mRNArchitect: sequence design of mRNA medicines

Magdalena A. Budzinska[1,2], Mohammad Zardbini[1,2], Michelle, P. Siddharta[1,2], Jonathon Waterhouse[1,2], Theodore E. Leonard[1,2], Aayushi Ghodasara[1,2], Sneha Lakshman Das[1,2], Rachel Y. Chang[1,2], Natasha Chaudhary[1,2], Seth W. Cheetham[1,2], Timothy R. Mercer[1,2], Helen M. Gunter[1,2,*]

1: Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane, QLD, Australia

2: BASE Facility, University of Queensland, Brisbane, QLD, Australia

*: Corresponding Author

Abstract

Sequence design is the critical first step in developing a new mRNA medicine. The mRNA primary sequence is assembled from multiple elements, including the coding sequence of the target protein or gene, flanking untranslated regions that impact expression, and the poly(A) tail, which improves stability. The mRNA primary sequence can be optimized to improve the performance of an mRNA medicine, including increasing translation and stability, and reducing reactogenicity. Here, we introduce *mRNArchitect*, a software toolkit to assist in the design of mRNA sequences. *mRNArchitect* can rapidly assemble and optimize mRNA sequences based on user requirements, including GC content, secondary structure, codon optimization, and uridine depletion. The sequences generated by *mRNArchitect* can also be readily synthesized, and have been experimentally validated across a variety of applications. We present *mRNArchitect* as a software toolkit to help scientists design new mRNA medicines.

Availability and implementation: *mRNArchitect* is available via a web portal at www.basefacility.org.au/software/ and Github at https://github.com/BaseUQ/mRNArchitect. *mRNArchitect* is open-source and freely available under MIT License.

1. INTRODUCTION

Careful sequence design can markedly improve the performance of an mRNA by enhancing its translation, increasing stability and reducing reactogenicity[1,2]. Therefore, the primary sequence design is a key first step when developing a new mRNA vaccine or therapy.

mRNA medicines can encode any coding sequence (CDS), from small peptides to large structural proteins. The CDS of interest is reverse translated into an mRNA sequence using codons optimised for high expression in the target organism or tissue[3]. The CDS can also be depleted of uridines that are recognised by receptors that initiate the innate immune response[4].

An mRNA medicine also encodes untranslated regions (UTR) that flank the CDS and regulate translation and stability. The ribosome binds the 5' UTR and scans for the methionine start codon, while the 3' UTR contains binding sites for proteins and non-coding RNAs that impact cell-specific expression. The mRNA ends with a poly(A) tail that protects against 3' exonuclease degradation and improves stability.

Sequence design also considers the formation of RNA secondary structures that stabilise the mRNA and interact with proteins. RNA structures, such as hairpins, can slow ribosome binding and translation, but protect against nuclease digestion. Double-stranded RNA (dsRNA) structures can also be detected by RIG-I and toll-like receptors that activate the innate immune response[5].

In addition to improving performance, the mRNA sequence design should also be compliant with manufacturing requirements. Ideally, the output mRNA sequences should be easily synthesized into DNA templates, avoid error-prone, repetitive sequences, and produce high yields of mRNA during *in vitro* transcription.

Different mRNA design strategies may be preferred depending on the final clinical application. For example, mRNA vaccine sequences with extensive secondary structure may be preferred as they improve mRNA stability and confer an adjuvant effect[6]. Alternatively, therapeutic applications that require frequent dosing may avoid uridine and secondary structures to minimize potential inflammatory side effects[7].

To assist in the design of mRNA sequences for different applications, we have developed *mRNArchitect*, a software toolkit that can assemble and optimize mRNA sequences according to user specifications. Output sequences can be readily synthesised and have been experimentally validated in a wide range of cell-types and applications. We provide *mRNArchitect*, open-source software that assists in the design of new mRNA medicines.

