5 µL 15 mM ligation adapter, 30 µL NEBNext Ultra II FS Ligation Master Mix, and 1 µL NEBNext Ultra II FS Ligation Enhancer. The ligation reaction was incubated in a thermocycler (heated lid off) at 20 °C for 60 min. Ligated fragments were purified with 0.9X Select-a-Size DNA Clean and Concentrator MagBeads (Zymo Research) and eluted in 25 µL water. Dual-index sample barcodes were added through indexing PCR using the NEBNext QS Hot Start HiFi PCR Master Mix in a reaction with 0.5 µM each of iTru_5 and iTru_7 indexing primers (New England BioLabs) for a final volume of 15 µL. The reaction was incubated at 37 °C for 3 h. After transcription, the DNA template was removed through the addition of 2 µL of DNase I (add 2 µL of DNase I to 26 µL of reaction mix, followed by incubation at 37 °C for 30 min). After DNA digestion, the RNA was purified using a mixture of AMPure XP beads (Beckman Coulter) with 40% polyethylene glycol (mixed in a 7:3 ratio). Final elution into 25 µL of nuclease-free water yielded purified RNA.

Degradation in-line probing of RNA samples. For degradation experiments, 45–50 pmol of RNA was subject to four conditions: (1) 50 mM Na-CHES buffer (pH 10.0) at room temperature without added MgCl₂; (2) 50 mM Na-CHES buffer (pH 10.0) at room temperature with 10 mM MgCl₂; (3) phosphate buffered saline (PBS, pH 7.2); (4) Thermofisher Scientific-Gibco 20012027 at 30 °C without added MgCl₂ and (4) PBS (pH 7.2). For degradation reactions containing MgCl₂, RNA was collected at 0 and 24 h time points. For reactions without MgCl₂, timepoints were collected at 0 and 7 days. At each timepoint, the degradation reaction was quenched with 15 µL of 500 mM Tris-8HCl (pH 7) and 3 µL of 500 mM NaEDTA. Quenched samples were brought to a final volume of 100 µL with nuclease-free water. The RNA was purified through precipitation as follows. 1.5 µL of Glycine Blue (Thermo Fisher), 10 µL of 3 M sodium acetate (pH 5.2), and 330 µL of cold 100% ethanol were added, the reaction mixture mixed, and incubated on dry ice for 20–30 min. After incubation, the reaction was centrifuged at 21,000 rcf for 30 min. The resulting pellet was washed twice with 500 µL of cold 70% ethanol, and pelleted after each wash step. Finally, the reaction was dried for 10 min at room temperature to remove any residual ethanol, and resuspended in 5 µL of nuclease-free water.

Structure probing of RNA samples. In parallel with the in-line hydrolytic degradation conditions above, we carried out SHAPE structure probing experiments68 as follows. 1 µM of purified RNA was added to 2 µL of 500 mM HEPS buffer (pH 8.0) and denatured at 90 °C for 3 min. The reaction was then cooled down to room temperature over 10 min. Two microliters of 100 mM MgCl₂ was then added, followed by incubation at 50 °C for 30 min. The sample was cooled down to room temperature over 20 min before addition of 5 µL of 1-methyl-7-nitroisatoic anhydride (1M7, 8.48 mg/mL of DMSO) followed by incubation at room temperature for 15 min, and brought to a final volume of 20 µL with nuclease-free water. The reaction was quenched with 5 µL of 500 mM Na-MES pH 6.0, the reaction was adjusted to be 100 µL, and purified with ethanol precipitation as above. As a control, a sample was prepared in parallel without the addition of 1M7 but subject to the same protocol described above.

Preparation of cDNA and sequencing. cDNA was prepared from the six RNA samples (two from structure probing, four from degradation). Five microliters of purified RNA was added to a reaction mixture containing 1X First Strand buffer (Thermo Fisher), 5 mM dithiothreitol (DTT), 0.8 mM dNTPs, 0.6 µL of SuperScript III First strand synthesis buffer (Thermo Fisher), and 0.5 mg/ml of BSA, and a final volume of 15 µL. The reaction was incubated at 48 °C for 40 min and stopped with 5 µL of 0.4 M sodium hydroxide. The reaction was then incubated at 90 C for 3 minutes, cooled on ice for 3 min, and neutralized with 2 µL of quench mix (prepared as 2 mL of 5 M sodium chloride, 3 mL of 3 M sodium acetate, 2 mL of 2 M hydrochloric acid).

