PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Analysis of 11,430 recombinant protein production experiments reveals that protein yield is tunable by synonymous codon changes of translation initiation sites

Bikash K. Bhandari[1☯], **Chun Shen Lim**[1☯], **Daniela M. Remus**[2], **Augustine Chen**[1], **Craig van Dolleweerd**[3], **Paul P. Gardner**[1,3*]

1 Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand, **2** Callaghan Innovation Protein Science and Engineering, University of Canterbury, Christchurch, New Zealand, **3** Biomolecular Interaction Center, University of Canterbury, Christchurch, New Zealand

☯ These authors contributed equally to this work.
* paul.gardner@otago.ac.nz

Citation: Bhandari BK, Lim CS, Remus DM, Chen A, van Dolleweerd C, Gardner PP (2021) Analysis of 11,430 recombinant protein production experiments reveals that protein yield is tunable by synonymous codon changes of translation initiation sites. PLoS Comput Biol 17(10): e1009461. https://doi.org/10.1371/journal.pcbi.1009461

Editor: Eugene I. Shakhnovich, Harvard University, UNITED STATES

Received: June 2, 2021

Accepted: September 19, 2021

Published: October 5, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1009461

Data Availability Statement: Our code and data can be found in our GitHub repository (https://

Abstract

Recombinant protein production is a key process in generating proteins of interest in the pharmaceutical industry and biomedical research. However, about 50% of recombinant proteins fail to be expressed in a variety of host cells. Here we show that the accessibility of translation initiation sites modelled using the mRNA base-unpairing across the Boltzmann's ensemble significantly outperforms alternative features. This approach accurately predicts the successes or failures of expression experiments, which utilised *Escherichia coli* cells to express 11,430 recombinant proteins from over 189 diverse species. On this basis, we develop TIsigner that uses simulated annealing to modify up to the first nine codons of mRNAs with synonymous substitutions. We show that accessibility captures the key propensity beyond the target region (initiation sites in this case), as a modest number of synonymous changes is sufficient to tune the recombinant protein expression levels. We build a stochastic simulation model and show that higher accessibility leads to higher protein production and slower cell growth, supporting the idea of protein cost, where cell growth is constrained by protein circuits during overexpression.

Author summary

Recombinant proteins are widely used as therapeutics, such as vaccines, monoclonal antibodies, hormones and enzymes. However, the success rate of recombinant protein production is about 50%. To address this problem, we propose optimising the unpairing propensities of nucleotides around translation initiation sites using a thermodynamic quantity called mRNA accessibility. Our study shows that this method is generalisable across prokaryotic and eukaryotic expression hosts. Importantly, we validated this method using laboratory experiments and computational modelling. Furthermore, we propose a

github.com/Gardner-BinfLab/TIsigner_paper_2019). These include the scripts and Jupyter notebooks to reproduce our results and figures. The source code of TIsigner is available at https://github.com/Gardner-BinfLab/TISIGNER-ReactJS. The public web version of this tool runs at https://tisigner.com/tisigner. The experimental data, analysis and results are available at https://github.com/bkb3/TIsignerExperiment/tree/master/Jupyter and an interactive version of results are available at https://bkb3.github.io/TIsignerExperiment/.

Funding: This work was supported in part by the Ministry of Business, Innovation and Employment (https://www.mbie.govt.nz/) [MBIE Smart Idea grant: UOOX1709 to P.P.G. and C.D., and MBIE Data Science Programmes grant: UOAX1932 to P. P.G.] and the Royal Society of New Zealand Te Apārangi (https://www.royalsociety.org.nz/) [Marsden grant: 19-UOO-040 to P.P.G.]. B.K.B was also supported in part by the University of Otago Postgraduate Publishing Bursary (Doctoral). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

low cost technique to tune protein expression by engineering minimal changes to genes of interest through our web application (https://tisigner.com/tisigner).

Introduction

Recombinant protein expression has numerous applications in biotechnology and biomedical research. Despite extensive refinements in protocols over the past three decades, half of the experiments fail in the expression phase (http://targetdb.rcsb.org/metrics/). Notable problems are the low expression of 'difficult-to-express' proteins such as those found in, or associated with, membranes, and the poor growth of the expression hosts, which may relate to toxicity of heterologous proteins [1] (see [2, 3] for detailed reviews). Despite these issues, mRNA abundance only explains up to 40% of the variation in protein abundance, presumably due to variation in translation and turnover rates [4–10].