2. FEATURES

2.1. Sequence input

mRNArchitect enables the design of mRNAs based on user-supplied sequences. First, an open-reading frame encoding the user's protein of interest is supplied (**see Figure 1A**). The coding sequence can be provided in either amino acid, RNA, or DNA format, and should initiate with a methionine codon (M, AUG) and terminate with one or more stop codons (e.g., *, UAA, UAG, UGA).

Users can also supply custom untranslated regions. By default, we use 5' and 3' UTRs based on the human alpha-globin gene (*HBA1*; Gene ID 3039) that has been validated in a wide range of cell types and applications. Finally, users can include a poly(A) tail in mRNA designs, either specifying its length, or supplying more complex designs, such as segmented poly(A) tails[8].

2.2. Optimisation parameters

mRNArchitect can optimize mRNA sequences according to a range of parameters. *mRNArchitect* is built upon the *DNA Chisel* framework[9], which determines the optimal sequence based on a set of criteria, including GC content, codon translation speed, and mRNA secondary structure (e.g. hairpins; **Figure 1B**). Default design parameters are derived from expression vector optimisation (https://basebuddy.lbl.gov/), however custom parameters can be specified, including both hard- and soft constraints. Hard constraints, such as avoidance of specified sequences are strictly enforced, while for soft constraints, such as codon optimisation, the constraint score is maximized. These parameters can be modified according to the user's requirements and often depend on the target organism, tissue and intended application.

There are several key parameters that impact mRNA expression and reactogenicity. To enhance translation, the CDS is codon-optimized to use preferred codons of the specified target organism (e.g. human or mouse). *mRNArchitect* also allows uridine depletion (UD), which excludes uridine at wobble positions in codons. Difficult sequence motifs, such as motifs associated with ribosomal frameshifting[10], can also be avoided. *mRNArchitect* also allows users to impose a minimum and maximum GC fraction across a defined window, which is a key variable in the formation of stable RNA secondary structures. In addition to these performance considerations, it is necessary for mRNA sequences to be manufactured to a consistently high quality. *mRNArchitect* allows users to avoid restriction enzyme binding sites used during DNA template assembly and avoid low-complexity, repetitive sequences that cause errors during DNA synthesis and *in vitro* transcription.

2.3. Output mRNA sequences

The *mRNArchitect* output provides fully assembled and optimized mRNA sequence candidates that are ready for synthesis (**Figure 1C**). Sequence outputs are accompanied by statistics describing their nucleotide- and structural composition.

Reported statistics include the Codon Adaptation Index (CAI) score, which measures how closely the codon usage of the sequence matches the codon preference of a particular organism[3,11]. The MFE, which is calculated using the *ViennaRNA* package, represents the lowest Gibbs free energy change associated with the formation of secondary structures in RNA molecules due to intramolecular base pairing[12].

3. RESULTS

3.1. Running *mRNArchitect*

To demonstrate the use of *mRNArchitect*, we designed mRNAs encoding the Luciferase protein[13]. Sequence inputs were based on a modified version of the *P. pyralis* Luciferase amino acid sequence (GenBank: AB644228.1; **Supplementary Table 1**). We used 5' and 3' UTRs based on human- and mouse alpha-globin sequences respectively (*HBA1*; Gene IDs 3039; 15122) and a 120 nt poly(A) tail.

First, *mRNArchitect* was used to design *Luciferase* mRNAs using the human codon model[3] and UD, with all other parameters set to default values. This strategy yielded an mRNA design with a CAI of 0.859 a Minimum Free Energy (MFE) of -534.2, which we used as a point of comparison for further optimisation strategies (**Supplementary Figure 1A-B**). Different nucleotide inputs (encoding identical amino acid sequences) had a limited impact on the CAI and MFE of output sequences.

Altering the *mRNArchitect* optimisation parameters impacted our designs more significantly. *Luciferase* mRNA designs optimised without UD showed a higher CAI and a lower MFE than designs with UD, despite only a modest increase in overall Uridine content (~4%) (**Supplementary Table 2**). Moreover, increasing the maximum GC content to 70% or more,

progressively increased the CAI and decreased the MFE of our designs. However, these designs were more difficult to manufacture with DNA synthesis (**Supplementary Figure 1C**).