cDNA was pooled down with 1.5 µL of Oligo C′ beads65 (in house magnetic beads prepared by immobilizing 2X Biotin oligonucleotides with Dynabeads™ MyOne Streptavidin C1; Thermo Fisher Scientific 65001), washed twice with 70% ethanol, then resuspended in 3.0 µL of water. We pooled 1.5 µL of each sample together to make 9 µL of cDNA. We used our custom protocol for sequencing library preparation by using Circ. Ligase I (Lucigen) at 68 °C for 2 h. The reaction was stopped by incubation at 80 °C for 10 min. cDNA was added to 10 µL of 5 M NaCl and pulled down with a magnetic bead and washed with 70% ethanol; the ligated product was resuspended in 15 µL H₂O.

The ligated product was quantified by qPCR. dsDNA at 3 nM concentration was sequenced using an Illumina Miseq (High output, Read 1 = 101 cycles and Read 2 = 51 cycles). The resulting data were analyzed using MAPseeker69 following the recommended steps for sequence assignment, peak fitting, background subtraction of the no-modification control, correction for signal attenuation, and reactivity profile normalization.

In-line and SHAPE probing by capillary electrophoresis. One-by-one follow-up to profile degradation of RNA fragments at single-nucleotide resolution through capillary electrophoresis was carried out with some differences to the pooled In-line-seq experiments above. We describe these steps below; see ref. 70 for further details and Supplementary Data 7 for primers.

DNA template preparation. DNA templates were designed to include the 20-nt T7 RNA polymerase promoter sequence (5′-TCTAATACGACTCACTATAGG-3′) followed by the remaining sequence encoding desired RNA. Double-stranded template were prepared by extension of 60-nt DNA oligomers (IDT, Integrated DNA Technologies) with Phusion DNA polymerase (Finnzymes), using the following thermocycler protocol: denaturation for 30 s at 98 °C, 35 cycles of denaturation for 10 s at 98 °C annealing for 30 s at 60–64 °C, extension for 30 s at 72 °C final extension for 10 min at 72 °C and cooling to 4 °C.

DNA samples were purified with AMPure magnetic beads (Agencourt, Beckman Coulter) following the manufacturer’s instructions. Sample concentrations were estimated based on UV absorbance at 260 nm measured on Nanodrop 100 or 8000 spectrophotometers. Verification of template length was
accomplished by electrophoresis of all samples and 10-bp and 20-bp ladder length standards (Thermo Scientific O’RangeRuler SM313 & SM312) in 4% agarose gels (containing 0.5 μg/μL ethidium bromide) and 1 x TBE (100 mM Tris, 83 mM boric acid, 1 mM disodium EDTA). All sample manipulations, including the following steps, were carried out in 96-well V-shaped polypropylene microplates (Greiner).

**Preparation of RNA templates.** In vitro transcription reactions were carried out in 40 μL volumes with 10 pmol of DNA template, using the TranscriptAid T7 High Yield Transcription Kit (Thermo Fisher, K0441). In some experiments, pseudouridine-5'-triphosphate (TriLink Biotechnologies, N-1019) or N1-methylpseduouridine-5'-triphosphate (TriLink Biotechnologies, N-1081) were used to replace regular NTP. Reactions were incubated for 3 h at 37°C. Following degradation of DNA template with 2 μL of DNase I at 37°C for 30 min. RNA samples were purified with 1.8 x volume of AMPure XP beads (Beckman Coulter) mixed with 40% PEG-8000 (ratio of 7:3). Following the manufacturer's instructions. Concentrations were measured by absorbance at 260 nm on Nanodrop 100 or 8000 spectrophotometers.

**In-line probing (hydrolytic degradation profiling).** For degradation experiments, 750 ng of RNA was subjected to 50 μM Na-CHES buffer (pH 10) at room temperature with 10 mM MgCl₂ at a concentration of 0.08 mM) was used in the reverse-transcription reaction.