For *Escherichia coli*, mainstream models that may explain the lower-than-expected correlation between mRNA and protein levels are codon-usage and mRNA structure. Codon analysis is based on the frequency of codon usage in highly expressed proteins using codon adaptation index (CAI) [11] or tRNA adaptation index (tAI)—these are thought to capture tRNA availability which may influence translation rates [12, 13]. On the other hand, stable mRNA structures are thought to impede the assembly and progress of ribosomes on mRNAs [14–16]. More recent studies show stronger support for models based on mRNA folding, in which the stability of RNA structures, usually estimated using nearest-neighbour minimum-free energy (MFE) models, around the Shine-Dalgarno sequence and translation initiation site (e.g., AUG start codon) inversely correlates with protein expression [15–20]. We recently proposed a third model in which the avoidance of inappropriate interactions between mRNAs and non-coding RNAs (ncRNAs) has a strong effect on protein expression [21]; in addition, we evaluate a related measure 'accessibility' which considers all possible intramolecular base-unpairing probabilities [22]. Many of these features are interdependent, which presents a major challenge for identifying useful features.

The existing algorithms for gene optimisation that sample synonymous protein-coding sequences use models based on CAI, tAI, MFE, and/or G+C content (%) [23–27]. However, these models are usually evaluated on relatively small numbers of endogenous proteins, reporter proteins, or heterologous proteins with synonymous variants, often with poor separation of training and test datasets. It is unclear whether these features are generalisable to explain the expression of all heterologous proteins. To address this question, we have used data from multiple large-scale, non-redundant recombinant protein production experiments (N = 11,430), proteomics (N = 3,725), and fluorescent reporters (N = 82,002) from bacterial and eukaryotic species in order to identify mRNA features that best explain variation in protein abundance. This problem has not previously been investigated with this scale of heterogeneous datasets in any previous study.

We find that mRNA accessibility is a single best predictor of protein expression across the datasets, and accurately predicts the successes and failures of 11,430 experiments of recombinant protein expression in *E. coli*. Specifically, the accessibility of translation initiation sites outperform other mRNA features by capturing all possible optimal or suboptimal structures beyond translation initiation sites. With this information, we propose how accessibility can be exploited to fine-tune recombinant protein expression at a low cost. Specifically, we built a web server called TIsigner (Translation Initiation coding region designer), which optimises a protein-coding sequence by suggesting synonymous codon changes within the first nine

codons. Therefore, our approach makes gene optimisation accessible, as PCR can be used rather than an expensive full-length gene synthesis.

Results

Accessibility of translation initiation sites strongly correlates with protein abundance

To identify an accurate model of mRNA structure that explains protein expression, we examined an *E. coli* expression dataset of green fluorescent protein (GFP) fused in-frame with a library of 96-nt upstream sequences (N = 244,000 variants) [16]. These 96-nt sequences were generated to achieve a full factorial design by varying A+T content (%), CAI, codon ramp bottleneck position and strength, hydrophobicity of the encoded peptide, and minimum free energy (MFE). We removed redundancy of these 96-nt upstream sequences by clustering on sequence similarity, giving rise to 14,425 representative sequences. We calculated the accessibility (also known as 'opening energy' based on unpairing probability) for all the corresponding sub-sequences (Fig 1, see the definitions and equations in Methods).

Previous studies have defined mRNA accessibility using different terms [14, 24, 28–33] (see Additional notes, S2 File). Here we use a partition function method implemented in RNAplfold to calculate accessibility [22] (Fig 1, see Methods). We examined the correlation between the opening energies and GFP levels. We found that the opening energies of translation initiation sites, in particular from the nucleotide positions −30 to 18 (−30:18), shows the highest correlation with protein abundances (Fig 2A; Spearman's correlation, $R_s = -0.65$, $P < 2.2 \times 10^{-16}$). This is stronger than the highest correlation between the MFE at the region −30:30 (MFE −30:30) and protein abundance, which was previously reported as the highest