To investigate the limits of the mRNA sequences generated by *mRNArchitect*, we prepared designs that maximised and minimised CAI (**Figure 1D, Supplementary Table 1**). The *mRNArchitect* maximum CAI design (termed "*mRNArchitect*" here on) was generated by relaxing the optimisation constraints of *mRNArchitect* (i.e. no UD, Min GC = 0, Max GC = 1.0; Avoid repeat length = 1000) so that only CAI value is considered for sequence optimisation. As a negative control, we designed a "minimum CAI" codon model, that assigns a high score to rare codons, thereby enriching them in the resulting design (**Supplementary Table 2**). This codon model was used to prepare a Minimum CAI *Luciferase* mRNA sequence (termed "*MinCAI*" here on).

We benchmarked the *mRNArchitect* and *MinCAI* designs against different publicly-available mRNA design tools, including mRNAid and LinearDesign[1,2] (**Supplementary Figure 1D-E**). mRNAid was used to design a *Luciferase* mRNA, with parameters that include CAI optimization with UD, MFE estimation using ViennaRNA, avoiding BsaI, Xba and XhoI cleavage sites, and a target GC content of 0.4 – 0.6 across a 60nt GC content window with all other parameters kept as default. LinearDesign, was also used to design a *Luciferase* mRNA, with the default parameters, and with lambda set to zero[2].

The *mRNArchitect Luciferase* design showed the highest similarity to mRNAid (94%), then LinearDesign (81%), then the Wild-type *Luciferase* (73%) and *MinCAI* designs (66%) (**Figure 1D, Supplementary Figure 1G**). With the exception of mRNAid, all designs required specialized DNA synthesis due to difficult and repetitive sequences (**Supplementary Figure 1F**).

We next compared the CAI and MFE of the *Luciferase* mRNA designs (**Figure 1E; Supplementary Figure 1D-E**). The *mRNArchitect* and *MinCAI Luciferase* designs showed the highest (1.0) and lowest (0.62) CAI values, while mRNAid and LinearDesign had intermediate (0.965) and lower (0.734) values respectively. The MFE values were also variable, where the LinearDesign sequence had the lowest MFE (-1111.5), *MinCAI* had the highest (-351.9) and *mRNArchitect* and mRNAid designs showed intermediate MFE values (-685.3 and -593.0 respectively).

3.2. *Luciferase* mRNA

We hypothesized that the different *Luciferase* mRNA designs would exhibit distinct expression levels and stability. To test this hypothesis, we synthesized the mRNA constructs at the BASE mRNA facility using standard methods[14] and transfected them into HEK293-T cells using Lipofectamine™ MessengerMAX™ (Invitrogen™). Luciferase activity was measured at 24 and 48 hours using the Luciferase Assay System (Promega) (**Figure 1F**).

At 24 hours post-transfection, Luciferase expression peaked and varied significantly across the mRNA designs. The highest expression was observed for the *mRNArchitect* mRNA, while the lowest was recorded for the *MinCAI* design. The LinearDesign and mRNAid sequences exhibited intermediate expression levels. Specifically, *mRNArchitect* expression was 21-fold higher than *MinCAI*, sixfold higher than LinearDesign, and fivefold higher than mRNAid. Similar trends were observed at 48 hours, however expression of the *mRNArchitect* mRNA was now four-fold higher than the LinearDesign mRNA and 34-fold higher than the *MinCAI* mRNA.

3.3. *eGFP* mRNA

Next, we prepared *eGFP* mRNA designs using *mRNArchitect* to validate its performance on a different amino acid sequence. First, an mRNA was designed using the default *mRNArchitect* parameters and UD, yielding an mRNA design with a CAI of 0.9 and an MFE of -217.4 (**Supplementary Table 3**). Like the *Luciferase* optimisations (above), relaxing the *eGFP* design constraints (e.g. increasing GC content and avoiding UD), produced mRNA designs with progressively higher CAI and lower MFE values (**Supplementary Figure 2A-B**). The highest CAI

was observed for an *mRNArchitect* Maximum CAI *eGFP* design (termed "*mRNArchitect*" here on), that was prepared using the same method described for the *Luciferase* Maximum CAI design (above). Further *eGFP* mRNA designs were also prepared using mRNAid and LinearDesign, using the optimisation parameters described for Luciferase (above). All designs can be easily manufactured using DNA synthesis, except for the LinearDesign mRNA (**Supplementary Figure 2C**).