**Polysome selection and library preparation.** The variant 5' UTR is composed of: fixed first 29 nt of HBB, variable 35 nt (initially degenerate) and 6 nt (Kozak consensus). See Supplementary Data 7 for the detailed construct layout. To generate the reporter mRNA pool containing the variant 5' UTR library, IVT template was first assembled by PCR under the following conditions: 4 μL 10x AccuPrime PfX Reaction Mix, 0.4 pmol HB29_35Npmp, 0.4 pmol Nluc_HBB_3UTR, 0.4 μL of Primerase in 40 μL total reaction volume. Cycling conditions are: 95°C for 120 s, and 45 cycles of 95°C for 15 s, 66°C for 30 s, 68°C for 75 s. PCR product was purified on silica columns (NEB T1034) and amplified with the following conditions: 4 μL 10x AccuPrime PfX Reaction Mix, 4 μL 10 μM TT_28_HBB_30_F, 4 μL 10 μM Nanoluc_ORF_R, 0.4 μL AccuPrime PfX Polymerase in 40 μL total reaction volume. Cycling conditions are 95°C for 120 s, and 4 cycles of 95°C for 15 s, 66°C for 30 s, 68°C for 75 s. The mRNA pool was run on an agarose gel and confirmed as described above. This yields an estimated initial starting degenerate pool complexity of ~2.4 x 10¹¹.

**Measurement of in-solution mRNA stability by capillary electrophoresis.** For each time point, 1.6 pmol of mRNA brought to 10 μL in a buffer containing 50 mM Na-CHES at pH 10 with 10 mM MgCl₂, and the reaction was incubated at 25°C. When the incubation period was reached for each time point, 5 μL of Tris-HCl at pH 7 and 1 μL of 500 mM EDTA in nuclease free water was added to quench the degradation reaction, and frozen for further analysis. After the final time point (24 h), 4 μL of each RNA degradation sample (out of a total stored volume of 16 μL) was taken, and mixed with 1 μL of a control RNA at a concentration of 50 ng/μL. For these experiments, we used the P4-P6 domain of the Tetrahymena ribozyme with two added hairpins (-239 nt) as our control. The RNA mixture was then purified using a mixture of AMPure XP beads (Beckman Coulter) with 40% polyethylene glycol (mixed in a 7:3 ratio). The resulting RNA was eluted into 4.5 μL of RNase-free water for analysis on the 2100 Bioanalyzer (Agilent) using the RNA Nano Eurokaryote protocol.

The data from the Bioanalyzer were analyzed using a custom script that performs the following analysis. We first converted elution times to nucleotides based on a ladder control (25, 200, 500, 1000, 2000, and 4000 nts). We then estimated relative mRNA amounts based on peak areas at expected band lengths (in our case, ~900 nucleotides for the mRNAs of interest and ~265 nucleotides for the control). When calculating peak areas, we performed background subtraction, where the background was defined as the area under a linear in the range of no-modification peaks. Normalization was performed using different methods to cross-validate. First, we normalized the peak areas of full-length mRNA to the control P4-P6 domain RNA that was spiked into the samples after degradation was performed. Second, we also normalized peak areas of full-length mRNAs to the total amount of RNA in the lane less the peak area of the band. This was done by running a ladder control (25, 200, 500, 1000, 2000, and 4000 nts), assuming that the majority of the other RNA in the lane were degradation products from the mRNA of interest. These different approaches to normalizing the data gave the same results for half-life relative to reference mRNA within estimated error (see below). After calculations of normalized peak areas, we then calculated fraction intact values for each mRNA by dividing the normalized area across the ten time points by the normalized area at the start (t=0).

\[
\text{Fraction Intact} = \frac{\text{Normalized Area}_{\text{sample}}}{\text{Normalized Area}_{\text{control}}} \tag{6}
\]

For each sample, we fit fraction intact values across the different timepoints to an exponential function,

\[
F_i = A e^{-t/T} \tag{7}
\]

where \(F_i\) is an array of fraction intact values across multiple time points, \(A\) is the amplitude of the exponential decay function, \(r\) is the time constant, and \(t\) is an array of time points in hours. We then used the time constant to calculate the in vitro half-life of mRNA:

\[
\text{Half-life} = \ln(2)t \tag{8}
\]

Scripts and data are available at https://github.com/DasLab/openvaccine-CE-analysis.