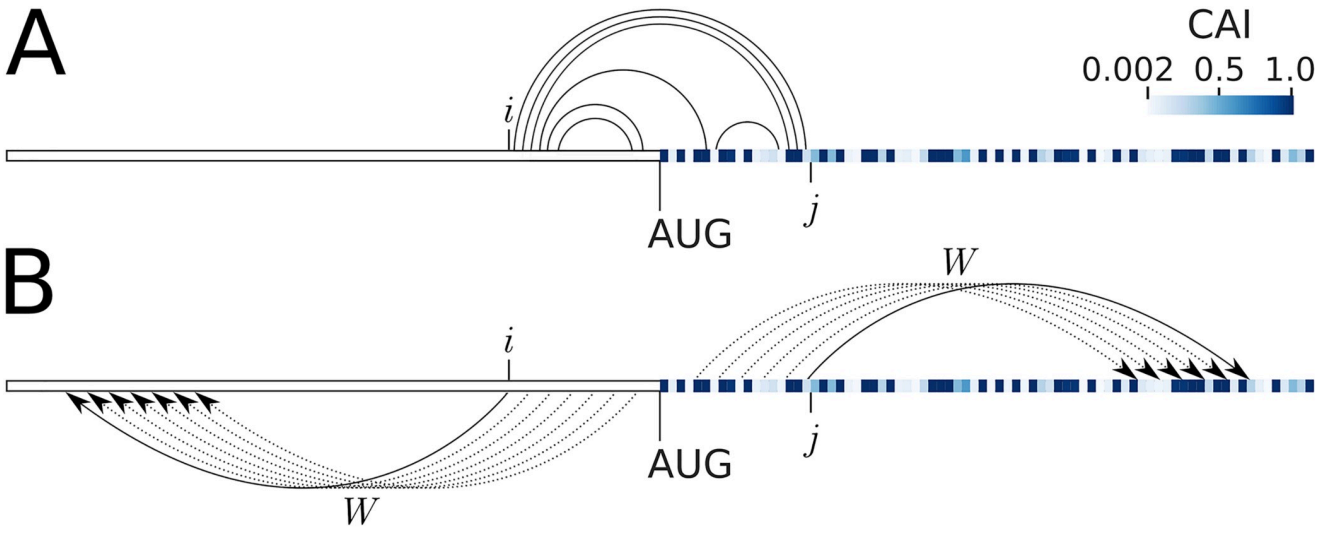

Fig 1. Opening energy (accessibility) has greater contextual information than minimum free energy (MFE). Schematic representation and interpretation of MFE and opening energy of a mRNA sequence (GFP). Codons are color coded using the weights of Codon Adaptation Index (CAI) from Sharp and Li (1987) [11]. A: The computation of MFE of the region *i*. . .*j* results in finding a single most stable structure. This stable structure contains pairings within the region as indicated by arcs. Hence, this approach is unable to detect any information beyond the target region (initiation site in this case). For example, a change in CAI after the nucleotide at position *j* does not affect the MFE. In addition, the predicted single structure may not be present under physiological conditions. B: The computation of opening energy of the region *i*. . .*j* uses several windows, each of length *W* nucleotides, shown by dotted and solid arrows (flanking window). The partition function used to determine the opening energy is computed over all possible structures (optimal and suboptimal) from the Boltzmann's ensemble where the region *i*. . .*j* is unpaired (Methods). As these windows could be extended well beyond the target region, opening energy contains additional contextual information. For example, a change of CAI beyond the nucleotide *j* could influence the opening energy (Results, Accessibility captures the full ensemble average energy of a sequence).

