# *mRNArchitect*: optimized design of mRNA sequences

Magdalena A. Budzinska[1,2], Theodore E. Leonard[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**

**Summary: mRNA design is the critical first step in the development of a new vaccine or therapy. The mRNA primary sequence is assembled from multiple elements, including the coding sequence of the target protein or gene, flanking untranslated regions that enhance translation, and the poly(A) tail, which improves mRNA stability. The coding sequence can also be optimized to improve translation and stability, and the depletion of uridines and double-stranded RNA secondary structures can reduce reactogenicity. Here, we introduce *mRNArchitect*, a software toolkit to assist in the design of mRNA sequences according to user requirements. *mRNArchitect* can rapidly assemble and optimize mRNA sequences based on criteria including GC content, secondary structure, codon optimization, and uridine depletion. The sequences generated by *mRNArchitect* can also be readily synthesized into DNA templates, and have been validated across a variety of applications. We offer *mRNArchitect* as an open-source 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

The design of an mRNA primary sequence is the critical first step in developing a new vaccine or therapy. Optimized sequence design can markedly improve the performance of mRNA by enhancing translation and protein expression, increasing mRNA stability, and reducing reactogenicity[1,2].

Any coding sequence (CDS) can be encoded in mRNA, from small peptides to large structural proteins. The CDS is reverse translated into an mRNA sequence using codons that have been optimized for high expression in the target organism or tissue[3]. The CDS may also be depleted of uridines to avoid detection by toll-like receptors in the endosome, evading the innate immune response[4].

The coding sequence is flanked by untranslated regions (UTRs) that mediate protein expression. The ribosome binds to the 5' UTR and scans for the methionine start codon. The 3' UTR also contains binding sites for proteins and microRNAs and is crucial for cell-specific expression and stability.

The mRNA also folds into secondary structures that can interact with proteins and can further extend the half-life of the mRNA but can also bind to dsRNA receptors that activate the innate immune response[5]. The mRNA sequence ends with a poly(A) tail that loops back to enhance translation initiation and protects against ongoing 3' exonuclease degradation.

In addition to these performance considerations, mRNA sequences should also be compliant with manufacturing requirements. Ideally, the output mRNA sequences should be readily synthesized into DNA templates that produce high mRNA yields during *in vitro* transcription.

Different mRNA design strategies may be preferred depending on the final application. For example, mRNA vaccine sequences that include 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 mRNAs for different applications, we have developed *mRNArchitect*, an open-source software toolkit that can assemble and optimize mRNA sequences based on user specifications. *mRNArchitect* can optimize the expression, stability and immunogenicity of mRNA sequences for manufacture. *mRNArchitect* has been extensively used to design mRNA for a wide range of applications, and is freely available for use by researchers.

## 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 (**see Figure 1A**) is supplied. 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).

The user 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 wide range of cell types and applications. Finally, the user 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**). Design parameters can be specified and include 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 constraints score is maximized.

The parameters can be modified according to the user's requirements and often depend on the target organism, tissue and intended application. To enhance translation, the mRNA sequence is optimized to use preferred codons of the specified target organism (e.g. human or mouse).

*mRNArchitect* also allows uridine depletion (UD), with the exclusion of codons with uridine at wobble positions. Difficult sequence motifs, such as motifs associated with ribosomal slippage, can also be avoided. *mRNArchitect* also allows users to impose a minimum and maximum GC fraction across a defined window, with the formation of stable RNA secondary structure and GC content being key variables that impact stability.

In addition to these functional considerations, it is necessary for mRNA sequences to be easily manufactured. *mRNArchitect* allows users to avoid restriction enzyme binding sites, hairpins and low-complexity, repetitive sequences that may impede DNA template manufacture and mRNA *in vitro* transcription.

### 2.3. Output mRNA sequences

The *mRNArchitect* output provides fully assembled and optimized mRNA sequence candidates ready for synthesis (**Figure 1C**). In addition, these candidate sequences are accompanied by statistics describing the features of the optimized sequence.

