Computational and Structural Biotechnology Journal 23 (2024) 3050–3064

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

A statistical-physics approach for codon usage optimisation

David Luna-Cerralbo [a,b], Irene Blasco-Machín [c], Susana Adame-Pérez [c], Verónica Lampaya [c], Ana Larraga [c], Teresa Alejo [c], Juan Martínez-Oliván [c], Esther Broset [c,*], Pierpaolo Bruscolini [a,b,**]

[a] *Department of Theoretical Physics, Faculty of Science, University of Zaragoza, c/ Pedro Cerbuna s/n, Zaragoza, 50009, Spain*
[b] *Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, c/ Mariano Esquillor s/n, Zaragoza, 50018, Spain*
[c] *Certest Pharma, Certest Biotec S.L, Polígono Industrial Río Gallego II, Calle J, 1, San Mateo de Gállego, 50840, Spain*

ARTICLE INFO

Keywords:
Statistical-physics model
Nearest-neighbour interaction
Codon-optimisation
Protein expression
mRNA-vaccine

ABSTRACT

The concept of “codon optimisation” involves adjusting the coding sequence of a target protein to account for the inherent codon preferences of a host species and maximise protein expression in that species. However, there is still a lack of consensus on the most effective approach to achieve optimal results. Existing methods typically depend on heuristic combinations of different variables, leaving the user with the final choice of the sequence hit. In this study, we propose a new statistical-physics model for codon optimisation. This model, called the Nearest-Neighbour interaction (NN) model, links the probability of any given codon sequence to the “interactions” between neighbouring codons. We used the model to design codon sequences for different proteins of interest, and we compared our sequences with the predictions of some commercial tools. In order to assess the importance of the pair interactions, we additionally compared the NN model with a simpler method (Ind) that disregards interactions. It was observed that the NN method yielded similar Codon Adaptation Index (CAI) values to those obtained by other commercial algorithms, despite the fact that CAI was not explicitly considered in the algorithm. By utilising both the NN and Ind methods to optimise the reporter protein luciferase, and then analysing the translation performance in human cell lines and in a mouse model, we found that the NN approach yielded the highest protein expression *in vivo*. Consequently, we propose that the NN model may prove advantageous in biotechnological applications, such as heterologous protein expression or mRNA-based therapies.

1. Introduction

The process of transcribing genetic information, encoded within the 4-letter code of DNA, into the elaborate 20-letter alphabet that constitutes proteins is orchestrated by “codons” i.e., sequences of three neighbouring nucleotides that encode an amino acid. There are 64 different codons but only 20 amino acids, causing degeneracy of the genetic code, since the same amino acid corresponds to different codons. This degeneracy results in many possible ways of encoding the same protein. Nonetheless, not all potential codons of an amino acid possess an equal likelihood of occurring in nature, as certain synonymous codons are more frequently employed to represent a specific amino acid in a particular organism than others. This trend is referred as codon usage bias [1]. Furthermore, the utilisation of sequential codon pairs is not random [2] and exhibits unique patterns specific to each species. This phenomenon is referred to as codon-pair usage bias [3] and cannot be related simply to individual codon bias.

Species-dependent codon bias is an important factor to consider in the framework of biotechnological applications where *in vivo* maximisation of protein expression is desired, as is the case, for instance, of mRNA vaccine design. However, no clear and reliable recipe is known to reach this goal, since the connection between gene translation and protein expression *in vivo* is not straightforward [4]. Rather, it is mediated by complex processes [5] that include cell-type dependent expression preferences [6,7] and potentially other mechanisms at the organism level (e.g. immune response).

A reasonable starting point when designing a coding sequence for heterologous expression is to assume that the use of codons which

* Corresponding author.
** Corresponding author at: Department of Theoretical Physics, Faculty of Science, University of Zaragoza, c/ Pedro Cerbuna s/n, Zaragoza, 50009, Spain.
E-mail addresses: dluna@unizar.es (D. Luna-Cerralbo), ibmachin@certest.es (I. Blasco-Machín), sadame@certest.es (S. Adame-Pérez), vlampaya@certest.es (V. Lampaya), alarraga@certest.es (A. Larraga), talejo@certest.es (T. Alejo), jmartinez@certest.es (J. Martínez-Oliván), ebroset@certest.es (E. Broset), pier@unizar.es (P. Bruscolini).

https://doi.org/10.1016/j.csbj.2024.07.020
Received 14 May 2024; Received in revised form 25 July 2024; Accepted 25 July 2024
Available online 30 July 2024

are more frequently observed in the host organism, rather than those rarely observed, would increase the translation efficiency and the level of protein expression. However, a hyper-optimisation of codon usage has shown to be detrimental for the desired protein expression [8] or function [9]. Remarkably, several suboptimal codons or pairs are often necessary for proper protein function [10–12]. This counterintuitive fact may be related to the appearance of translational errors when only the most frequent codons for each amino acid are used, something that is believed to be related to the imbalanced utilisation and consequent unavailability of a subset of tRNAs [13]. Conversely, the use of suboptimal codons at crucial positions could induce a slowdown in translation, preventing the misfolding of the nascent protein [14,4,15]. Codon usage might also be controlled by the need to avoid an excess of stability of the secondary structure at the 5'-end, which would hinder the attachment of the ribosomes [16,17]. These considerations suggest that a conservative strategy for codon optimisation should rely on choosing patterns of codon usage that reproduce that of the host organism.

Regarding codon optimisation tools, many publicly available ones primarily are based on the utilisation of the most common codon for a specific amino acid [18–21], even if they often give the possibility to generate alternative sequences based on the natural codon frequencies in preselected, and often user-specified, databases [20–22]. Thus, several steps of optimisation are usually needed to refine the raw predictions of these tools, such as mutations on the target sites of common restriction enzymes, elimination of repeating sequences, and tuning of extreme GC content regions. More advanced optimisation tools are available at the web pages of commercial suppliers specialised in gene synthesis. However, in such cases, the internal algorithm is generally undisclosed in its details, and there is limited flexibility for fine-tuning the sequence.

Recently, new algorithms have been developed and made publicly available. Some of them focus on multi-objective (Pareto) optimisation, following the path of several reports [23–25] and apply Mixed Integer Linear Programming approaches [26], as well as more traditional Dynamic Programming [27], to optimise different objective functions related to the quality of the codon design (CAI, CPB, CPS, RCB, RCPB, among others (see the precise definitions in Methods 5.4). These approaches extend the focus beyond the single codon frequencies, accounted for in CAI, to include codon context, i.e., the frequencies of codon pairs; however, they leave the user with the decision on what weight to give to each quality indicator. Other proposals tackle the problem using Neural Networks [28], that may be very efficient, but, as a black-box, do not allow to understand the relevance of different features and its relation to the different biological facets of the codon selection problem.

Here, we propose to tackle codon design using a different angle, which does not aim at finding the optimum result of any objective function. Instead, we implement a probabilistic approach, inspired by simple models of statistical physics, where the probability of any codon sequence for a given protein is expressed as a Boltzmann probability. Thus, it depends fundamentally on two ingredients: 1) an “energy function”, accounting for single site codon preferences and “interactions” between neighbouring codon, and 2) a “temperature”, that increases the probability of accepting solutions that departs from the one that minimises the energy functions. In our approach, codon and codon-pair biases are introduced in an integrated framework and the parameters describing the interaction between neighbouring codons are learnt upon maximising the probability of the whole codon sequence database. This allows, in principle, to account for finer details and correlations that could be disregarded in the usual approach, as it just focuses on precalculated codon and pair frequencies.

In the following sections of the paper, a brief introduction of the model and its principles is given, followed by the results obtained upon designing several codon sequences for different proteins. We assessed the model's effectiveness using both *in silico* and experimental tests. Firstly, through *in silico* analysis, we compared the predicted codon sequences to those generated by a simplified version of the model in which pair interactions are omitted, thereby elucidating the significance of such interactions. We also compared our sequences with the codon optimisations proposed by some commercial tools and analysed the role of the temperature parameter in reproducing different scenarios. Secondly, we focused on the firefly luciferase protein, and redesigned its codon sequence under different temperature conditions and using the NN and Ind models, and the resulting sequences were tested experimentally. Thus, we tested luciferase optimisations both *in vitro*, on HeLa and HepG2 cell lines, and *in vivo*, on a mouse model, with the aim of establishing a criterion relating the design protocol and the protein expression. An overall scheme for the workflow, from model definition to testing on luciferase, is presented in Fig. 1

2. Results

2.1. Definition of the method

We build our model on a few basic assumptions:

1. The observed codon sequences $S(P)$ for a given protein P and species $\mathcal{A}$ are random events, associated with a probability of the form $p(S(P)|P,\mathcal{A}) \propto e^{-\beta\mathcal{H}(S(P)|\mathcal{A})}$ at $\beta = 1$ (see Eq. (2) in Methods). That is, we introduce a Boltzmann probability associated to an energy function $\mathcal{H}(S(P)|\mathcal{A})$, with species-dependent parameters; in the following, since we will always work with one species at a time, we drop the indication of the species $\mathcal{A}$.
2. In principle, the energy function above implies interactions between codons at arbitrary positions along the sequence; however, we will limit ourselves to interactions between neighbouring codons $(i, i+1)$. These interactions need not be regarded as true, physical interactions, like, e.g., Watson-Crick couplings. Instead, they could be indirect interactions, such as mediated by each codon's contact with the ribosome; we do not enter into a microscopic justification of the form and origin of such energy function, but simply assume its existence.
3. The variable $\beta = 1/T$ plays the role of an inverse temperature, and its value modulates the number of codon sequences that are acceptable to represent the protein P: at $\beta = 0$, all possible codon sequences $S(P)$ of protein P have the same probability; on the other hand, $T = 0$ will select only the codon sequence of minimum energy, while the probability of all other sequences will be zero. By construction, $T = 1$ will correspond to the distribution of the natural sequences, and will represent our “learning temperature”, at which the model parameters are adjusted to reproduce the observed sequence probabilities, as explained in Section 5.2.3.

The possibility to design codon sequences at different temperatures, using Monte Carlo simulations, allows us to study the interplay between the temperature parameter and the characteristics of the resulting designed sequences.

2.2. *In silico* tests of the Ind and NN models

2.2.1. Codon usage

We first compared the codon and nucleotide usage as predicted by the full model with nearest-neighbour interactions (“NN”) and the individual codons model (“Ind”), as described in Sections 5.2.1, 5.2.2, to see how the introduction of interactions between neighbours affects the results. Fig. 2 shows the frequency of codon usage (normalised so that the sum over codons of the same amino acid is 1).