We compared the expression of eGFP for mRNA designs generated by *mRNArchitect*, mRNAid and LinearDesign (**Supplementary Figure 2D-F**). The mRNAs were synthesised at the BASE mRNA facility, and transfection followed the previously described methods (above). eGFP expression was assessed using flow cytometry (CytoFlex). At 24 hours post-transfection, there was no significant difference in the expression of the different mRNA designs. However, by 48 hours, significant differences emerged: the *mRNArchitect* mRNA exhibited the highest expression, the mRNAid mRNA the lowest, and the LinearDesign mRNA an intermediate level. The *mRNArchitect* mRNA expression was fivefold higher than mRNAid and twice as high as LinearDesign.

Together, our results show that by modulating the *mRNArchitect* optimisation parameters, users can generate candidate mRNA sequences with differing sequence and expression properties. We observe a trade-off between the manufacturability and performance of mRNAs designs, where mRNAs with higher CAI and lower MFE tend to be more highly expressed and more difficult to manufacture (**Figure 1G, Supplementary Table 1, 3**). Users can specify different sequence optimisation parameters to suit the required performance and manufacturability of mRNA sequences according to their intended application.

4. CONCLUSIONS

The optimisation of mRNA sequences can markedly improve their translation and stability, and reduce their reactogenicity. *mRNArchitect* can be used to design optimized mRNAs that encode a wide range of proteins, including vaccine antigens, peptides, antibodies, gene- and cell therapies. We have used *mRNArchitect* to design hundreds of mRNAs that have been experimentally validated across a wide range of applications. We provide *mRNArchitect* as a freely-available software toolkit for the assembly and optimization of mRNA sequences.

Our results show that mRNAs composed entirely of preferred human codons, optimised for maximum CAI, show high and sustained expression in human cell culture. This strategy is particularly effective for the design of mRNAs encoding short- to medium-length proteins. mRNAs composed of preferred codons are translated efficiently due to the high availability of complementary tRNAs, and tend to form stable secondary structures due to their high GC content (**Supplementary Figure 1B,E; 2B**)[15].

However, while a Maximum CAI optimisation supports robust translation of short- to medium-length proteins, it poses challenges for the manufacturability of mRNAs encoding longer proteins. It is therefore unlikely that a single mRNA sequence optimisation strategy will be universally appropriate. Instead, mRNA designs should consider the length and amino-acid composition of encoded proteins, and the intended application of the mRNA medicine.

The field of mRNA design is advancing, with the identification of new features and sequences that regulate translation rates, subcellular localization, and stability. We anticipate that ongoing improvements in mRNA design, combined with an increased understanding of mRNA biology within the cell, will further enhance these designs. In concert, we will continue to improve *mRNArchitect* by incorporating new elements and features, supporting scientists in designing innovative mRNA medicines.

Acknowledgements

The authors acknowledge the facilities and the scientific and technical assistance of the BASE mRNA Facility at The University of Queensland. BASE is supported by Therapeutic Innovation Australia (TIA). TIA is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS) program.

Supplementary data

Supplementary data are available online.

Conflict of interest

None.

Funding

We acknowledge the following sources of funding and support: National Health and Medical Research Council (GNT2014002 and GNT1161832) to T.R.M., Australian Research Council (DE230100036) to S.W.C., Medical Research Future Fund (MRFCRI000063 and MRFCRI000089) to S.W.C. and T.R.M., National Collaborative Research Infrastructure Strategy (NCRIS) to T.R.M., Therapeutic Innovation Australia (TIA) to S.W.C and T.R.M., Tour de Cure to S.W.C. and T.R.M., and The University of Queensland to S.W.C. and T.R.M.