**Polysome selection and library preparation.** The variant 5' UTR is composed of: fixed first 29 nt of HBB, variable 35 nt (initially degenerate) and 6 nt (Kozak consensus). See Supplementary Data 7 for the detailed construct layout. To generate the reporter mRNA pool containing the variant 5' UTR library, IVT template was first assembled by PCR under the following conditions: 4 μL 10x AccuPrime PfX Reaction Mix, 0.4 pmol HB29_35Npmp, 0.4 pmol Nluc_HBB_3UTR, 0.4 μL of Primerase in 40 μL total reaction volume. Cycling conditions are: 95°C for 120 s, and 45 cycles of 95°C for 15 s, 66°C for 30 s, 68°C for 75 s. PCR product was purified on silica columns (NEB T1034) and amplified with the following conditions: 4 μL 10x AccuPrime PfX Reaction Mix, 4 μL 10 μM TT_28_HBB_30_F, 4 μL 10 μM Nanoluc_ORF_R, 0.4 μL AccuPrime PfX Polymerase in 40 μL total reaction volume. Cycling conditions are 95°C for 120 s, and 4 cycles of 95°C for 15 s, 66°C for 30 s, 68°C for 75 s. The mRNA pool was run on an agarose gel and confirmed as described above. This yields an estimated initial starting degenerate pool complexity of ~2.4 x 10¹¹.

Transfection of HEK-293 cells and sucrose gradient fractionation were performed as described above. The predominant peak between 10–16 was pooled and RNA was by acid phenol chloroform extraction followed by column purification (Zymo Research, R1013) as described above. 1/3 lysate volume was kept as input before layering onto the sucrose gradient and RNA was extracted from the input lysate by TriZol extraction followed by column purification. 1.5 μg RNA in 5.5 μL was further purified with 0.5 μL 2x R_T Nuc26, UM12_Read1Partial (Supplementary Data 7) and 0.5 μL 100 μM dNTPs each. The RNA samples were then denatured at 65°C for 5 min and chilled to 4°C. 3.5 μL reverse transcription mix was added to

(https://github.com/ribokit/HITRACE), following the recommended steps for sequence assignment, peak fitting, background subtraction of the no-modification control, correction for signal attenuation, and reactivity profile normalization.
10 µl total reaction volume; 2 µl 5x Superscript IV buffer, 0.5 µl 10 mM DTT, 0.5 µl Superscript I (Thermo 19091050). The reaction was incubated at 25 °C for 5 min, inactivated at 95 °C for 10 min. Variant 5' UTR amplicon was amplified from the reverse transcription reaction via PCR under the following reaction conditions: 4 µl RT reaction, 40 µl 2x Q5 Hot Start Master Mix (NEB M0494S), 0.8 µl 100X SYBR (Thermo S7563), 4 µl 10 µM T7_28_HBB_29_F, 4 µl 10 µM Nanoluc_ORF_R, in a 20 µl total reaction volume. Cycling conditions were as follows: 98 °C for 60 s, and 15 cycles of 98 °C for 10 s, 68 °C for 10 s, 72 °C for 10 s. PCR product was purified on silica columns (NEB T1034) and assembly with Nucl_HBB_3UTR fragment was performed as described above for initial preparation using HBB29_N35 primer set. The mRNA structure was in vitro transcribed, capped, and polyadenylated as described above. The same process of transcription, fractionation, reverse transcription, PCR amplification, assembly and in vitro transcription was repeated.

For sequencing library preparation, the RT reaction was PCR amplified under the following conditions: 1 µl RT reaction, 10 µl 2x Q5 Hot Start Master Mix (NEB M0494S), 0.2 µl 100X SYBR (Thermo S7563), 1 µl 10 µM Read1, 1 µl 10 µM Read2Partial_HBB29 in 20 µl total reaction volume. Cycling conditions were as follows: 98 °C for 60 s, and 15 cycles of 98 °C for 10 s, 68 °C for 10 s, 72 °C for 10 s. Sequencing adapters were added using the following conditions for final round PCR: 1 µl first round PCR reaction, 10 µl 2x Q5 Hot Start Master Mix, 0.2 µl 100X SYBR, 1 µl 10 µM NEBNext Index Primer (NEB E7335, NEB E7500, NEB E7710, NEB E7730, NEB E6609), 1 µl 10 µM NEBNext Universal PCR Primer in 20 µl total volume. Cycling conditions are: 98 °C for 60 s, and 5 cycles of 98 °C for 10 s, 72 °C for 10 s, 72 °C for 10 s. All barcoded samples were then pooled at equal volumes and purified with 1.1x SPRIselect beads Beckman Coulter B23317). Sequencing was performed at the Stanford Functional Genomics Facility (SFGF) at Stanford University, on the Illumina NextSeq 550 instrument, using a high output kit, 1 × 81 cycles. Primer sequences and the sequencing construct layout are provided in Supplementary Data 7.