We noticed that, at $\beta = 1$, the Ind model does not accurately reproduce the codon usage for all amino acids (with notable deviations for C, H, K, L, T, and V), implying that a model in which neighbouring codons have independent probabilities, cannot reproduce natural sequences, even when the probability of each codon is learned from the

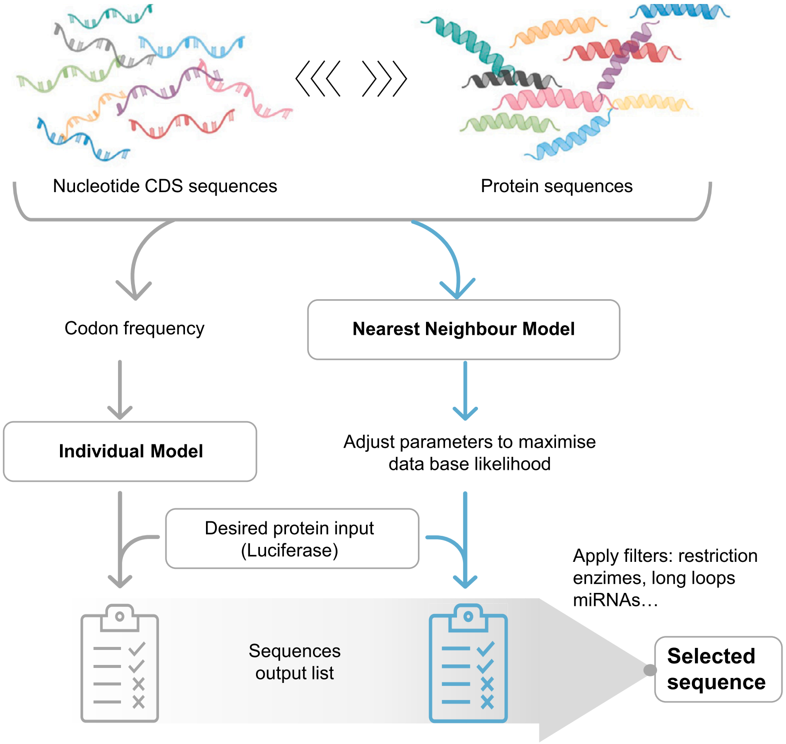
Nucleotide CDS sequences
Protein sequences
Codon frequency
Nearest Neighbour Model
Individual Model
Adjust parameters to maximise data base likelihood
Desired protein input (Luciferase)
Apply filters: restriction enzimes, long loops miRNAs…
Sequences output list
Selected sequence

Fig. 1. Scheme of the construction of the models. Workflow of the Nearest-Neighbour interaction (NN) model compared to the Individual codons (Ind) model. Reference nucleotide coding sequences (CDS) from the human genome were extracted and matched to their corresponding amino acid sequences. For the Individual model (left panel), the frequency of each codon was used to fit its energy, with $\beta = 1$, from Eq. (10). For the NN model (right panel), the likelihood of the natural codon sequence database was maximised at $\beta = 1$ to determine the model parameters. For the case of luciferase as a model protein, codon sequences are generated with the Ind and NN models at different values of β. Such sequences are subjected to a final filtering process based on various criteria, such as sequence restriction and the absence of long loops. The aim of this refinement step is to obtain the most suitable optimised sequence for the target organism.

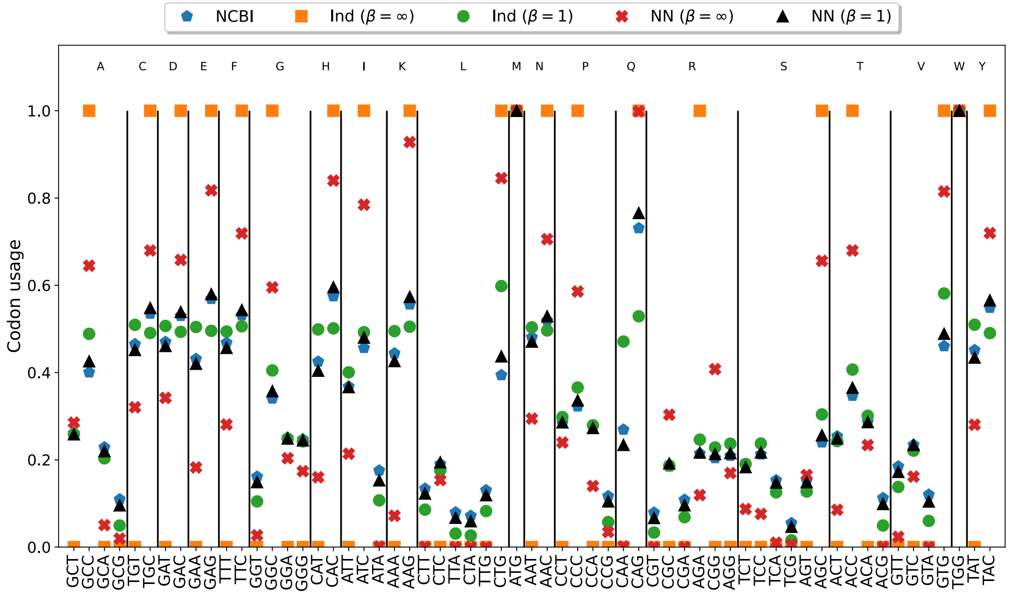

Fig. 2. Codon usage: NCBI vs Ind and NN algorithms with different β values. Human-observed (NCBI) codon usage, and predictions with the NN and Ind models, at $\beta = 1$ (the "learning" inverse temperature, see Methods 5.2.3) and at $\beta = \infty$ (minimum energy solution). The NCBI codon usage line is derived by extracting codon usage information from the 116,487 sequences available in the NCBI database. For the NN and Ind designs at $\beta = 1$, codon usage is determined by considering a randomly selected dataset of 10,000 protein sequences from NCBI, and selecting one codon sequences for each of them, from a thermalised Monte Carlo run, according to the probabilities Eqs. (2), (8), respectively. In the case of $\beta = \infty$, another set of 10,000 sequences is randomly selected from NCBI, and the sequences with the minimum energy are determined for both the NN and Ind' models, see Methods 5.3.

natural NCBI database. Additionally, we observed that the NN model can predict the frequencies of each codon well enough that the lines of NCBI and NN overlap in Fig. 2, except for the amino acid C, where small differences are noticeable. At $\beta = \infty$ ($T = 0$), it can be seen that the Ind model resorts to using only the most likely codon for each amino acid, as expected. However, this is not the case for NN, due to the influence

Table 1
The abbreviated protein name, the full protein name, the organism, and the length in amino acids (aa) of the sequences used.

Short name	Protein Name	Organism	Length
LUC	Luciferase	Allobacillus saliphilus	551
GFP	Green fluorescent protein	Aequorea victoria	239
DsRed	Red fluorescent protein	Discosoma sp	226
bGAL	beta galactosidase	Escherichia coli str. K-12	1025
Cas9	Nuclease cas9	Streptococcus pyogenes	1369
Cas13d	Nuclease cas13	Ruminococcus flavefaciens	968
OVA	Ovoalbumin	Gallus gallus	387
CRE	Cre recombinase	Escherichia phage P1	344
ADIPOQ	Adiponectin	Homo sapiens	245
BDNF	Brain Derived Neurotrophic Factor	Homo sapiens	248
CNTF	Ciliary Neurotrophic Factor	Homo sapiens	201
EGF	Epidermal Growth Factor	Homo sapiens	1208
FGF-4	Fibroblast Growth Factor-4	Homo sapiens	207
IL-1RA	Interleukin-1 Receptor Antagonist	Homo sapiens	178
TNF-α	Tumor Necrosis Factor-alpha	Homo sapiens	234
VEGF-D	Vascular Endothelial Growth Factor	Homo sapiens	355
EREG	Epiregulin	Homo sapiens	170
EPO	Erythropoietin	Homo sapiens	194
IL-4	Interleukin-4	Homo sapiens	154
HGF	Hepatocyte Growth Factor	Homo sapiens	729
IL-2	Interleukin-2	Homo sapiens	154
KGF	Keratinocyte Growth FactorA	Homo sapiens	195
MB	Mioglobina	Homo sapiens	155
BTC	Betacellulin	Homo sapiens	179
GM-CSF	Granulocyte Macrophage-Colony Stimulating Factor	Homo sapiens	145
LIF	Leukemia Inhibitory Factor	Homo sapiens	203
LEP	Leptin	Homo sapiens	168
BAFF	B-cell Activating Factor	Homo sapiens	286

of the interactions. In this case, even though the preferred codons for each amino acid are more frequently used, none of them is exclusively used.

The differences are somewhat blurred when looking at the nucleotide and nucleotide pairs frequencies, see Fig. S3 in S.I. There, at $\beta = 1$, both the nucleotide and nucleotide-pair fraction are similar to those obtained from the human NCBI database. However, the $\beta = \infty$ sequences (for the NN and even more for the Ind model), entail an increase in the content of C and G nucleotides (a commonly used parameter in codon sequence design) over A and T. This suggests that, at least for the human database, a design criterion based on the most frequent codons is not independent of one that maximises the amount of C, G nucleotides.

2.2.2. Comparison with other design tools

Next, we focused on the NN model due to its superior results in mimicking human codon distribution, and compared its predictions to those of some commercial sequence-design services. This was done by calculating a series of common indicators (CAI, CPB, RCB, RCPB), defined in Methods 5.4, for a set of 28 sequences commonly produced in the laboratory. Eight of these proteins correspond to non-human proteins, while the rest are human, so that a human wild type codon-sequence is also available for them. The set of sequences, along with their corresponding organism and their length (in amino acids), is reported in Table 1.

In order to rationalise the differences between the prediction tools, we referred to the human NCBI database to estimate the typical variability of such indicators when calculated on human sequences. Since these indicators depend on the length of the sequences, for each of the 28 sequences we filtered the human coding sequences listed in the NCBI database to obtain a set of sequences of the same length, and calculated the mean and variance of each indicator on that set. Following this approach, the bias related to the length of the sequence is removed, and the reference values that are obtained can be used to compare the different predictions.

Fig. 3 shows the CAI and CPB for the designed sequences, by different tools, of the 28 sample proteins, together with the values for the wild type sequences, and the average values for the human sequences of the same length. For CAI, we see that wild-type values are generally within one standard deviation from the human average, with the obvious exception of the first 8 sequences, whose wild type is not human. We also see that some tools (i.e. Vendor 1, 3) are clearly designed to generate sequences with high CAIs, and their predictions reach extremely high values. Other tools' designs are close to (Vendor 4) or systematically below (Vendor 2) the average obtained for natural human sequences (dashed blue line), suggesting that the design criteria for these are essentially different. It is notable that the NN model is able to interpolate between these disparate behaviours. At $\beta = 1$, it reproduces the human average with remarkable precision, outperforming all other models. Conversely, at $\beta = \infty$ it generates sequences with a CAI that is marginally lower than those obtained by Vendor 1 and 3. On the other hand, coming to CPB, we notice that the NN model at $\beta = \infty$ is the only one that produces values significantly above the human average, while again reproducing the average at $\beta = 1$, as can be seen when comparing the dashed black and blue lines. It is noteworthy that the value obtained for the human sequences, of the same length as the 28 sample sequences, remains approximately 0.075 and is not compatible with zero. Keeping in mind the definition of the CPB as a probability ratio (see Section 5.4) between the natural observed pair frequency and the value obtained in the independent-codons case, the positive values observed for natural sequences are to be expected. However, the fact that these values are significantly greater than zero indicates that in natural sequences codons are at least pair-correlated, and should not be considered as independent. It is noteworthy that all the others methods produce values that are significantly below the human average, with negative values that are comparable with the wild-type values of the non-human sequences. In contrast, the NN model at $\beta = \infty$ generates codons and codon pairs that are highly probable within the context of the natural human database, resulting in high values for both CAI and CPB.

Similar observations can be drawn from Fig. 4, where the trends of the indicators RCB and GC for the same 28 proteins are studied. Similarly to what is observed for CAI, we detect a separation between the predictions by different vendors. Thus, the designs from Vendor 1 and

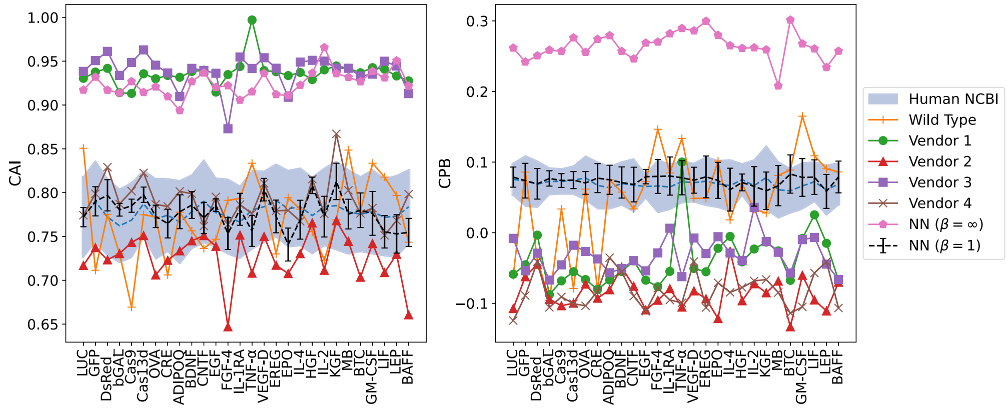

Fig. 3. Codon Adaptation Index (CAI, left panel) and Codon Pair Bias (CPB, right panel) for the codon sequences proposed by different commercial tools for the 28 proteins reported in Table 1, as well as for the wild type protein and for our NN predictions (at $\beta = 1$ and $\beta = \infty$). For comparison, the average values (dashed lines) and standard deviations (grey areas) for codons sequences in the human NCBI database are also presented. These correspond to proteins of the same length (±2 codons) as those considered in Table 1. This is because both indicators depend of sequence length. In the case of $\beta = 1$, a database of 1000 codon sequences is generated for each protein, and the mean and standard deviation (represented by an error bar) is reported. The colour code is the same in the two panels.