Data availability

mRNArchitect is available through www.basefacility.org.au/software/ and can be downloaded for local installation from Github at https://github.com/BaseUQ/mRNArchitect.

Figure Legend

Figure 1. Design of mRNA medicines using *mRNArchitect*. (A) Sequence inputs. *mRNArchitect* accepts 5' and 3' UTR sequences, a coding sequence encoding the protein of interest (supplied as DNA, RNA or amino acid) and a poly(A) tail. **(B) Optimisation parameters.** CDS sequences are optimized to maximize CAI values within a given GC% range. Different sequence motifs can be avoided, including Restriction Enzyme binding sites or repeats which can both impede manufacture. Additionally, secondary structure is optimized according to Minimum Free Energy (MFE) calculations. **(C) Outputs.** mRNA candidate sequences are provided with an accompanying output table summarising the properties of the mRNA designs, which can be plotted to assist in sequence selection. **(D) Sequence alignment.** Multiple Sequence Alignment and sequence logo plot, showing the similarity between *Luciferase* mRNA designs generated by *mRNArchitect*, mRNAid and LinearDesign. **(E) Codon optimality and mRNA secondary structure.** Scatter plot comparing CAI and MFE of *Luciferase* mRNA designs generated by *mRNArchitect*, mRNAid and LinearDesign. **(F) mRNA expression.** Relative expression of different *Luciferase* mRNA designs was assayed in HEK293 cells at 24h and 48h post transfection. Results were analysed using One-way ANOVA (****= $p<0.0001$). **(G) Template manufacture.** *mRNArchitect* designs include mRNAs with a range of CAI and MFE values, which impact manufacturability and expression.

Supplemental Figure legends

Supplementary Figure 1. *Luciferase* mRNA sequences generated using different mRNA design tools and parameters. (A-C) Different performance and manufacturability characteristics of *Luciferase* mRNA designs generated by *mRNArchitect*, with different CDS input sequences and optimisation parameters. These include mRNAs designed to meet stringent DNA template manufacturing requirements (Different inputs), without Uridine Depletion (No U depletion), with a maximum GC content of 70% (Max 70% GC), or 80% (Max 80% GC), or mRNAs designed to without any further design constraints, to maximise CAI (*mRNArchitect*), or with a rare codon model (*MinCAI*). (**D-F**) Different performance and manufacturability characteristics of *Luciferase* mRNA designs generated by different software including *mRNArchitect*, mRNAid and LinearDesign, and the *Luciferase* Wild Type sequence (*P. pyralis*). **(G)** Percent identity matrix from Multiple Sequence Alignment of *Luciferase* mRNA sequences.

Supplementary Figure 2. *eGFP* mRNA sequences, generated using different mRNA design tools and parameters. (A-C) Different performance and manufacturability characteristics of *eGFP* mRNA designs generated by *mRNArchitect*, using different optimisation parameters. These include mRNAs designed to meet stringent DNA template manufacturing requirements (low complexity), without Uridine Depletion (No U Depletion), with a maximum GC content of 70% (Max 70% GC), or 80% (Max 80% GC), or mRNAs designed to without any further design constraints, to maximise CAI (*mRNArchitect*). Further *eGFP* mRNA designs were prepared using mRNAid and LinearDesign. **(D)** Expression of different eGFP mRNA designs in HEK293 cells, determined by Flow cytometry, at 24h and 48h post transfection. Results were analysed using One-way ANOVA (ns=$p>0.05$; *=$p<0.05$; **=$p<0.01$; ****=$p<0.0001$). **(E)** Fluorescence imaging of HEK293 cells transfected with *mRNArchitect* eGFP mRNA. **(F)** eGFP mRNA designs showed similar transfection rates in HEK293 cells, determined by Flow cytometry. Results were analysed using One-way ANOVA (****=$p<0.0001$).

Supplementary Tables

Supplementary Table 1. *Luciferase* mRNA designs.

Supplementary Table 2. Codon model used for *MinCAI Luciferase* mRNA design.