https://doi.org/10.1371/journal.pcbi.1009461.g001

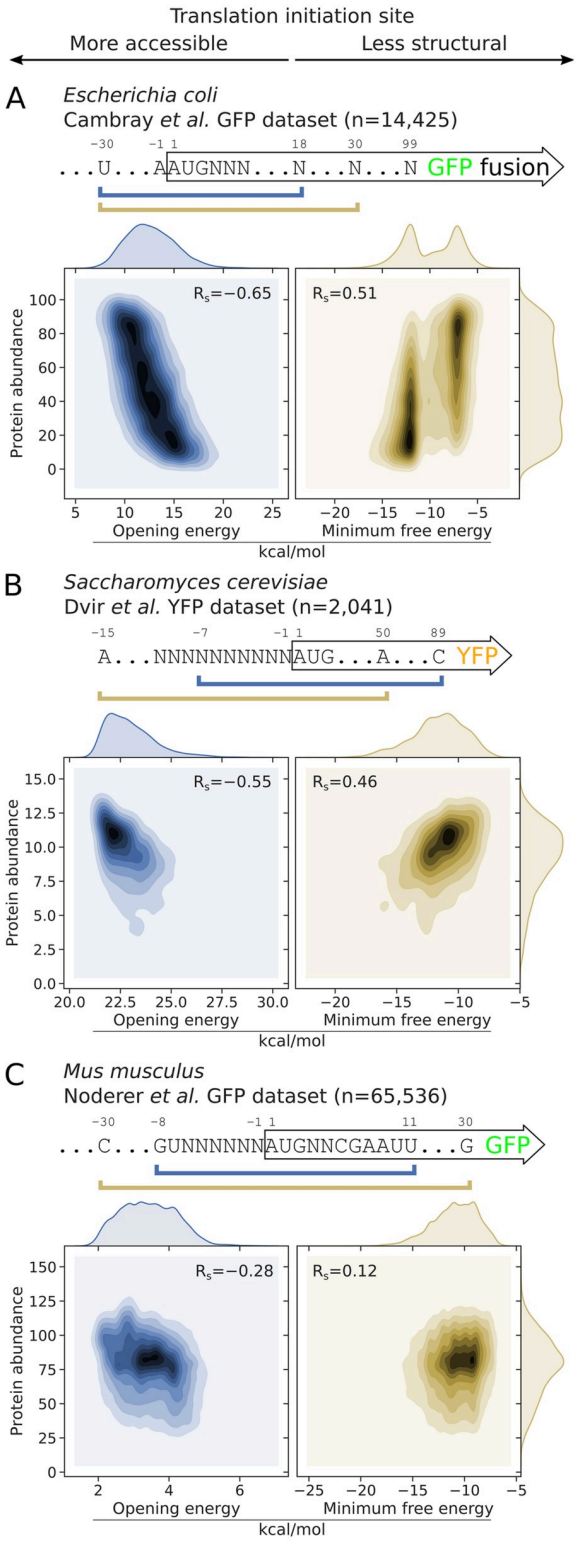

Fig 2. Correlations between the opening energies of translation initiation sites and protein abundances are stronger than that of minimum free energies (MFE). A: For *E. coli*, the opening energy at the region −30:18 shows the strongest correlation with protein abundance (see also Fig 3A and 3B, or Fig A in S1 Fig, sub-sequence l = 48 at position i = 18). For this analysis, we used a representative green fluorescent protein (GFP) expression dataset from Cambray et al. (2018) [16]. The reporter library consists of GFP fused in-frame with a library of 96-nt upstream sequences (N = 14,425). The MFE at the region −30:30 (MFE −30:30) shown was determined by Cambray et al. (right panel). B:

For *S. cerevisiae*, the opening energy −7:89 shows the strongest correlation with protein abundance (see also Fig B in S1 Fig, sub-sequence l = 96 at position i = 89. For this analysis, we used the yellow fluorescent protein (YFP) expression dataset from Dvir et al. (2013) [19]. The YFP reporter library consists of 2,041 random decameric nucleotides inserted at the upstream of YFP start codon. The MFE −15:50 was previously shown to correlate the best with protein abundance (right panel). C: For *M. musculus*, the opening energy −8:11 shows the strongest correlation with protein abundance (see also Fig C in S1 Fig, sub-sequence l = 19 at position i = 11). For this analysis, we used the GFP expression dataset from Noderer et al. (2014) [34]. The GFP reporter library consists of 65,536 random hexameric and dimeric nucleotides inserted at the upstream and downstream of GFP start codon, respectively. The MFE −30:30 was shown (right panel). See also S1 File. R_s, Spearman's rho. The Bonferroni adjusted P-values are below the machine's underflow level for the correlations between opening energies and protein abundances shown in the left panels.

https://doi.org/10.1371/journal.pcbi.1009461.g002

ranked feature [Fig 2A; $R_s = 0.51$, $P < 2.2 \times 10^{-16}$ (right panel)]. To account for multiple-testing, the P-values were adjusted using Bonferroni's correction and reported to machine precision.

We repeated the analysis for a dataset of yellow fluorescent protein (YFP) expression in *Saccharomyces cerevisiae* [19]. This dataset corresponds to a library of 5′UTR variants, in which the 10-nt sequences preceding the YFP translation initiation site were randomly substituted (N = 2,041 variants). In this case, the opening energy −7:89 showed a stronger correlation with protein abundance than that of the MFE −15:50 reported previously (Fig 2B; $R_s = -0.55$ versus 0.46).