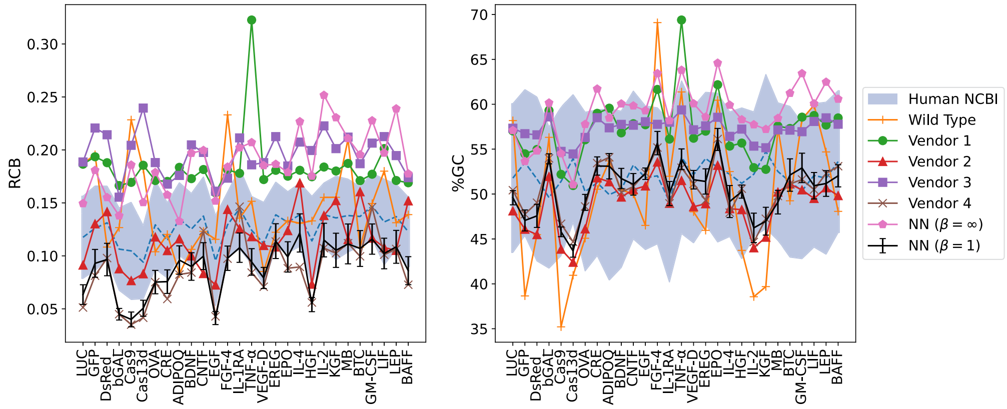

Fig. 4. Relative Codon Bias (RCB, left panel) and Percentage of GC (Guanine and Cytosine) (GC, right panel) for the same sequences as in Fig. 3. The colour codes and abbreviations are the same as in Fig. 3.

3 are found well above those from Vendor 2 and 4, while again the NN model switches between the two groups, depending on the value of β. Remarkably, the NN model at $\beta = 1$ reproduces the predictions of Vendor 4 quite closely, and both are found at the lower end of the 1-standard-deviation region.

This observation is significant as, by construction, RCB (RCPB) equal to zero corresponds to having a frequency of codon (codon pair) usage, within the sequence, that is precisely equal to the frequency observed in the database. Therefore, the predictions of Vendor 4 and NN at $\beta = 1$ more closely resemble the "background" usage in the whole human database, than the human sequences themselves. Furthermore, the average RCB of natural human sequences is not consistent with the null value, which lies outside the 1-standard-deviation region. This seemingly paradoxical phenomenon may have two potential explanations:

- The human natural codon distribution is not homogeneous but instead depends on the gene length. This could happen, for instance, if the 5'- or 3'-terminal parts of the sequence showed a different codon usage than the bulk of the sequence. This end-effect would be dependent on the sequence length: the shorter the sequence, the more relevant it would become.
- The human codon distribution in the NCBI database is not unimodal, which implies that when a database probability ψ_α/ϕ_α of each codon is extracted (see Eq. (15)), we are averaging the codon usage of different distributions. This could be the case, for example, if codon bias was function- or organ-specific (however, other causes for heterogeneity could be also possible).

To test the first hypothesis, we proceeded as follows: We divided the human NCBI database in subsets according to sequence length, from 95 to 7505 nucleotides, in intervals of 100 nucleotides. For the genes in each subset, we generated an alternative codon sequence with the NN model at $\beta = 1$. This process yielded 75 subsets of the NCBI database, and the corresponding 75 databases of NN sequences. Then, we calculated the average RCB and standard deviation within each database. The results are plotted in Fig. S4 of S.I, and show that the RCB values from both the NCBI and NN sequences drop rapidly upon increasing the sequence length, due to the corresponding reduction in the statistical error, implicit in the definition of Eq. (15) (actually, for very short sequences, there were not enough residues to yield a significant sample of codon usage). However, the RCB indicator for NN sequences continues to decline until reaching values below 0.025, while the NCBI stabilises at around 0.1. Additionally, we see that in the case of the NN sequences, the standard deviation decreases with sequence length, indicating that the heterogeneity of the RCB within the same database is diminishing, whereas the deviations remain constant for the NCBI subsets.

The above suggests that the RCB behaviour in natural sequences is not caused by a length-dependent codon bias. Instead it seems to point to heterogeneities in codon usage related to other factors, although further exploration of this topic is necessary to identify which factors are involved. In order to provide another test for this hypothesis, we artificially joined the 28 codon sequences for each dataset (namely, the NCBI database, the set of wild-type sequences and the sets corresponding to the six different methods) and calculated the RCB thereof. It can be seen, in Fig. S5, that the RCB value for the NCBI set is much smaller than

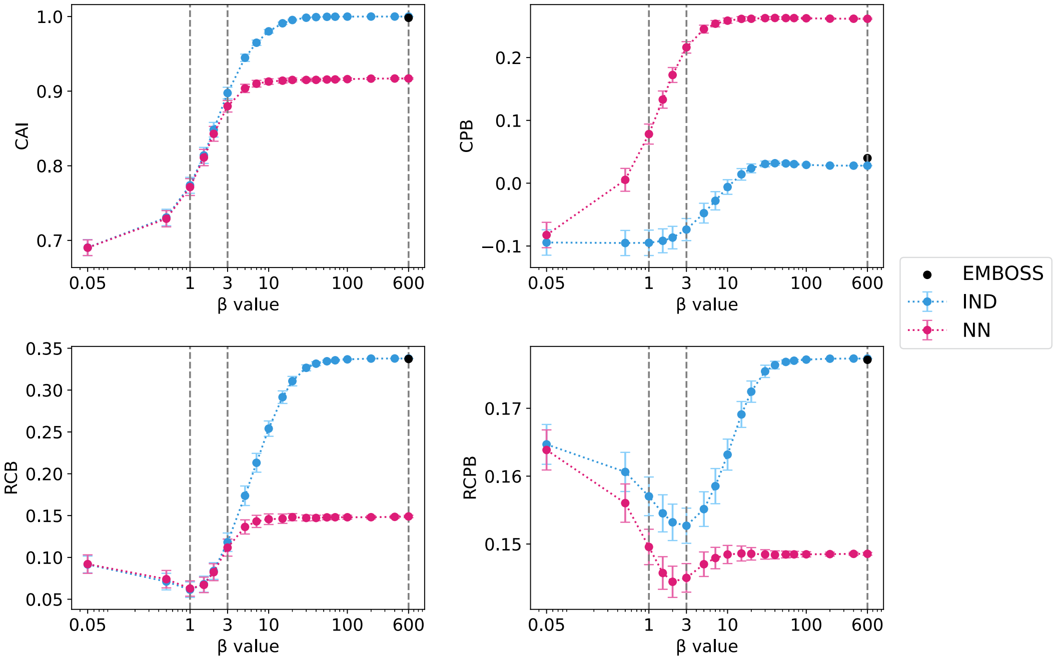

Fig. 5. Temperature dependence of the different indicators for protein luciferase, calculated with the two models. The averages of Codon Adaptation Index (**CAI**; Eq. (13)), Codon Pair Bias (**CPB**; Eq. (14)), Relative Codon Bias (**RCB**; Eq. (15)) and Relative Codon Pair Bias (**RCPB**; Eq. (16)) were calculated across databases consisting of 1000 sequences each, obtained as explained in Sec. 5.5. Error bars denote standard deviations. In each panel a black dot corresponds to the value for the sequence proposed by EMBOSS web server [18,29], and the vertical dashed lines indicate the β values finally chosen, $\beta = 600$ representing $\beta \to \infty$; see text.

when using separate sequences. This suggests that the use of the joint sequence for calculations smeared out the differences in codon usage, intrinsic to each separate sequence. On the other hand, RCPB plots (Fig. S6 in S.I.) do not reveal great differences among all methods and the human sequences, and no special conclusions can be drawn from them. Finally, we found that the GC content (Fig. 4, right panel) reproduces the patterns found for CAI. Once again, Vendor 1, 3 and the designs of NN at $\beta = \infty$ are higher and separated from the rest, with NN at $\beta = 1$ closely resembling the designs of Vendor 4. The average human fraction for GC content is slightly higher than 50%, while the more extreme ones are found slightly above 55%, still within the one-standard-deviation region around the human average. As previously noted, the maximisation of CAI entails an increase in GC content.

2.3. Application of the NN and Ind model to the optimisation of luciferase mRNA

To verify whether using the NN model would increase the expression of the desired protein, we conducted protein expression assays in different human cell lines. The Firefly Luciferase protein sequence (Uniprot: P08659) was selected as a reporter and the codon optimisation over the sequence was performed by using the Ind or the NN model.

2.3.1. Selection of codon sequences for experimental tests

The models were run at different inverse temperatures β, resulting in a set of different designs for each β. In order to select a few sequences for experimental testing, the characteristics of each set were examined in terms of the values of different indicators.

Fig. 5 reveals the dependence on β ofthe CAI, CPB, RCB and RCPB indicators. It can be observed that the parameters of the NN model stabilise at lower values of β than those of the Ind model, due to the pair interactions.

In both models, the CAI (Eq. (13)) increases with β, something that is related to the fact that, at high β, just the sequences with the lowest energies are selected.

For the Ind model, this implies that codons with high probabilities are chosen, since the energies are directly related to their frequency in the learning dataset. Consequently, the CAI of the Ind model reaches 1 at $\beta = \infty$; in contrast, the CAI of the NN model never reaches 1, due to the contribution of codon pairs. Contrary to CAI, RCB (Eq. (15)) exhibits a minimum at $\beta = 1$ for both models. This is in agreement with the fact that their parameters have been tuned to reproduce the natural codon distribution precisely at that temperature (it should be reminded that the RCB is always non-negative, with the null value corresponding to the case where the codon usage reproduces perfectly the natural one). Notice that for both CAI and RCB, the values obtained with the two models are essentially the same up to $\beta = 3$. This is due to the fact that the indicators based on single codon usage cannot reflect the relevance of pair interactions, until β values sensibly higher than the learning value are reached.

Conversely, CPB and RCPB (Eqs. (14), (16)) start at roughly the same value at the highest temperature considered (lowest β), indicating that the sequences are essentially random at that β, in the sense that for each given amino acid, any of its codons is chosen with uniform probability. However, their values quickly diverge. Indeed, considering CPB, we observe that the sequences generated by the NN model present positive values already at $\beta = 1$ (i.e., the pairs they contain are more likely found than expected by random selection of single codons with their natural frequency), while the Ind model shows negative values up to $\beta = 10$. Similarly, RCPB plots for the Ind and NN models exhibit different characteristics. The NN values are consistently smaller (i.e. “more natural”) across all temperatures, and with minima close to the learning inverse temperature $\beta = 1$.

The GC percentage (see Fig. S7 in S.I.) exhibits a similar behaviour to that of CAI as the codon frequency correlates with its GC content. However, in the case of the Ind at high β, we see that not all the amino acids have the GC-richest codon as their most likely one. Indeed, GC fraction has a maximum at around $\beta = 30$, where suboptimal codons are still sampled, and stabilises at a slightly lower value when only the most probable codon of each amino acid is selected.