Supplementary Table 3. *eGFP* mRNA designs.

References

1 Vostrosablin, N. *et al.* mRNAid, an open-source platform for therapeutic mRNA design and optimization strategies. *NAR Genomics and Bioinformatics* **6**, lqae028 (2024).

2 Zhang, H. *et al.* Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**, 396-403 (2023).

3 Holcomb, D. D., Alexaki, A., Katneni, U. & Kimchi-Sarfaty, C. The Kazusa codon usage database, CoCoPUTs, and the value of up-to-date codon usage statistics. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases* **73**, 266-268 (2019).

4 Karikó, K., Buckstein, M., Ni, H. & Weissman, D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* **23**, 165-175 (2005).

5 Pichlmair, A. *et al.* Activation of MDA5 requires higher-order RNA structures generated during virus infection. *Journal of virology* **83**, 10761-10769 (2009).

6 Lee, J., Woodruff, M. C., Kim, E. H. & Nam, J.-H. Knife's edge: Balancing immunogenicity and reactogenicity in mRNA vaccines. *Experimental & Molecular Medicine* **55**, 1305-1313 (2023).

7 Kiaie, S. H. *et al.* Recent advances in mRNA-LNP therapeutics: immunological and pharmacological aspects. *Journal of nanobiotechnology* **20**, 276 (2022).

8 Trepotec, Z., Geiger, J., Plank, C., Aneja, M. K. & Rudolph, C. Segmented poly (A) tails significantly reduce recombination of plasmid DNA without affecting mRNA translation efficiency or half-life. *Rna* **25**, 507-518 (2019).

9 Zulkower, V. & Rosser, S. DNA Chisel, a versatile sequence optimizer. *Bioinformatics* **36**, 4508-4509 (2020).

10 Mulroney, T. E. *et al.* N 1-methylpseudouridylation of mRNA causes+ 1 ribosomal frameshifting. *Nature* **625**, 189-194 (2024).

11 Kames, J. *et al.* TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. *Journal of molecular biology* **432**, 3369-3378 (2020).

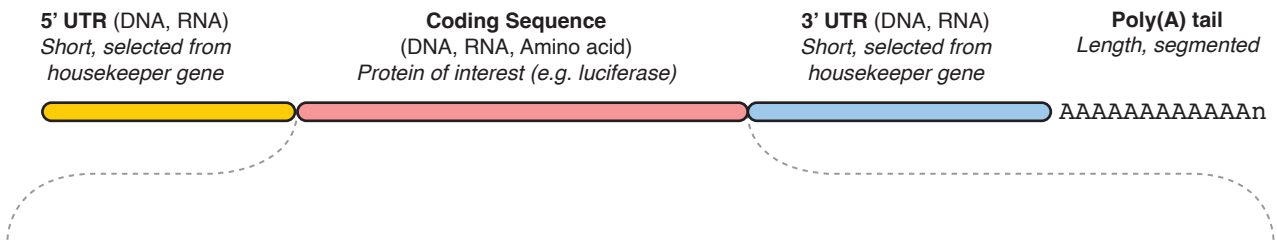
12 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for molecular biology* **6**, 26 (2011).

13 De Wet, J. R., Wood, K. V., DeLuca, M., Helinski, D. R. & Subramani, S. Firefly luciferase gene: structure and expression in mammalian cells. *Molecular and cellular biology* **7**, 725-737 (1987).

14 Gunter, H. M. *et al.* mRNA vaccine quality analysis using RNA sequencing. *Nature Communications* **14**, 5663 (2023).