To examine the usefulness of accessibility in complex eukaryotes, we analysed a dataset of GFP expression in *Mus musculus* [34]. The reporter library was originally designed to measure the strength of translation initiation sequence context, in which all possible substitutions were made at the flanking regions of the GFP translation initiation site (6-nt upstream region and 2-nt downstream region of initiation codon; N = 65,536 variants). Here the opening energy −8:11 showed a maximum correlation with expressed proteins, which again, is stronger than that of the MFE −30:30 (Fig 2C; $R_s = -0.28$ versus 0.12).

Taken together, our findings suggest that the accessibility of translation initiation sites strongly correlates with protein abundance across species. Interestingly, our findings in *E. coli* also suggest that the surrounding region of initiation sites, including the Shine-Dalgarno sequence [35] at −13:−8, should be accessible, presumably in order to recruit ribosomes. In contrast, the Shine-Dalgarno sequence is absent in yeasts and complex eukaryotes, which may explain why the computed accessibility regions begin at positions ≥ -8. In eukaryotes, the 43S preinitiation complexes scan from the 5′-cap end of the mRNAs [36]. This mechanism employs helicases such that the RNA structures preceding initiation codons are scanned through. However, caution is in order here, as large-scale recombinant protein production datasets for these eukaryotes are not available to validate these findings. Further investigation into the differences in the mechanisms of translation initiation between prokaryotes and eukaryotes would be useful to explain why these mRNA regions are distinct.

Theoretically, bacterial 30S subunits can initiate at any position as non-AUG initiation codons are more common in bacteria than eukaryotes, and most bacterial mRNAs are polycistronic. However, in agreement with previous high-throughput RNA structural probing studies, we found that the regions −30:30 of bacterial mRNAs are significantly less structured and are A-rich [37–40]. We reasoned that accessibility is likely more important in bacteria than eukaryotes (Fig 2, stronger correlation between accessibility and protein abundance). High accessibility of initiation sites likely improves a greater selectivity in translation initiation in bacteria.

Accessibility predicts the outcome of recombinant protein expression

We investigated how accessibility performs in the real world in prediction of recombinant protein expression. For this purpose, we carefully curated and analysed 11,430 expression

experiments in *E. coli* from the 'Protein Structure Initiative:Biology' (PSI:Biology) [41–44]. These PSI:Biology targets were expressed using the pET21_NESG expression vector that harbours the T7lac inducible promoter and a C-terminal His tag [43].

We divided the experimental results of the PSI:Biology targets into protein expression 'success' and 'failure' groups that were previously curated by DNASU (8,780 'Protein_Confirmed' and 2,650 'Tested_Not_Found' determined by SDS-PAGE analysis, respectively; see S2 Fig). These PSI:Biology targets span more than 189 species and the failures are representative of various problems in heterologous protein expression. Only 1.6% of the targets were *E. coli* proteins, which is negligible (N = 179; see S2 Fig).

We calculated the opening energies for all possible sub-sequences of the PSI:Biology targets as above (Fig 3, positions relative to initiation codons). For each sub-sequence region, we used the opening energies to predict the expression outcomes and computed the prediction accuracy using the area under the receiver operating characteristic curve (AUC; see Fig 3C). A closer look into the correlations between opening energies and expression outcomes, and AUC scores calculated for the sub-sequence regions reveals a strong accessibility signal of translation initiation sites (Fig 3B and 3C, Cambray's GFP and PSI:Biology datasets, respectively). We matched the correlations and AUC scores by sub-sequence regions and confirmed that sub-sequence regions that have strong correlations are likely to have high AUC scores (Fig 3D). In contrast, the sub-sequence regions that have zero correlations are not useful for predicting the expression outcomes (AUC approximately 0.5).

We then asked how accessibility manifests in the endogenous mRNAs of *E. coli*, for which we studied a proteomics dataset of 3,725 proteins available from PaxDb [45]. As expected, we observed a similar accessibility signal, with the region −25:16 correlated the most with protein abundance (Fig 3E, sub-sequence l = 41 at position i = 16). However, the correlation was rather low ($R = -0.17$, $P < 2.2 \times 10^{-16}$), which may reflect the limitation of mass spectrometry to detect lower abundances [46, 47]. Furthermore, the endogenous promoters have variable strength, which gives rise to a broad range of mRNA and protein levels [48, 49]. Taken together, our results show that the accessibility signal of translation initiation sites is very consistent across various datasets analysed (Fig 3 and S1 Fig).