In addition to the codon usage indicators mentioned above, we also considered the RNA (secondary structure) energy, as predicted by the RNAFold server [30], to see the amount of secondary structure present in the different designed sequences. Fig. S7 in S.I. reveals that the RNA energy for the NN sequences decreased (almost monotonically) upon increasing β, and was consistently lower than, or equal to, the energy of the Ind sequences. These exhibited a minimum and subsequent increase at low temperatures. This is due to the fact that the sequence composed by the most probable codons has a minor GC content than some populated sequence at intermediate temperatures (see Fig. S8), and GC coupling is stronger than AU. On the other hand, Fig. S7 illustrates that the NN sequences consistently exhibit more secondary structure than Ind ones. This is evidenced by the number of bonded nucleotides, which indicates that the energy per bonded nucleotide (i.e., the average "strength" of the contact) is not significantly different between the two models. Ultimately, since it was not very clear whether the RNAfold energy or the energy per bonded nucleotide would constitute a good selection criterion, we decided not to use any structural criterion to guide our choice of sequences.

According to the above analysis, the databases $\beta = 1$ for the Ind model, $\beta = 3$ and $\beta = \infty$ for both models were selected. Using $\beta = 1$ for the Ind model leads to a human distribution, while $\beta = 3$ for both models allows a slight shift from the human distribution, increasing the GC percentage and decreasing the energy, so that the mRNA will be more stable, but without reaching the minimum. Lastly, employing $\beta = \infty$ yields the sequences (one per model) minimising the energy Eq. (4). These sequences, in both models, are characterised by having the highest CAI values and being furthest from the human distribution in terms of RCB. Additionally, in the case of Ind ($\beta = \infty$), the most frequent nucleotide is used to encode each amino acid, establishing it as our benchmark due to its simplicity and widespread adoption among researchers. To enhance realism, we employed the commonly accepted EMBOSS web server [29] to optimise the Luciferase sequence. This was done under the assumption that its results would align with those of our algorithm, since it uses the most likely codon for each amino acid [18]. However, disparities were observed, particularly in the codon for Arginine; while our method utilised AGA as the most frequent codon, EMBOSS utilised AGG. This discrepancy is related to the use of different versions of the Human Genome sequence. The version utilised by EMBOSS is considerably older than the one employed in our method, resulting in slight variations in the frequencies of AGA and AGG codons. However, upon plotting the indicators' values for the EMBOSS sequence in Fig. 5, we observed that they closely resemble those of Ind ($\beta = \infty$). Thus, to align more closely with standard bench science practices, we opted to adopt the EMBOSS sequence as our benchmark, under the assumption that it would closely approximate the conditions of Ind ($\beta = \infty$).

Finally, the selected sets of sequences were filtered, as detailed in Section 5.5, using the parameters described in Table S1, in order to identify the most promising candidates, for each β value, for experimental tests.

2.3.2. *In vitro performance of the Ind and NN models*

The selected optimised sequences of luciferase were synthesised and inserted into a plasmid for *in vitro* transcription into mRNA. The resulting mRNAs were then transfected into HeLa and HepG-2 cell lines, and the produced luminescence was quantified. Due to the negatively charged nature of mRNA, which repels the anionic cell membrane, encapsulation into lipid nanoparticles (LNPs) is essential for optimal *in vivo* functionality. Furthermore, LNP formulation also contributes to maintain mRNA integrity, promote endocytosis, and facilitate endosomal escape [31]. Accordingly, two experiments were conducted to validate the generated mRNAs: one using a commercial cationic transfection reagent (Fig. 6A), and another involving the encapsulation of mRNAs into LNPs (Fig. 6B). The production of LNPs does not impact protein production, as Table S2 and S3 reveal: the observed variability in the quality parameters for LNPs does not significantly affect the experimental results.

The results obtained from both cell lines and utilising both transfection methods revealed a slightly superior performance of the Ind_3 and EMBOSS optimisation method (Fig. 6). However, statistically significant differences were observed only when comparing Ind_3 with NN_∞ using a transfection reagent in Hela cell line (Fig. 6A). We also conducted assays with a lower final mRNA quantity for transfection to discern potential differences (Fig. S9) and obtained similar results, with a superior performance of the EMBOSS method. This suggests that the Ind_3 and EMBOSS optimisation methods appear to be more efficient in generating higher levels of Luciferase *in vitro*.

2.3.3. *In vivo performance of the Ind and NN models*

Since our objective is to optimise mRNA coding sequences for therapeutic applications *in vivo*, we proceeded to analyse the production of luciferase in a mouse model. Assuming that codon usage statistics are comparable between humans and mice (Fig. S10), we injected the optimised mRNAs intramuscularly and monitored luminescence production at 4- and 24-hours post-injection (Fig. 7). Notably, codon optimisation using the NN model significantly enhanced protein production compared to the Ind model. Specifically, employing the NN model with $\beta = \infty$ resulted in the highest protein production levels at both 4- and 24-hours post-inoculation. This result suggests that a codon optimisation based on "interactions" between neighbouring codons can improve protein production.

3. Discussion

While codon usage has a significant impact on the levels of heterologous protein production, there remains a lack of consensus on the most effective approach to maximising such production. Accordingly, different vendors use their own (often undisclosed) method to optimise gene design, and there is ongoing research aimed at improving the algorithms and understanding the key factors that affect protein production. Here we have developed a new codon optimisation method, inspired by statistical physics, in which we assume that natural codon sequences are not the result of some kind of optimisation process, carried out by evolution. Instead, adopting a neutral evolution perspective, we consider them as random samples of a general sequence probability, for which we postulate the form of a Boltzmann probability depending on an energy function and on a value of the temperature. This approach is usually adopted to study a physical system with clearly identified interactions that, at least at the probabilistic level, completely determine its behaviour (e.g., protein folding). This is not the case here: the choice of the codons (the states of our system) does not depend on an interaction between amino acids (even if we can think that the necessity for interaction with the ribosome may effectively "couple" the codons of groups of amino-acids). Therefore, our approach is not based on ab-initio observations on the nature of the system. However, it proves to be quite fruitful: we have found that, after fitting the energy parameters at the inverse temperature $\beta = 1$, the sequences obtained at any given β share common characteristics in terms of CAI, RCB, GC content, etc, as indicated by the small values of the variance of these quantities, see for example Fig. 5.

Thus, it appears that assuming the existence of an energy function and using it to generate a family of different probability distributions at different β, captures some fundamental features of the complex processes that determine the codon usage patterns, that are species dependent. In addition, we have seen that the results on CPB of Fig. 3 confirm the role of pair interactions, as a key factor in explaining the observed positive average for natural sequences. Note also that the parameters of the energy function are fitted, at $\beta = 1$, to maximise the likelihood of the natural sequences, so that, differently to all other methods that we know, the observed codon and codon pair frequencies here are an output, and not an input of the model. Finally, we have found that in the low temperature region, the NN model roughly reproduces the CAI and RCB indicators, as well as the GC fraction obtained by commercial design tools. This suggests that, by decreasing the temperature parameter, our model can gradually switch from producing natural sequences to

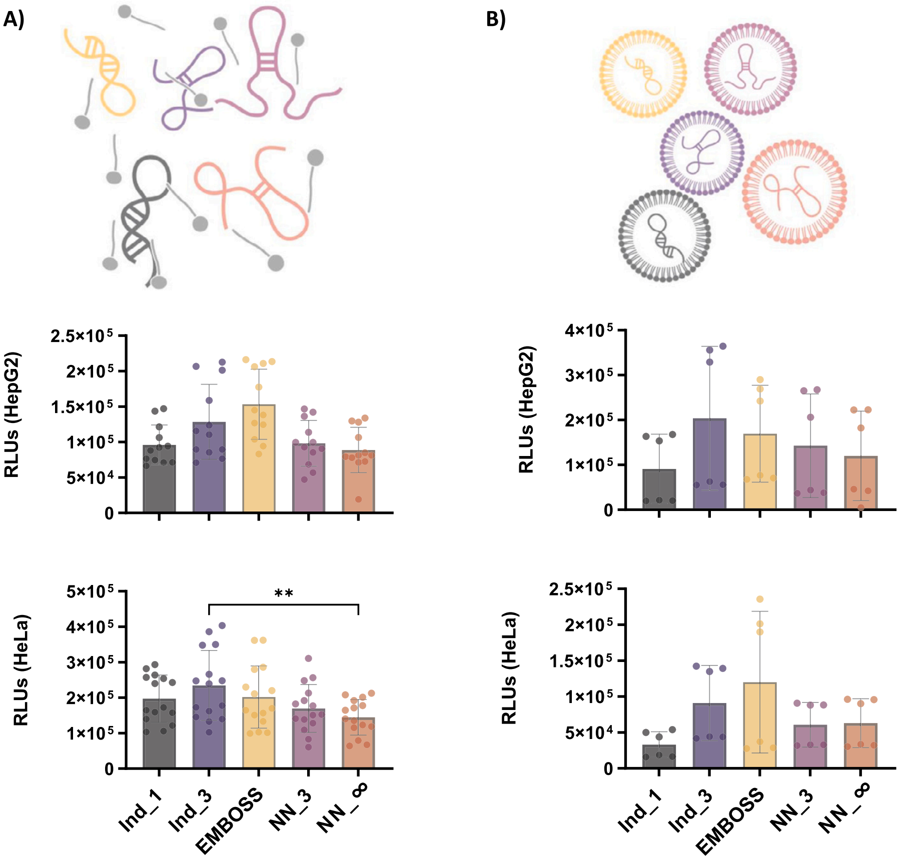

Fig. 6. ***In vitro luciferase production from mRNA sequences optimised by*** **the Ind and the NN model.** Luciferase production in HepG2 (upper panel) or Hela (lower panel) cell line, was measured in Relative Luminescence Units (RLU). The transfection reagent was a commercial cationic lipid in **A)** and a Lipid Nanoparticle in **B)**. In all cases a final quantity of 100 ng of mRNA per well was used. Each bar represents the mean of at least 2 independent experiments with triplicates in each experiment, where each point represents the result for an independent well. Ind_1, Ind_3 represent individual codon optimisations using $\beta = 1$ and $\beta = 3$ respectively, and EMBOSS corresponds to using always the most frequent codon. NN_3 represents the nearest-neighbour interaction model at $\beta = 3$ and NN_∞ is the sequence obtained with the NN model at $\beta = 600$.

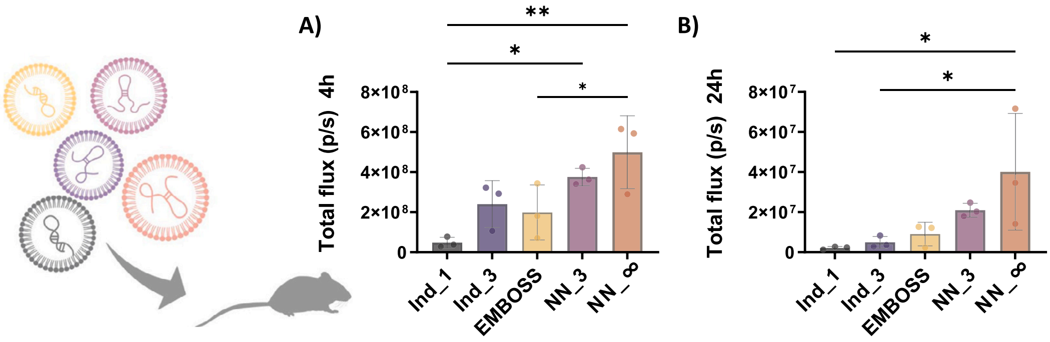

Fig. 7. ***In vivo luciferase production from mRNAs sequences optimised*** **by the Ind and the NN model.** Lipid Nanoparticles, encapsulating the optimised mRNAs, were used to inoculate 1 µg of total mRNA intramuscularly to each mouse. The luminescence yield was measured by total flux quantification in photons per second (p/s) at **A)** 4 hours and **B)** 24 hours post inoculation. Each bar represents the mean of two independent experiments, one using two mice per group and the other 3 mice per group. Ind_1 represents individual codon optimisation using $\beta = 1$ and Ind_3 using $\beta = 3$ and EMBOSS using always the most frequent codon. NN_3 represents the nearest-neighbour interaction model using $\beta = 3$ and NN_∞ using β 600 which is equivalent as ∞.

generating "extreme" ones, with a bias towards high values of CAI and GC content.