15 Trotta, E. On the normalization of the minimum free energy of RNAs by sequence length. *PloS one* **9**, e113380 (2014).

a. Sequence Input



b. Optimisation Parameters

Codon optimisation

Amino acid	Met	Pro	Glu	Tyr
Possible codon	AUG	CCG	GAA	UAU
		CCU	GAG	UAC
		CCC		
		CCA		
Optimised codon	AUG	CCG	GAA	UAU

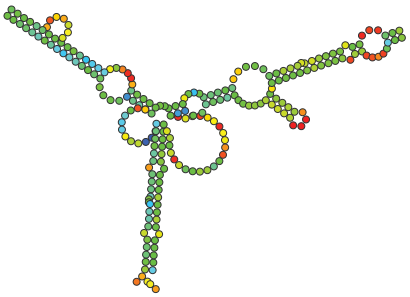
Avoid sequences

Amino acid	Phe	Leu	Arg	Ser	Leu
Avoid *BsaI*	TTC	CTA	CGG	TCT	CTG
Optimised codon	TTC	CTA	CGG	TCG	CTG

Avoid repeats

Amino acid	Phe	Leu	Arg	Phe	Leu	Arg
Avoid repeat	TTC	CTA	CGG	TTC	CTA	CGG
Optimised codon	TTC	CTG	CGG	TTC	TTA	CGG

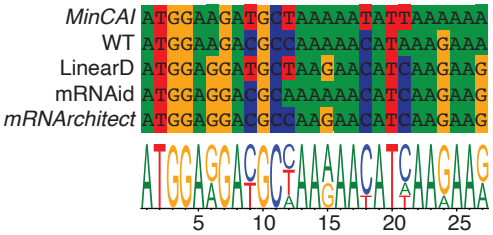
Secondary structure



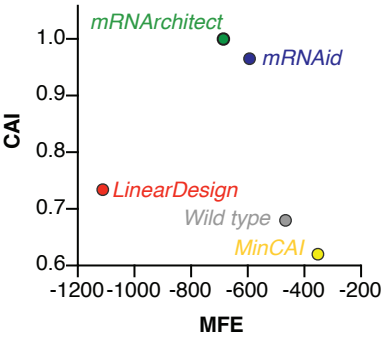
c. Output mRNA Statistics

	Input	Optimised
A ratio	0.20	0.24
T/U ratio	0.14	0.20
G ratio	0.31	0.28
C ratio	0.34	0.28
AT ratio	0.35	0.44
GA ratio	0.52	0.52
GC ratio	0.65	0.56
Uridine depletion	0	0.15
CAI	1	0.89
CDS MFE (kcal/mol)	-690.7	-596.6

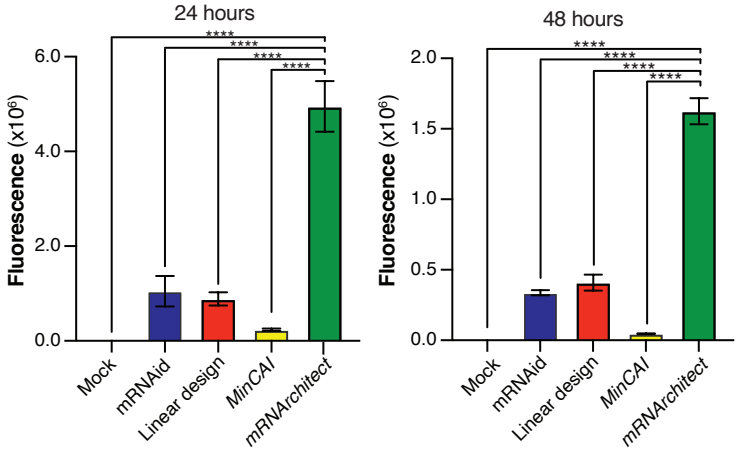
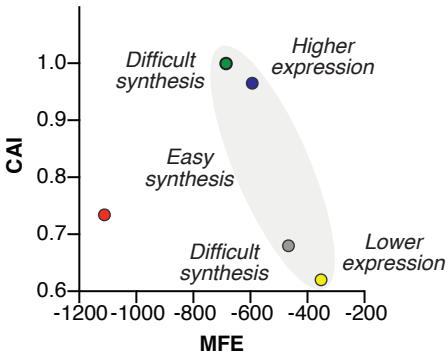
d. Multiple Sequence Alignment of mRNA designs



e. Codon optimality and secondary structure of mRNA designs



f. Expression of mRNA designs



g. Template Manufacture