Polysome selection library sequencing data analysis. Following adapter trimming, 670,440 sequences with at least 10 unique read count across all libraries combined were set as the reference. Each library was aligned to this indexed reference using Bowtie2. Only uniquely mapping reads with edit distance ≤6 were retained. Alignments were further deduplicated using UMIcollapse (-p 0.05, -k 10)106. This results in the matrix of read count where rows are different sequence candidates, the criteria of adjacent groups (R package ordPens)100. False discovery rate was estimated by Taylor series method. Error and significance tests for single-exponential in-solution half-lives were estimated by non-parametric bootstrapping, except for data on the 24 nanoluciferase measurements in Fig. 4, which made use of error estimates derived from triplicate measurements.

Design algorithms/analysis. The protein sequences for each target were used to generate DNA sequences at Integrated DNA Technologies (IDT, https://www.idtdna.com/CodonOpt), Twist Biosciences (https://ecommerce.twistdna.com/app), and GENEVIZ (https://climsg4.geneviz.com/ToolBox/CodonOptimization). Solutions from LinearDesign were obtained using the LinearDesign server (http://rna.baidu.com/). Ribotree solutions were terminated after 6600 iterations. The Ribotree code is available for noncommercial use at https://eternagame.org/about/software, with example usage to replicate runs for this work provided.

Structure prediction and ensemble-based calculations were performed using LinearFold and LinearPartition with ViennaRNA, CONTRAfold, and EternaFold parameters. Secondary structure features were calculated from predicted MFE with ViennaRNA, CONTRAfold, and EternaFold. The RiboTree code is available for noncommercial use at https://eternagame.org/about/software. Scripts and data for the measurement of in-solution mRNA stability by capillary electrophoresis are available at https://github.com/DasLab/openvaccine-CE-analysis. Codes and formatted datasets to reproduce the linear DegScore model and degradation prediction calculations is available at https://github.com/eternagame/DegScore. PERSIST-seq processing pipeline is available at https://github.com/barnalab/persist.

Received: 8 April 2021; Accepted: 7 February 2022; Published online: 22 March 2022

References
1. Kowalski, P. S., Rudra, A., Miao, L. & Anderson, D. G. Delivering the messenger: advances in technologies for therapeutic mRNA delivery. Mol. Ther. 27, 710–728 (2019).
2. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines - a new era in vaccinology. Nat. Rev. Drug Discov. 17, 261–279 (2018).
3. Sahin, U., Kanki, K. & Türeci, O. mRNA-based therapeutics: developing a new class of drugs. Nat. Rev. Drug Discov. 13, 759–780 (2014).
4. Jackson, N. A. C., Kester, K. E., Casimiro, D., Gurunathan, S. & DeRosa, F. The promise of mRNA vaccines: a biotech and industrial perspective. npj Vaccines 5, 11 (2020).
5. Weng, Y. et al. The challenge and prospect of mRNA therapeutics landscape. Biotechnol. Adv. 40, 1075534 (2020).
6. Crommelin, D. J. A., Anderson, T. J., Volkin, D. B., Jasko, W. & Mastrobattista, E. Addressing the cold reality of mRNA vaccine stability. J. Pharm. Sci. 110, 997–1001 (2021).
7. Waymanton-Steele, H. K. et al. Theoretical basis for stabilizing messenger RNA through secondary structure design. Nucleic Acids Res. 49, 10604–10617 (2021).
8. Mauger, D. M. et al. mRNA structure regulates protein expression through changes in functional half-life. Proc. Natl Acad. Sci. USA 116, 24073–24083 (2019).

9. Thess, A. et al. Sequence-engineered mRNA without chemical nucleoside modifications enables an effective protein therapy in large animals. Mol. Ther. 23, 1456–1465 (2015).

10. Lee, J. et al. RNA design rules from a massive open laboratory. Proc. Natl Acad. Sci. USA 114, 21222–21237 (2017).

11. Li, Y. & Breaker, R. R. Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2′-hydroxyl group. J. Am. Chem. Soc. 121, 5364–5372 (1999).

12. Jia, L. et al. Decoding mRNA translatability and stability from the 5′ UTR. Nat. Struct. Mol. Biol. 27, 814–821 (2020).