By design, our approach generates several sequences, except at T rigorously zero, where the solution is unique (unless two different codon sequences have exactly the same energy, which is very unlikely). However, the resulting sequences need not all be good hits, and some filtering is recommended, in order to satisfy other important criteria that are not implicit in our model. For example, requirements on secondary structure formation and its energy, or on forbidden nucleotide sequences, as well as on the absence of long GC-rich regions, must be implemented on top of the pool of predicted sequences. However, the computational time is extremely low (Fig. S11), which makes our method widely applicable.

Other methods can also return different solutions, apart from the optimal one, using Monte Carlo, Genetic Algorithms, or other approaches, and they also resort to post-generation filtering to implement other criteria. For example, Refs. [20,22] use Monte Carlo to explore alternatives to the most frequent codon, while Ref. [21] presents a scaling factor to tune the codon frequencies; a bee-colony algorithm is used to perform a multi-objective optimisation in Ref. [32]. However, to the best of our knowledge, none of them uses an energy-based approach, with the temperature as a control parameter, which provides a simple and intuitive framework for sequence design, and also allows to generate sequences with homogeneous values of different indicators.

In terms of the efficiency in protein production in both *in vitro* and *in vivo* settings, a notable disparity has been observed, with the Ind model outperforming the NN model *in vitro, but the NN model showing significantly better performance in vivo. A* plausible explanation for this difference could be that mRNA translation encounters biological barriers upon entering the cell, which vary depending on the cell type and context within the body [33]. We have attempted to reduce variability by using the same 5' and 3'UTR [34], and the same polyA tail length for all mRNA because those features are the ones that have more impact on mRNA stability and translation efficiency [35,36]. However, many factors within the coding sequence could contribute to disparities between *in vitro* and *in vivo* performance. One such factor is the structure of the mRNA, which influences its stability in response to metabolite concentrations, such as Mg^{2+} [37], which can vary between *in vitro* conditions or among tissues. In an effort to clarify the role of mRNA stability in our designs, we performed a stability experiment. Unexpectedly, the results showed that the design with the lowest stability (Ind_3) in solution (Fig. S12B) outperformed others *in vitro* (Fig. S12C), despite being evidently more degraded than its counterparts. Hence, it appears that additional factors are playing a pivotal role *in vivo*. We hypothesise that these factors are associated with the differential presence of specific nucleotide motifs among the optimised sequences. These motifs could potentially influence ribosome entry and translocation [38], or the binding of microRNAs, which may exhibit different expression patterns in different cell types. Furthermore, the emerging field of ribosome heterogenicity among tissues could provide insights into variations in mRNA translation across different tissues [39]. Given that our NN model extracts probabilities and patterns from human sequences, we could be implicitly accounting for undiscovered patterns that contribute to superior *in vivo* performance.

4. Conclusions

We have proposed and tested, *in silico*, *in vitro* and *in vivo*, a new algorithm for gene design, inspired by statistical physics. The algorithm differs significantly from the existing methods, and can be tuned to generate sequences with any degree of bias, as measured by common indicators like CAI, CPB, RCB, and GC content ranging from natural sequences to extremely biased ones. In fact, at $\beta = 1$ it outperforms other methods in generating sequences that are closely similar to natural human sequences, in terms of the CAI, CPB, and GC indicators. This encourages the use of the NN model as a null model to answer biological questions, as we did when investigating the cause of the distribution of RCB values of natural sequences.

Our method does not set the last word in elucidating the relationship between codon usage and protein expression, that remains an elusive subject of research. Our findings indicate that the results can be different *in vitro and in vivo, both for the designs that mimick natural sequences,* and for those with a high level of bias towards the most common codons (i.e., the low-T sequences, presenting high CAI and GC content). This points at the relevance, for protein expression, of factors that are not simply related to the choice of codons. On the other hand, we note that our approach, designed to mimic the patterns of human codon sequences, was able to bypass the multifactorial in vivo regulatory elements that affect mRNA translation. Although we lack an explanation for that, we note that our NN method generates an adequate translation efficiency *in vivo*, that could be applied to the future design of mRNA therapeutics for vaccines or gene replacement. The algorithm at the moment is implemented as a series of Python and Fortran codes, that are probably not very handy for the casual user. Nevertheless, we are developing a user-friendly tool, potentially in the form of a web server, with the objective of enhancing its usability and obtaining independent feedback.

Finally, we notice that in this work we have tuned our model to reproduce the probability of sequences from the entire NCBI human database, without selecting highly expressed genes or distinguishing among tissues or function. While this is the correct approach to start with, in order to characterise the general human database, we note that the learning database could be restricted and the model tailored to address specific tasks and questions.

5. Methods

5.1. Database preparation

In order to build the protein sequence database, we started from the Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13) from Ensembl 107 [40] identifiers database. With the identifiers, we obtained the codon sequences from NCBI's Nucleotide database [41,42], resulting in a database of 116,487 sequences.

For computational simplicity and to increase the statistics of the codons for each peptide sequence, we took advantage of the fact that the parameters in our model are assumed to be site-independent to split the protein sequences, obtained from the NCBI gene sequences, into shorter peptides of length L; thus, the same peptide is represented by several codon sequences (even though they come from different proteins). Splitting started alternatively at the N- or C-terminal end of the peptide, to avoid biases. For each sequence, we discarded the last (N- or C-terminal) peptide, if it was shorter than L residues. This procedure possibly introduces some border effect (that will be less important the longer the peptides are); we will discuss in the following section how to (partially) cope with this problem when calculating the probabilities.

Hence, by randomly selecting sequences from the entire database, and splitting them as explained above, we built a database $\mathcal{S}$ of 44827 codon sequences[1] representing a set of N different peptides $\{p^k, k = 1,\dots,N\}$ of length $L = 50$:

$$\mathcal{S} = \{\{S_j(p^k), j = 1,\dots,M(p^k)\}, k = 1,\dots,N\}, \tag{1}$$

where $M(p^k)$ is the number of codon sequences in the database that translates to the same peptide p^k (notice that sequences $S_j(p^k)$ need not be different: $\mathcal{S}$ may contain repeated codon sequences for p^k, and $M(p^k)$ will count all of them).

5.2. Data-driven statistical physics models

In the following, we discuss two probabilistic models designed to describe the experimental database prepared in the previous section. More details on the methods can be found in the S.I.

5.2.1. Nearest-neighbour interaction model

Let $P = \{a_i, i = 1\dots L(P)\}$ indicate a protein sequence, where $a = \{1\dots 20\}$ are the amino acids, so that a_i is the amino acid at the position i. Let $\mathcal{K}(a)$ be the set of codons $\alpha^{(a)}$ codifying for the amino acid a; $\mathcal{K} = \cup_{a=1}^{20}\mathcal{K}(a)$ is the set of all codons, of cardinality $K = 61$. Let $S(P) = \{\alpha_i \equiv \alpha^{(a_i)}, i = 1,\dots,L(P)\}$ be a codon sequence encoding the protein sequence P. Inspired by statistical physics, we associate, to each codon sequence $S(P)$ of a given protein P, a Boltzmann probability

$$p(S(P)|\mathbf{h},\mathbf{J},P) = \frac{1}{\mathcal{Z}(P)} e^{-\beta \mathcal{H}(S(P)|\mathbf{h},\mathbf{J})}, \tag{2}$$

i.e., the ratio of a Boltzmann weight $\exp(-\beta H(S|\mathbf{h},\mathbf{J}))$ and the corresponding partition function (for a fixed protein sequence), as a normalisation factor:

[1] This number was just a trade-off accounting for computational limitations.

$$\mathcal{Z}(P) = \sum_{\{\alpha_i \in \mathcal{K}(a_i), i=1,\dots,L(P)\}} e^{-\beta \mathcal{H}(S(P)|\mathbf{h},\mathbf{J})}. \tag{3}$$

Thus, in this approach, we map each codon sequence to a state of a physical system, ruled by the energy function $H(S|\mathbf{h},\mathbf{J})$:

$$\mathcal{H}(S(P)|\mathbf{h},\mathbf{J}) = \sum_{i=1}^{L(P)} h_{\alpha_i} + \sum_{i=1}^{L(P)-1} J_{\alpha_{i+1},\alpha_i}, \tag{4}$$

depending on the parameters $\mathbf{h} = \{h_\alpha\}$, $\mathbf{J} = \{J_{\alpha\beta}\}$, that are a priori unknown and must be optimised to recover the observed codon sequences in a given, species-dependent, database.

Notice that we assume that h_{α_i} depends only on the codon type α_i and not on the sequence position i in where the codon is located. The same holds true for $J_{\alpha_{i+1},\alpha_i}$, so that the number of parameters is independent from the length $L(P)$ of protein P, and does not exceed $K^2 + K$. The reason for these assumptions is that we do not expect the parameters to depend strongly on the position along the sequence, so, as a first approximation, we assume them to be completely position-independent. Actually, it is known that codon usage is related to the species-dependent availability of the different tRNAs, which is site independent, but also to site-dependent constraints related to the mRNA secondary structure, as well as to the need to control specific signals to modulate translation; in taking the parameters as position-independent, we are assuming that the latter effects are less relevant than the former ones.

When calculating the probability of a peptide in the database $\mathcal{S}$ introduced in Sec. 5.1, we slightly modify the expression Eq. (2), to cope with the fact that the peptides are fragments of a longer protein sequence, by considering the peptide p' of length $L+2$, obtained from peptide p adding a codon α_0 at the N-term, and α_{L+1} at the C-term, and computing the marginal probability, over all possible codons α_0 and α_{L+1} of any amino acid:

$$\begin{aligned}\mathcal{P}(S(p)|\mathbf{h},\mathbf{J}) &= \sum_{\alpha_0,\alpha_{L+1}=1}^{K} p(S(p')|\mathbf{h},\mathbf{J}) \\ &= \frac{1}{\mathcal{Z}(p')} \sum_{\alpha_0,\alpha_{L+1}=1}^{K} e^{-\beta(\sum_{i=0}^{L+1} h_{\alpha_i} + \sum_{i=0}^{L} J_{\alpha_{i+1},\alpha_i})},\end{aligned} \tag{5}$$

where now $\mathcal{Z}(p')$ is given by:

$$\mathcal{Z}(p') = \sum_{\alpha_0,\alpha_{L+1}=1}^{K} \sum_{\{\alpha_i \in \mathcal{K}(a_i), i=1,\dots,L\}} e^{-\beta \mathcal{H}(S(p')|\mathbf{h},\mathbf{J})}. \tag{6}$$

Notice that in the latter expressions, the sum over α_0, α_{L+1} at the ends of the peptides is over all possible codons, and not limited to a particular amino acid. In this way we avoid disregarding interactions (that indeed exist) at the ends, at the price of assuming that all amino acids can appear close to a given one (and each with a probability that will be proportional to the number of corresponding codons).

This hypothesis does not consider that a minority of peptides (those corresponding to the N- and C-term of the protein) do lack one neighbour. This could cause some problems, especially for Methionine, that is preferentially found at the N-terminal of proteins; however, we ignore this, for the sake of simplicity.

Probabilities Eqs. (2), (5) can be efficiently calculated using a "transfer matrix" formalism; for instance, Eq. (3) can be recast as:

$$\mathcal{Z}(P) = \sum_{\{\alpha_i \in \mathcal{K}(a_i)\}} \left(e^{-\beta h_{\alpha_L}/2} \left(\prod_{i=1}^{L-1} T_{\alpha_{i+1},\alpha_i} \right) e^{-\beta h_{\alpha_1}/2} \right), \tag{7}$$

where: $T_{\alpha,\alpha'} = e^{-\beta(h_\alpha/2 + J_{\alpha\alpha'} + h_{\alpha'}/2)}$ can be seen as the elements of the transfer matrix; care must be paid to considering just the codons α, α' available for the amino acids at positions $i+1, i$.