Reported statistics include the Codon Adaptation Index (CAI) score, which measures how closely the codon usage of the sequence matches the preference of a particular organism or tissue[3,10]. 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[11].

## 3. RESULTS

### 3.1. Running *mRNArchitect*

To demonstrate the use of *mRNArchitect*, we designed an mRNA encoding the Luciferase protein[12]. Sequence inputs were based on a modified version of the *P. pyralis* Luciferase amino acid sequence (GenBank: AB644228.1), and 5' and 3' UTRs based on human alpha-globin (*HBA1*; Gene ID 3039) and a 120 nt poly(A) tail with no linker (**Supplementary Table 1**). The mRNA sequence was optimized using the human codon model[3] with UD, a GC content of 0.4 – 0.6 across a GC content window of 60 nt. We avoided *BsaI*, *Xba* and *XhoI* cleavage sites that are used in plasmid-based DNA template manufacture, and to optimize DNA template manufacturability, we avoided repeat lengths of 10 nt or more, as well as repeats of 9 nt poly(U) or poly(A), and repeats of 6 nt poly(C) or poly(G). Hairpins with a stem size of 10 nt or more were avoided across a window size of 60 nt.

The resulting mRNA sequence was compared to the original *P. pyralis* wild-type *Luciferase* nucleotide sequence (AB644228) showing 74% similarity (**Supplementary Figure 1A,B**). In comparison to the wildtype *Luciferase* sequence, the optimized mRNA sequence showed a reduced uridine content (from 26% to 16%), increased GC content (from 45% to 56%), a reduction in the MFE (from -466.4 to -534.2), and an increase in CAI (from 0.68 to 0.859).

### 3.2. Benchmarking *mRNArchitect* outputs

We next benchmarked *mRNArchitect* against a range of publicly-available mRNA design tools, including mRNAid and LinearDesign[1,2]. 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 resulting mRNA designs were then compared to the *mRNArchitect Luciferase* mRNA design described above (**Figure 1D**). As a control, we also included a 'rare-codon' *Luciferase* mRNA sequence with low CAI values that maximizes uridine at the third position (**Supplementary Table 2**).

The *mRNArchitect Luciferase* design showed the highest similarity to the mRNAid design (89%) then LinearDesign (78%) and the rare codon design (66%) (**Supplementary Figure 1A**). With the exception of *mRNArchitect*, all other designs required specialized DNA synthesis due to difficult and repetitive sequences (**Supplementary Figure 1E**).

We next compared the predicted translation and stability of the *Luciferase* mRNA designs (**Figure 1D; Supplementary Figure 1C-D**). The *mRNArchitect* design had an intermediate CAI (0.859) and a relatively high MFE (-0.534.2) in comparison to mRNAid design which has a higher CAI (0.965) and lower MFE (-593.0), while the LinearDesign sequence had a lower CAI (0.734) and highest MFE (-1111.5). The control 'rare codon' design had the lowest CAI (0.62) and highest MFE (-351.9). These predicted differences in translation dynamics were empirically confirmed when the different *Luciferase* mRNA designs were transfected into HEK293-T cells, using Lipofectamine™ MessengerMAX™ (Invitrogen™). Here, we observed a linear relationship between the CAI values of each mRNA and their corresponding Luciferase protein expression at 4 hours post transfection (**Supplementary Figure 2G**).

### 3.3. Benchmarking *mRNArchitect* parameters

*mRNArchitect* has default parameters that have been selected to provide optimal, low complexity mRNA sequences that can be readily manufactured into DNA templates. However, these parameters can be modulated to effect translation, stability and manufacturability of different mRNA sequences.