13. Asrani, K. H. et al. Optimization of mRNA untranslated regions for improved expression of therapeutic mRNA. RNA Biol. 15, 756–762 (2018).

14. Sultana, N. et al. Optimization of 5′ untranslated region of modified mRNA for use in cardiac or hepatic ischemic injury. Mol. Ther. Methods Clin. Dev. 17, 622–633 (2020).

15. Leppke, K., Das, R. & Barma, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat. Rev. Mol. Cell Biol. 19, 158–174 (2018).

16. 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).

17. Cao, J. et al. High-throughput 5′ UTR engineering for enhanced protein production in non-viral gene therapies. Nat. Commun. 12, 4138 (2021).

18. Zhao, W. et al. Massively parallel functional annotation of 3′ untranslated regions. Nat. Biotechnol. 32, 387–391 (2014).

19. Orlandini von Nissen, A. G. et al. Improving mRNA-based therapeutic gene delivery by expression augmenting 3′ UTRs identified by cellular library screening. Mol. Ther. 27, 824–836 (2019).

20. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. Nature 473, 337–342 (2011).

21. Hsieh, A. C. et al. The translational landscape of mTOR signalling steers cancer initiation and metastasis. Nature 485, 55–61 (2012).

22. Thoren, C. C. et al. A unifying model for mTORC1-mediated regulation of mRNA translation. Nature 485, 109–113 (2012).

23. Leppke, K. et al. Gene- and species-specific Hox mRNA translation by ribosome expansion segments. Mol. Cell 80, 980–995.e13 (2020).

24. Patel, M., Siegel, A. J. & Berry, J. O. Untranslated regions of PbsRs1 mRNA mediate bundle sheath cell-specific gene expression in leaves of a C4 plant. J. Biol. Chem. 281, 25485–25491 (2006).

25. Wilusz, J. E. et al. A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly(A) tails. Genes Dev. 26, 2392–2407 (2012).

26. Hickey, K. L. et al. GIGYF2 and 4EHP inhibit translation initiation of defective messenger RNAs to assist ribosome-associated quality control. Mol. Cell 79, 950–962.e6 (2020).

27. Wollner, C. I. et al. A mRNA-LNP vaccine against Dengue Virus elicits robust, serotype-specific immunity. J. Virol. 95, e00820-20 (2021).

28. Ferriz, M. et al. Human cellular CYBA UTR sequences increase mRNA translation without affecting the half-life of recombinant RNA transcripts. Sci. Rep. 6, 39149 (2016).

29. Zeng, C. et al. Leveraging mRNA sequences express SARS-CoV-2 antigens vivo. Adv. Mater. 32, e2004452 (2020).

30. Kim, D. et al. The architecture of SARS-CoV-2 transcriptome. Cell 181, 914–921.e10 (2020).

31. Huston, N. C. et al. Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. Mol. Cell 81, 584–598.e5 (2021).

32. Chiu, W.-W., Kinney, R. M. & Dreher, T. W. Control of translation by the 5′ UTR. Annu. Rev. Microbiol. 47, 1–24 (1993).

33. Sarnow, P. & Miller, W. A. 3′ cap-independent translation enhancers of plant viruses. Annu. Rev. Microbiol. 67, 21–42 (2013).

34. Nicholson, B. L. & White, K. A. 3′ Cap-independent translation enhancers of positive-strand RNA plant viruses. Curr. Opin. Virol. 1, 373–380 (2011).

35. Gallic, D. R. The 5′-leader of tobacco mosaic virus promotes translation through enhanced recruitment of eIF4F. Nucleic Acids Res. 30, 3401–3411 (2002).

36. Gallic, D. R., Tanguay, R. L. & Leathers, V. The tobacco etch viral 5′ leader and poly(A) tail are functionally synergistic regulators of translation. Gene 165, 233–238 (1995).

37. Shirokikh, N. E. & Spirin, A. S. Poly(A) leader of eukaryotic mRNA bypasses the dependence of translation on initiation factors. Proc. Natl Acad. Sci. USA 105, 10738–10743 (2008).

38. Sokolowski, K. J. et al. Sindbis virus usurps the cellular HuR protein to stabilize its transcripts and promote productive infections in mammalian and mosquito cells. Cell Host Microbe 8, 196–207 (2010).