5.2.2. Individual codons model

A simpler model can be extracted from the one exposed in Section 5.2.1, by simply setting all interaction parameters $J_{\alpha,\beta} = 0$, for all codons α, β at neighbouring sites. In this case, the codons are independent of each other and the probability Eq. (2) becomes:

$$p(S(p)|\mathbf{h}) = \frac{1}{\mathcal{Z}(p)} e^{-\beta \sum_{i=1}^{L} h_{\alpha_i}} = \prod_{\gamma=1}^{K} p_\gamma(h_\gamma)^{n_\gamma(S(p))} \tag{8}$$

where

$$p_\gamma(h_\gamma) = \frac{e^{-\beta h_\gamma}}{\sum_{\gamma \in \mathcal{K}(a)} e^{-\beta h_\gamma}}, \tag{9}$$

and $n_\gamma(S(p))$ is the number of times the codon γ appears in sequence $S(P)$. Thus, the probability for all the database will read:

$$p(\mathcal{S}|\mathbf{h}) = \prod_{\gamma=1}^{K} p_\gamma(h_\gamma)^{n_\gamma(\mathcal{S})}, \tag{10}$$

with $n_\gamma(\mathcal{S})$ the number of times γ appears in the whole database. The h_γ are found maximising the above probability.

5.2.3. Data-driven determination of the parameters

Equation (2) (or (5)) allows to calculate the probability of any protein (or capped peptide) codon sequence given the values of the parameters $\mathbf{h}, \mathbf{J}$ and β.

We regard the variability of natural codons sequences, expressing the same protein sequence, as a manifestation of the accessibility of different states of the physical system, with different probability, when the inverse temperature is set to $\beta = 1$ (for simplicity and without loss of generality, since this is equivalent to setting the energy scale).

Assuming independence among all sequences in the database, from Eq. (5), we obtain for the probability of the whole peptide database $\mathcal{S}$:

$$p(\mathcal{S}|\mathbf{h},\mathbf{J}) = \prod_{k=1}^{N} \left(\prod_{j=1}^{M(p^k)} \mathcal{P}(S_j(p^k)|\mathbf{h},\mathbf{J},\beta=1) \right), \tag{11}$$

where $j = \{1 \dots M(p^k)\}$ labels all codon sequences in our dataset that code the same peptide p^k, with $k = \{1 \dots N\}$ labeling the different peptides. Notice that we have one partition function for each amino acid sequence, since these sequences differ from each other.

Bayes' theorem states that the probability of the parameters given the database can be related to that of the database given the parameters and the prior probabilities of the parameters:

$$p(\mathbf{h},\mathbf{J}|\mathcal{S}) \propto p(\mathcal{S}|\mathbf{h},\mathbf{J}) p^{\text{prior}}(\mathbf{h},\mathbf{J}). \tag{12}$$

Assuming uniform priors, we can reduce the determination of the parameters to maximising the likelihood $p(\mathcal{S}|\mathbf{h},\mathbf{J},\beta=1)$ on the values of $\mathbf{h},\mathbf{J}$. Actually, for convenience, we will maximise $\ln p(\mathcal{S}|\mathbf{h},\mathbf{J})$ instead, using the L-BFGS-B algorithm, implemented in the function scipy.optimize.minimize of the Python Scipy library; as a convergence protocol, we consider convergence achieved when there is a variation of 0.001% in the value of the objective function or a 0.1% change in the gradient between one step and the next).

For the case of the Individual Codons model, an analogous approach is applied, using the likelihood Eq. (10), with the probabilities calculated from Eq. (9) at $\beta = 1$.

5.3. Sequence generation

By setting the optimal values for $\mathbf{h}$ and $\mathbf{J}$, we can generate species-specific codon sequences for a given amino acid sequence. On the one hand, we want to find the codon sequences that encode a given amino acid sequence at low temperatures and, ultimately, at zero temperature due to its significance as the sequence that minimises the energy, i.e., the "best" sequence in terms of the $\mathbf{h}$ and $\mathbf{J}$ parameters. On the other hand, we want to explore the sequence space at $T = 1$, that should present the same natural variability observed in nature.

We will tackle both tasks using the Simulated Annealing (SA) algorithm [43] implemented on top of a standard Metropolis Monte Carlo scheme, where, at any temperature, codon changes are always accepted if they lower the energy Eq. (4), and are accepted with probability $e^{-\Delta H/T}$ depending on the energy variation upon the change.

We start simulations at a high temperature, with a random codon sequence for the selected protein, and gradually decrease the temperature, to avoid getting trapped in local minima. At low temperatures, we finally find the lowest energy solution. Further details on the sequence generation protocol can be found in the Supplementary Information.

5.4. Indicators of codon bias

We use some common indicators of the codon usage bias, that is, the “fitness” of a codon sequence to be used by a certain species. Let $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ be a protein sequence, represented by the codon sequence $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$. Upon calling ψ_α, $\psi_{\alpha,\beta}$ the observed number of occurrences for codon α or neighbouring codon pair (α, β) respectively, and φ_a, $\varphi_{a,b}$ the observed number of instances of amino acid a or neighbouring amino acid pair (a, b), in the reference dataset being used (in our case, the human dataset from NCBI), we will consider:

- The Codon Adaption Index (CAI) [44], representing the bias of a codon sequence towards the most used codon for each amino acid:

$$\mathrm{CAI}(\gamma, \sigma) = \left(\prod_{i=1}^{n} \frac{\psi_{\sigma_i}}{\psi^{max}(\gamma_i)} \right)^{1/n}, \tag{13}$$

where $\psi^{max}(a)$ is the number of occurrences of the most likely codon for amino acid a. Notice that the maximum value for CAI is 1, attained when the most common codon is used for each amino acid of the sequence.

- The Codon Pair Bias (CPB) [45], defined as:

$$\mathrm{CPB}(\gamma, \sigma) = \frac{1}{(n-1)} \sum_{i=1}^{n-1} \ln \left(\frac{\psi_{\sigma_i,\sigma_{i+1}} \varphi_{\gamma_i} \varphi_{\gamma_{i+1}}}{\varphi_{\gamma_i,\gamma_{i+1}} \psi_{\sigma_i} \psi_{\sigma_{i+1}}} \right). \tag{14}$$

Notice that CPB, although similar in spirit to CAI, has a different meaning: the ratios ψ/φ, with one or two indices, can be interpreted as a probability of using a codon or codon pair respectively (indeed, the denominator $\varphi_{a,b}$ is equal to the sum over all possible codons pairs compatible with the amino acid pair (a, b) in the database, and analogously for φ_a). Thus, CPB takes into account the ratio of the codon pair joint probability to the codon pair probability in the independent codon case: a deviation from zero implies that the sequence contains predominantly codons that are more (if positive) or less likely (if negative) to occur together than in a random choice. Notice that this does not imply that, in natural sequences, all pairs $(i, i+1)$ will contribute positively: CPB just reflects the deviation of a sequence from the natural independent codon distribution.

- The Relative Codon Bias (RCB) [1]:

$$\mathrm{RCB}(\gamma, \sigma) = \sum_{a=1}^{20} \frac{\eta_a(\gamma)}{n} \sum_{\alpha \in \mathcal{K}(a)} \frac{1}{K(a)} \left| \frac{\vartheta_\alpha(\sigma)}{\eta_a(\gamma)} - \frac{\psi_\alpha}{\varphi_a} \right|, \tag{15}$$

where $K(a)$ is the number of codons corresponding to amino acid a (i.e., the cardinality of $\mathcal{K}(a)$), $\eta_a(\gamma)$ is the number of times amino acid a appears in protein sequence γ and $\vartheta_\alpha(\sigma)$ is the number of times codon α appears in the codon sequence σ. This indicator is more “noisy” and sequence-length dependent than the previous ones: indeed, it is based on the deviation of the fraction $\vartheta_\alpha(\sigma)/\eta_a(\gamma)$ of a given codon α, within sequence σ, with respect to its overall fraction in the database, ψ_α/φ_a.

Table 2
Criteria used to select the optimised Luciferase sequences in the sets of possible solutions generated by the NN and Ind algorithms. The type of parameters and the values used to select the optimised sequences for each database are shown; see text for details.

Inverse temperature	$\beta = 1$	$\beta = 3$	$\beta = \infty$
Allowed restriction-enzyme sites	0		
Global percentage of GC	30-70%		
Local percentage of GC	35-65%		
Allowed complementary sequences (<10 bp)	Less than 10		
Allowed complementary sequences (>10 bp)	0		
Allowed length of repeated nucleotides	Less than 8 nucleotides		
RCB	0-0.2	0-0.4	0-0.5
Number of miRNAs binding sites	As low as possible		

- The Relative Codon Pair Bias (RCPB) [1]:

$$\mathrm{RCPB}(\gamma, \sigma) = \sum_{a,b=1}^{20} \frac{\eta_{ab}(\gamma)}{n-1} \sum_{\alpha \in \mathcal{K}(a), \beta \in \mathcal{K}(b)} \frac{1}{K(a)K(b)} \left| \frac{\vartheta_{\alpha\beta}(\sigma)}{\eta_{ab}(\gamma)} - \frac{\psi_{\alpha\beta}}{\varphi_{ab}} \right|, \tag{16}$$

where $\eta_{ab}(\gamma)$ is the number of times amino acid pair (a, b) appears in protein sequence γ and $\vartheta_{\alpha\beta}(\sigma)$ is the number of times codon pair (α, β) appears in the codon sequence σ. The same kind of caveats as for RCB hold for RCPB: the shorter the protein, the more likely that a given codon pair does not appear.

5.5. Application to the design of luciferase

The Firefly luciferase protein sequence (Uniprot: P08659) was selected, and its codons redesigned as explained in Section 5.3, using the parameter optimised from the human dataset. We selected sequences from $\beta = 1$, for the case of the Ind model, and $\beta = 3$, for both models, performing Monte Carlo simulations at this temperature as well as the lowest energy sequence for the NN (corresponding to $\beta = 600$). For $\beta = 1$ and 3, we wanted to select uncorrelated sequences that were representative of the variability of the solution space, so we saved sequences every 200 Monte Carlo sweeps.

The sequences obtained for each β value were filtered according to the criteria described below and summarised in Table 2. To avoid problems in the experimental protein expression, we discarded sequences that contain short nucleotide sequences targeted by the restriction enzymes used for plasmid linearisation during *in vitro* mRNA synthesis. Specifically, we excluded GCTCTTC (Seq_BspQI.1), GAAGAGC (Seq_BspQI.2), and GCGGCCGC (NotI). For the cases $\beta = 1$ and 3, this is done simply by discarding solutions containing these sub-sequences, while for the $T = 0$ case, we manually mutated the single solution, in such a way as to eliminate the forbidden sub-sequence, while producing the smallest increase in the energy.

Then, according to Ref. [46], the global and local GC percentage of the sequences were set between 30-70% and 35-65%, respectively. The global GC percentage is the average GC fraction of the whole sequence, whereas the local GC fractio was calculated using a fixed size window (30 nts). The number of complementary sequences of 10, 15 and 20 base pairs (bp) was then extracted from each sequence. This analysis was performed due to the potential formation of loops by complementary sequences, which could result in the generation of highly stable sequences, thereby potentially reducing translation efficiency. Therefore, the number of complementary sequences of 10 bp was limited to be less than 10, whereas for those of 15 and 20 bp, it was set to 0. Additionally, repeated nucleotides were avoided to reduce transcription errors, so the length of regions containing repeated nucleotides was set below 8 nucleotides.

The distance to human distribution was also calculated using the RCB (Eq. (15)) to determine the closeness to the human distribution.