We next examined the relationship between the optimisation parameters, and their impact on output mRNA sequences. Different input sequences encoding the same amino acid sequence had a limited impact on the CAI or MFE of output sequences, when identical design parameters were used (**Supplementary Figure 1F-H**). Varying the design parameters had a considerably larger impact on the output mRNA sequence properties. *Luciferase* mRNAs designed 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 Figure 1I-J**). Increasing the maximum GC content to 70% or 80% further increased CAI and decreased MFE values, however designs with 90% or 100% maximum GC content did not show further improvements to CAI or MFE values (**Supplementary Table S3**). Furthermore, mRNA designs with high (>60%) GC content were difficult to manufacture with DNA synthesis (**Supplementary Figure 1K**). Finally, relaxing all repeat and GC content constraints resulted in designs with a CAI of 1.

Open-reading frame length and amino acid composition can also impact output mRNA sequence composition. We used *mRNArchitect* to design mRNAs for the COVID-19 spike protein (Wuhan sequence, QJF75467.1) and compared them to the Comirnaty® (BioNTech-Pfizer) and Spikevax® (Moderna) COVID-19 mRNA vaccines (**Supplementary Table 4**). The *mRNArchitect* designs cover a range of MFE and CAI values depending on their optimisation parameters, including designs with a lower MFE and higher CAI than both the Comirnaty and Spikevax sequences (**Supplementary Figure 2A-C**).

We also evaluated *mRNArchitect*'s performance for designing mRNAs that encode shorter protein sequences. The *mRNArchitect* eGFP mRNA was designed to a modified eGFP amino acid sequence, using default mRNA design parameters and UD (**Supplementary Table 4**). As described above, relaxing the mRNA design constraints (e.g. higher GC content and without UD), led to progressively higher CAI and lower MFE values for our eGFP mRNA designs (**Supplementary Figure 2D-F**).

Together, our results show that by modulating the *mRNArchitect* optimisation parameters, users can generate candidate mRNA sequences with differing properties. We observe a trade-off between the

manufacturability and performance of mRNAs designs, where mRNAs with higher predicted expression and stability (i.e. higher CAI and lower MFE) tend to be more difficult to manufacture (Figure 1E). Users can specify different optimisation parameters to suit the required performance and manufacturability of mRNA sequences according to their intended application, and the unique properties of the amino acid sequence they wish to encode.

## 4. CONCLUSIONS

The optimisation of mRNA sequences can markedly improve their translation, stability, and 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.

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 parallel, 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**

T.R.M. H.M.G. and S.W.C. have received research funding, support for conference attendance, travel and accommodation from Oxford Nanopore Technologies. T.R.M. and S.W.C. have received research funding from Sartorius Stedim Australia, and Sanofi. T.R.M., H.M.G, R.Y.C., N.C. and S.W.C. have received support for conference attendance, travel and accommodation from Moderna. The other authors declare no competing interests.

**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.** *mRNArchictect* 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) Expression and RNA secondary structure.** CAI and MFE of *Luciferase* mRNAs was plotted for an *mRNArchitect* design, the *Luciferase* Wild Type sequence (*P. pyralis*) and different *Luciferase* mRNA designs generated by different software. **(E) Manufacture of difficult mRNA.** *mRNArchitect* designs include mRNAs with a range of CAI and MFE values, which impact manufacturability and expression. Through varying the optimisation parameters, users can identify designs that meet their manufacturability and expression requirements.

**Supplemental Figure legends**

**Supplementary Figure 1. *Luciferase* mRNA designs prepared using different tools and parameters. (A)** Bootstraped ML dendrogram shows the sequence similarity between mRNA designs generated by *mRNArchitect*, mRNAid, LinearDesign, and the rare codon design to the WT sequence. **(B)** Percent identity matrix from alignment generated by ClustalO. **(C-E)** Different performance and manufacturability characteristics of *Luciferase* mRNA designs generated by different software including *mRNArchitect*, mRNAid, LinearDesign, and the rare codon design and the *Luciferase* Wild Type sequence (*P. pyralis*). **(F-H)** Different performance and manufacturability characteristics of *Luciferase* mRNA designs generated by *mRNArchitect*, with different CDS input sequences alongside *Luciferase* mRNA designs generated by different software, and the *Luciferase* Wild Type sequence (*P. pyralis*). **(I-J)** Different performance and manufacturability characteristics of *Luciferase* mRNA designs generated by *mRNArchitect*, using different design constraints.