This parameter differed for each selected database because as the value of β increases, the range in which this parameter can be found increases (Fig. 5). A range of 0-0.2, 0-0.3 and 0-0.4 was defined for β 1, 3 and ∞, respectively. Finally, we checked the number of microRNAs (miRNAs) binding sites contained in each sequence. MicroRNAs are short RNAs, of approximately 21 nucleotides, whose function is to regulate gene expression. Previous studies [47] have demonstrated that miRNA target sites located in the coding sequence of mRNAs may have an inhibitory effect on translation. Therefore, the aim was to identify sequences that interact with miRNAs and subsequently exclude them from our databases. To achieve this, we used a database [48] containing the sequences of various miRNAs identified in humans and the corresponding sequences of the proteins to which they bind. We found that miRNA binding sites were so common that it was practically impossible to exclude them all. Thus, we ranked the sequences according to the number of miRNAs binding target they contained and selected the sequences with as few binding sites as possible for the experiments.

5.6. DNA template design for in vitro transcription

The luciferase (GenBank: WP_212371658.1) DNA sequences that were optimised by our methods were gene synthetised by Genscript. Subsequently, the luciferase coding sequences were cloned into a previously designed plasmid. This plasmid, based on pUC57, comprised the following elements in the 5'-3' orientation: the T7 promoter for RNA polymerase, the 5'UTR and 3'UTR from human beta globin (GenBank: NM_000518), a polyA tail consisting of 100 adenines and, immediately following, the BspQI restriction enzyme site. In particular, the luciferase coding sequence was inserted in-frame directly following the 5'UTR. The plasmids obtained were amplified and purified by Genscript, and subsequently employed for *in vitro* transcription.

5.7. In vitro transcription

Each plasmid containing the DNA template sequence of interest was digested with BspQI (HONGENE, ON-124), which cleaved the plasmid immediately after the segment to be transcribed. Subsequently, the linearisation reaction was then purified using the Wizard ® SV Gel and PCR Clean-Up (Promega, A7270), in accordance with the manufacturer's instructions.

The purified linear DNA was subsequently employed for mRNA production by *in vitro* transcription using T7 RNA polymerase following manufacturer's instructions. Transcription reactions were performed at 37 °C for a period of three hours, utilising the following materials:

- Template linear DNA (50 μg/mL)
- T7 RNA polymerase (5000 U/mL; HONGENE, ON-004)
- RNase inhibitor (1000 U/mL, HONGENE, ON-039)
- Inorganic Pyrophosphatase (2 U/mL, HONGENE, ON-025)
- ATP (5 μg/mL, HONGENE, R1331)
- GTP (5 μg/mL, HONGENE, R2331)
- CTP (5 μg/mL, HONGENE, R3331)
- N1-Methylpseudouridine (5 μg/mL, HONGENE, R5-027)
- CleanCap ® AG (4 μg/mL, TRILINK ® N-7113-10)
- RNAse free double distilled water

5.8. mRNA purification

Following a three-hour incubation period incubation, DNAse I (Hongene, ON-109), was added to the generated mRNA transcripts and the incubation was continued for 15 minutes at 37ºC. The crude RNAs were purified by affinity chromatography using POROS Oligo (dT) 25 column (ThermoFisher). In particular, the buffers employed were as follows: Buffer A which contained 50 mM disodium phosphate, 0.5 M NaCl, 5 mM EDTA, pH 7.0 and Buffer B which contained 50 mM sodium dihydrogen phosphate, 5 mM EDTA, pH 7.0. The mRNA samples were initially half-diluted in Buffer A 2x. Subsequently, the column was equilibrated with 100% Buffer A, loaded with mRNA, washed with Buffer B, and ultimately eluted using double-deionised water. To completely eliminate Buffer B, mRNA was washed with a 30 kDa Amicon filter and then equilibrated through a one-tenth dilution in citrate buffer 10x with a pH 6.5.

The concentration of mRNA was determined by measuring the optical density at 260 nm, then adjusted to a final concentration of 1 mg/ml, aliquoted and stored at -80 °C until required.

For quality assurance, all mRNAs underwent analysis through automated electrophoresis (2100 Bioanalyzer, Agilent, G2938B). Subsequently, the mRNA samples were aliquoted and stored at -80 °C until needed.

5.9. Encapsulation of mRNA into lipid nanoparticles (LNP)

For *in vivo* administration, the mRNAs were encapsulated into lipid nanoparticles (LNPs) as described in [49]. Briefly, the purified mRNAs were initially diluted in sodium citrate buffer at pH 4 reaching a final concentration of 266 μg/ml. Simultaneously, lipids: SM-102 (BOCSI, 2089251-47-6); DOPE (Merk, 850725P); Cholesterol (Sigma, C3045); DMG-PEG2000 (Cayman, 33945-1) were dissolved into ethanol at the respective molar ratios of 50:10:38.5:1.5, maintaining a molar N:P ratio of 4.6:1.

LNPs were prepared using the pipette mixing method. This involved rapidly combining the aqueous solution with the ethanol solution, followed by homogenisation through pipetting up and down for 4-5 cycles. The resulting LNPs were promptly diluted 1:1 with buffer to a final concentration of Tris 20 mM, pH 8 and Sucrose 15%. Each resulting LNP solution was then collected, and encapsulated mRNA was assessed by Quant-IT® Ribogreen (Invitrogen, R11490) following the manufacturer's instructions. Additionally, LNPs were analysed through agarose gel electrophoresis to assure proper mRNA integrity and encapsulation percentage.

Subsequently, the LNP solution was adjusted to a final concentration of mRNA of 100 μg/ml, filtered through a 0.22 mm filter and then stored at -80 °C until required.

5.10. LNP characterisation

In addition to mRNA encapsulation, other relevant quality parameters of LNPs were also evaluated. Size distribution, polydispersity (PDI) and Z-potential were measured by dynamic light scattering (DLS) in a Zetasizer Advance Lab Blue Label (Malvern). The data obtained for each LNP are presented in Tables S2 and S3.

5.11. Cell culture and mRNA transfection

HeLa (DSMZ GmbH, ACC57) cell line was cultured on DMEM high glucose (Merk, D6429) supplemented with Fetal Bovine Serum 10% (Sigma, F7524), Penicillin-Streptomycin Solution 1% (Gibco^{TM}, 15140122) and Glutamax 2 mM (ThermoFisher, 35050038). Hepatic HepG2 (ATCC, HB-8065) cell line was cultured on RPMI 1640 (Gibco^{TM}, 31870074) supplemented with Fetal Bovine Serum 10% (Sigma, F7524), Penicillin-Streptomycin Solution 1% (Gibco^{TM}, 15140122) and Glutamax 2 mM (ThermoFisher, 35050038). Both cell lines were cultivated in a 175 cm^2 flask.

The day before transfection, cells were detached from the flask by trypsinisation (ThermoFisher, 11590626), and subsequently, they were seeded into 96-well plates at a density of 1×10^4 cells/well.

For transfection with a commercial cationic lipid, culture media was replaced with 90 μL of fresh media. Subsequently, a mixture of each mRNA (100 ng/well) and Lipofectamine MessengerMAX^{TM} (Invitrogen, 15397974; 0.2 μL/well), pre-incubated in OptiMEM media. The mRNA-lipofectamine mixture was added to the corresponding well in triplicate, directly resulting in a final mRNA concentration of 100 ng/well. Alternatively, the mRNA-Lipofectamine mixture was diluted to half or quarter of its concentration and then added to the cell culture, achieving final mRNA concentrations of 50 ng/well and 25 ng/well, respectively. For transfection with the mRNA-LNPs, an initial series of one-half serial dilutions were prepared in culture media. Subsequently, 25 μL/well of the corresponding mRNA-LNP were added in triplicates to 100 μL of cells culture, resulting in final mRNA concentrations of 100 ng/well, 50 ng/well or 25 ng/well.

The cells were incubated for 24 hours at 37 °C in a 5% CO_2 atmosphere, along with the mRNA-Lipofectamine MessengerMAX mixture or the mRNA-LNPs.

5.12. Quantification of the in-vitro activity of firefly luciferase

Cells were lysed 24 h post-transfection by incubation during 10 minutes with 100 μL of PBS-Triton 0.1%. Then, 98 μL of cell lysate was transferred to an opaque 96-well white plate. Buffered d-Luciferin (GoldBio, LUCK-100) in 100mM Tris-HCl pH 7.8, 5 mM $MgCl_2$, 250 μM CoA, 150 μM ATP buffer was added to each well in a volume of 100 μL, resulting in a final concentration of 150 μg/mL. The negative control was comprised of cells that had not been incubated with any mRNA. Luminescence was measured after 5 minutes of incubation at room temperature in a FLUOstar Omega plate reader (BMG LABTECH).

5.13. In solution stability of mRNA

To evaluate the stability of transcribed mRNA, samples were incubated at 37 ºC in 20 mM CHES buffer with 4 mM $MgCl_2$ at pH 9.7 over various time intervals (0, 10, 30, 60, 120, and 1440 minutes). To prevent further degradation during the handling process, the samples were promptly quenched samples by adding 160 mM Tris buffer containing 110 mM EDTA. RNA integrity was analysed using automated electrophoresis (2100 Bioanalyzer, Agilent, G2938B). The degree of degradation was determined through image analysis of the electropherogram, quantifying the integrated intensity of the smear over the full-length mRNA band.

Simultaneously, at the specified time points, aliquots were quenched, and 100 ng per well of mRNA were transfected into HeLa cells under the aforementioned standard conditions.

5.14. Administration of the mRNA-LNP to mice and in-vivo quantification of the activity of luciferase

Female BALB/c mice (Charles River Laboratories), 8-10-week-old and weighting 18–23 g, were acclimated to new conditions upon arrival at the experimental facilities for 3-7 days. Housing conditions were maintained at a room temperature 20-24 ºC, humidity 50-70%, and light intensity 60 lux, with a light-dark cycle of 12 hours.

To measure Firefly Luciferase activity in mice, LNPs were produced as previously described, containing 1 μg of the indicated mRNA in a final volume of 30 μL. These were then injected intramuscularly. At 4 and 24 hours post mRNA-LPN inoculation, mice were anesthetised by inhalation with 4% of Isoflurane using a vaporiser. The maintenance of the anaesthesia was sustained at 1.5% of Isoflurane. Then, D-luciferin (Quimigen, 12507) was intraperitoneally injected at 150 mg/kg, typically 200 μL of the stock at 15 mg/mL in PBS for a 20 g mouse. Luciferase images were captured 10 minutes after luciferin inoculation using the IVIS Lumina XRMS Imaging System, following manufacturer's instructions.

5.15. Statistical analysis for in vitro and in vivo experiments

In experimental studies, GraphPad Prism 10 software was used for representation and statistical analysis. The presence of outliers in the experimental data was investigated using the ROUT test with a Q value of 1%. No outliers were identified. To assess the normality of the experimental data, the Kolmogorov-Smirnov test was employed. Although some specific groups did not pass the normality test in the case of *in vitro* mRNA evaluation, given the predominantly normal distribution in the overall evaluation, we assumed normality for all groups.

Finally, an ordinary one-way ANOVA with Tukey's multiple comparisons post-test was employed for the comparison of experimental groups. The statistical significance of the results is indicated in Fig. 6, 7 and S9 by asterisks, as follows: *: P-value < 0.05; **: P-value < 0.01; ***: P-value < 0.001. The absence of asterisks indicates that the result is not statistically significant, with a P-value > 0.05.

Funding

Research reported in this publication was supported by Gobierno de Aragón (Spain) through project IDMF/2021/0009 (Nuevas tecnologías para el diseño y obtención de vacunas de ARN en Aragón) and E30_20R (Supercomputación Y Física De Sistemas Complejos Y Biológicos – COMPHYS). P.B. and D.L. acknowledge support through Grant PID2020-113582GB-I00 funded by MCIN/AEI/10.13039/501100011033.

Ethics approval

All animal procedures were carried out under Project Licence 59/21 approved by the Ethic Committee for Animal Experiments from the University of Zaragoza. The care and use of animals were performed accordingly with the Spanish Policy for Animal Protection RD53/2013, which meets the European Union Directive 2010/63 on the protection of animals used for experimental and other scientific purposes.

CRediT authorship contribution statement

David Luna-Cerralbo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Irene Blasco-Machín:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Susana Adame-Pérez:** Writing – review & editing, Investigation. **Verónica Lampaya:** Writing – review & editing, Investigation. **Ana Larraga:** Writing – review & editing, Investigation. **Teresa Alejo:** Writing – review & editing, Investigation. **Juan Martínez-Oliván:** Supervision, Funding acquisition, Conceptualization. **Esther Broset:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Pierpaolo Bruscolini:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

Juan Martínez-Oliván, Esther Broset, Susana Adame-Pérez, Verónica Lampaya, Ana Larraga, Teresa Alejo and Irene Blasco-Machín are employees at Certest Pharma Department, Certest Biotec S.L.

Data availability

The wild-type codon sequences for proteins reported in Table 1 are included as a fasta file in the Supplementary Information.

The code used to generate Ind and NN codon optimization models is available from the corresponding authors PB and EB on reasonable request. The database with the $L = 50$-codons-long sequences extracted from human NCBI repository can be obtained from corresponding authors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the revision of this work, the authors used DeepL in order to improve the language quality of some sentences. After using this tool, the authors reviewed and edited the content as needed, rejecting the suggestions that did not reflect the original meaning of the sentences. The authors take full responsibility for the content of the publication.

Acknowledgements

We would like to express our gratitude to all the members of Certest Biotec and, in particular, the Certest Pharma group, whose insightful contributions and expertise greatly assisted in this research. We are grateful in particular to Dr. Javier Giménez Warren, for kindly reading and improving the English language of this manuscript. Figure Schemes 6, 7, S9 and S12 draws were created using Mind the Graph platform (www.mindthegraph.com).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2024.07.020.

References

[1] Şen A, Kargar K, Akgün E, Pınar MÇ. Codon optimization: a mathematical programing approach. Bioinformatics 2020;36(13):4012–20. https://doi.org/10.1093/bioinformatics/btaa248.

[2] Grosjean H, Westhof E. An integrated, structure- and energy-based view of the genetic code. Nucleic Acids Res 2016;44(17):8020–40. https://doi.org/10.1093/nar/gkw608.

[3] Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, et al. A new and updated resource for codon usage tables. BMC Bioinform 2017;18:391. https://doi.org/10.1186/s12859-017-1793-7.

[4] Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. Cell Commun Signal 2020;18(1):145. https://doi.org/10.1186/s12964-020-00642-6.

[5] Mauro VP, Chappell SA. A critical analysis of codon optimization in human therapeutics. Trends Mol Med 2014;20(11):604–13. https://doi.org/10.1016/j.molmed.2014.09.003.

[6] Meyer D, Kames J, Bar H, Komar AA, Alexaki A, Ibla J, et al. Distinct signatures of codon and codon pair usage in 32 primary tumor types in the novel database CancerCoCoPUTs for cancer-specific codon usage. Gen Med 2021;13(1):122. https://doi.org/10.1186/s13073-021-00935-6.

[7] Kames J, Alexaki A, Holcomb DD, Santana-Quintero LV, Athey JC, Hamasaki-Katagiri N, et al. TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. J Mol Biol 2020;432(11):3369–78. https://doi.org/10.1016/j.jmb.2020.01.011.

[8] Krafczyk R, Qi F, Sieber A, Mehler J, Jung K, Frishman D, et al. Proline codon pair selection determines ribosome pausing strength and translation efficiency in bacteria. Commun Biol 2021;4(1):1–11. https://doi.org/10.1038/s42003-021-02115-z.

[9] Le Nouën C, Luongo CL, Yang L, Mueller S, Wimmer E, DiNapoli JM, et al. Optimization of the codon pair usage of human respiratory syncytial virus paradoxically resulted in reduced viral replication in vivo and reduced immunogenicity. J Virol 2020;94(2). https://doi.org/10.1128/jvi.01296--19.

[10] Sanguinetti M, Iriarte A, Amillis S, Marín M, Musto H, Ramón A. A pair of nonoptimal codons are necessary for the correct biosynthesis of the aspergillus nidulans urea transporter, UreA. R Soc Open Sci 2019;6(11):190773. https://doi.org/10.1098/rsos.190773.

[11] Lyu X, Liu Y. Nonoptimal codon usage is critical for protein structure and function of the master general amino acid control regulator CPC-1. mBio 2020;11(5). https://doi.org/10.1128/mbio.02605--20.

[12] Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, et al. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature 2013;495(7439):111–5. https://doi.org/10.1038/nature11833.

[13] Kurland C, Gallant J. Errors of heterologous protein expression. Curr Opin Biotechnol 1996;7(5):489–93. https://doi.org/10.1016/S0958-1669(96)80050-4.

[14] Chaney JL, Steele A, Carmichael R, Rodriguez A, Specht AT, Ngo K, et al. Widespread position-specific conservation of synonymous rare codons within coding sequences. PLoS Comput Biol 2017;13(5):e1005531. https://doi.org/10.1371/journal.pcbi.1005531.

[15] Perach M, Zafrir Z, Tuller T, Lewinson O. Identification of conserved slow codons that are important for protein expression and function. RNA Biol 2021;18(12):2296–307. https://doi.org/10.1080/15476286.2021.1901185.

[16] Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in escherichia coli. Science 2009;324(5924):255–8. https://doi.org/10.1126/science.1170160.

[17] Nieuwkoop T, Terlouw BR, Stevens KG, Scheltema RA, de Ridder D, van der Oost J, et al. Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. Nucleic Acids Res 2023;51(5):2363–76. https://doi.org/10.1093/nar/gkad035.

[18] Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet 2000;16(6):276–7. https://doi.org/10.1016/S0168-9525(00)02024-2.

[19] Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res 2005;33(Web Server issue):W526. https://doi.org/10.1093/nar/gki376.

[20] Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res 2007;35(Web Server issue):W126–31. https://doi.org/10.1093/nar/gkm219.

[21] Wu G, Bashir-Bello N, Freeland SJ. The synthetic gene designer: a flexible web platform to explore sequence manipulation for heterologous expression. Protein Expr Purif 2006;47(2):441–5. https://doi.org/10.1016/j.pep.2005.10.020.

[22] Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S. Gene designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinform 2006;7(1):285. https://doi.org/10.1186/1471-2105-7-285.

[23] Gaspar P, Oliveira JL, Frommlet J, Santos MA, Moura G. EuGene: maximizing synthetic gene design for heterologous expression. Bioinformatics 2012;28(20):2683–4. https://doi.org/10.1093/bioinformatics/bts465.

[24] Chin JX, Chung BK-S, Lee D-Y. Codon optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. Bioinformatics 2014;30(15):2210–2. https://doi.org/10.1093/bioinformatics/btu192.

[25] Guimaraes JC, Rocha M, Arkin AP, Cambray G. D-tailor: automated analysis and design of DNA sequences. Bioinformatics 2014;30(8):1087–94. https://doi.org/10.1093/bioinformatics/btt742.

[26] Karaşan O, Şen A, Tiryaki B, Cicek AE. A unifying network modeling approach for codon optimization. Bioinformatics 2022;38(16):3935–41. https://doi.org/10.1093/bioinformatics/btac428.

[27] Taneda A, Asai K. COSMO: a dynamic programming algorithm for multicriteria codon optimization. Comput Struct Biotechnol J 2020;18:1811–8. https://doi.org/10.1016/j.csbj.2020.06.035.

[28] Fu H, Liang Y, Zhong X, Pan Z, Huang L, Zhang H, et al. Codon optimization with deep learning to enhance protein expression. Sci Rep 2020;10(1):17617. https://doi.org/10.1038/s41598-020-74091-z.

[29] Emboss backtranseq. https://www.ebi.ac.uk/jdispatcher/st/emboss_backtranseq.

[30] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA package 2.0. Algorithms Mol Biol 2011;6:26. https://doi.org/10.1186/1748-7188-6-26.

[31] Pardi N, Hogan MJ, Porter FW, Weissman D. mRNA vaccines — a new era in vaccinology. Nat Rev Drug Discov 2018;17(4):261–79. https://doi.org/10.1038/nrd.2017.243.

[32] Gonzalez-Sanchez B, Vega-Rodríguez MA, Santander-Jiménez S, Granado-Criado JM. Multi-objective artificial bee colony for designing multiple genes encoding the same protein. Appl Soft Comput 2019;74:90–8. https://doi.org/10.1016/j.asoc.2018.10.023.

[33] Sonneveld S, Verhagen BMP, Tanenbaum ME. Heterogeneity in mRNA translation. Trends Cell Biol 2020;30(8):606–18. https://doi.org/10.1016/j.tcb.2020.04.008.

[34] Niederer RO, Rojas-Duran MF, Zinshteyn B, Gilbert WV. Direct analysis of ribosome targeting illuminates thousand-fold regulation of translation initiation. Cell Syst 2022;13(3):256–2643.e. https://doi.org/10.1016/j.cels.2021.12.002.

[35] Kim SC, Sekhon SS, Shin W-R, Ahn G, Cho B-K, Ahn J-Y, et al. Modifications of mRNA vaccine structural elements for improving mRNA stability and translation efficiency. Mol Cell Toxicol 2022;18(1):1–8. https://doi.org/10.1007/s13273-021-00171-4.

[36] Torabi S-F, Chen Y-L, Zhang K, Wang J, DeGregorio SJ, Vaidya AT, et al. Structural analyses of an RNA stability element interacting with poly(A). Proc Natl Acad Sci 2021;118(14):e2026656118. https://doi.org/10.1073/pnas.2026656118.

[37] Chheda U, Pradeepan S, Esposito E, Strezsak S, Fernandez-Delgado O, Kranz J. Factors affecting stability of RNA – temperature, length, concentration, pH, and buffering species. J Pharm Sci 2024;113(2):377–85. https://doi.org/10.1016/j.xphs.2023.11.023.

[38] Grzybowska EA, Wakula M. Protein binding to cis-motifs in mRNAs coding sequence is common and regulates transcript stability and the rate of translation. Cells 2021;10(11):2910. https://doi.org/10.3390/cells10112910.

[39] Genuth NR, Barna M. The discovery of ribosome heterogeneity and its implications for gene regulation and organismal life. Mol Cell 2018;71(3):364–74. https://doi.org/10.1016/j.molcel.2018.07.018.

[40] Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Nucleic Acids Res 2023;51(D1):D933–41. https://doi.org/10.1093/nar/gkac958.

[41] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2021;50(D1):D20–6. https://doi.org/10.1093/nar/gkab1112.

[42] Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] - [cited YYYY Mmm DD]. Available from: https://www.ncbi.nlm.nih.gov/nucleotide/.

[43] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983;220(4598):671–80. https://doi.org/10.1126/science.220.4598.671.

[44] Sharp PM, Li WH. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15(3):1281–95.

[45] Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. Virus attenuation by genome-scale changes in codon pair bias. Science 2008;320(5884):1784–7. https://doi.org/10.1126/science.1155761.

[46] Fan Long. Codon optimization; Jun. 2020.

[47] Hausser J, Syed AP, Bilen B, Zavolan M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. Genome Res 2013;23(4):604–15. https://doi.org/10.1101/gr.139758.112.

[48] Sticht C, Torre CDL, Parveen A, Gretz N. miRWalk: an online resource for prediction of microRNA binding sites. PLoS ONE 2018;13(10):e0206239. https://doi.org/10.1371/journal.pone.0206239.

[49] Hassett KJ, Benenato KE, Jacquinet E, Lee A, Woods A, Yuzhakov O, et al. Optimization of lipid nanoparticles for intramuscular administration of mRNA vaccines. Molecular Therapy Nucleic Acids 2019;15:1–11. https://doi.org/10.1016/j.omtn.2019.01.013.