**Supplementary Figure 2. COVID-19 spike and eGFP mRNA designs, prepared using different tools and parameters. (A-C)** Different performance and manufacturability characteristics of Spikevax and Cominarty COVID-19 spike, compared to mRNA designs generated by *mRNArchitec*t using different design constraints. These include mRNAs designed to meet stringent DNA template manufacturing requirements (low complexity), without Uridine Depletion (No UD), with a maximum GC content of 70% (Max 70% GC), or 80% (Max 80% GC), or mRNAs designed to maximise CAI, without any further design constraints (Max CAI). **(D-F)** Different performance and manufacturability characteristics of eGFP mRNA designs generated by *mRNArchitect*, using different design constraints. These include mRNAs designed to meet stringent DNA template manufacturing requirements (low complexity), without Uridine Depletion (No UD), with a maximum GC content of 70% (Max 70% GC), or 80% (Max 80% GC), or mRNAs designed to maximise CAI, without any further design constraints (Max CAI).

**Supplementary Tables**

**Supplementary Table 1. *Luciferase* mRNA designs generated by *mRNArchitect*, mRNAid and LinearDesign.**

**Supplementary Table 2. Codon model used for the rare codon *Luciferase* mRNA design.**

**Supplementary Table 3. *Luciferase* mRNA designs generated by *mRNArchitect* using different optimisation parameters.**

**Supplementary Table 4. COVID-19 spike protein and eGFP mRNA designs generated by *mRNArchitect* using different optimisation parameters.**

**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 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).

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

12 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).

## A Sequence Input

**5'UTR** (DNA, RNA)
*Short, selected from highly expressed gene*

**Coding Sequence**
(DNA, RNA, Amino acid)
*Protein of interest (e.g. luciferase)*

**3'UTR** (DNA, RNA)
*Short, selected from highly expressed gene*

**Poly(A) tail**
*Length, segmented*

## 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 |

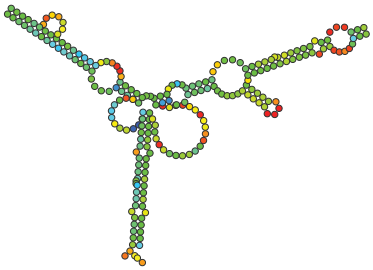
### Avoid sequences

| 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 |

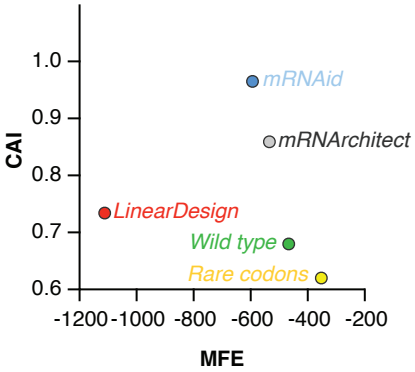
### Secondary structure



## C Output mRNA sequence 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 |
| 5'UTR/3'UTR MFE | -10.2/-29.5 | -10.2/-29.5 |

## D Predicted stability and translation of mRNA designs




## E Template manufacture

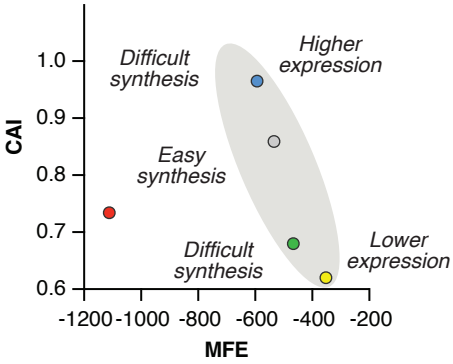



**Figure 1.**