Accessibility outperforms other features in prediction of recombinant protein expression

To choose an accessibility region for subsequent analyses, we selected the top 200 regions from the above correlation analysis on Cambray's GFP dataset (Fig 3B) and used random forest to rank their Gini importance scores in prediction of the outcomes of the PSI:Biology targets. The region −24:24 was ranked first (Fig 3B, l = 48 and i = 24), which is nearly identical to the region −23:24 with the top AUC score (Fig 3C, l = 47 and i = 24, AUC = 0.70). We therefore used the opening energy at the region −24:24 in subsequent analyses (Fig 4A). Interestingly, both the Shine-Dalgarno sequence (−13:−8 or l = 21 and i = −8) and initiation codons (l = 3 and i = 3) have weaker correlations and AUC scores than the region −24:24 (Fig 3B, 3C and 3E). This suggests that a slightly larger region around these key motifs provides a better context of accessibility and thus a better prediction of protein expression.

We asked how the other features perform compared to accessibility in prediction of heterologous protein expression, for which we analysed the same PSI:Biology dataset. We first calculated the MFE and avoidance at the regions −30:30 and 1:30, respectively. These are the local features associated with translation initiation rate. We also calculated CAI [11], tAI [50], codon context (CC) [51], G+C content, and Iχnos scores [52]. CC is similar to CAI except it takes codon-pairs into account, whereas the Iχnos scores are translation elongation rates

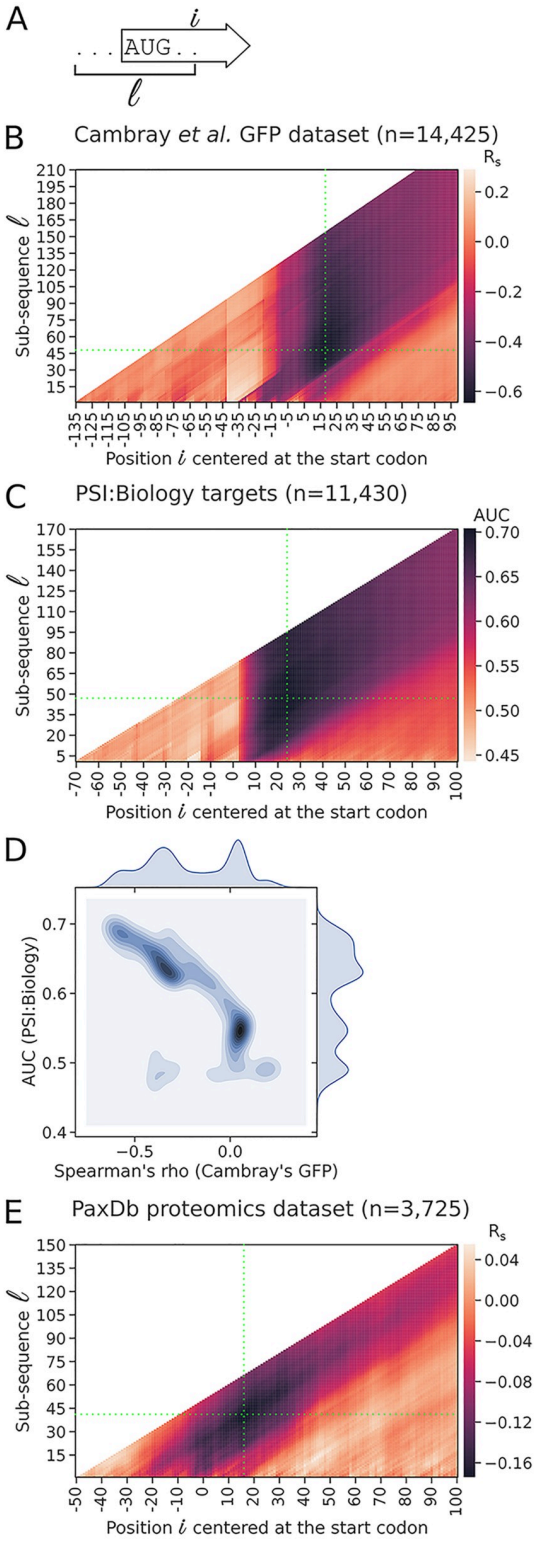

Fig 3. Opening energies of regions surrounding the Shine-Dalgarno and start codons are predictive of protein expression in *E. coli*. A: Schematic representation of a transcript sub-sequence l at position i for the calculation of opening energy. For example, the sub-sequence l = 10 at position i = 10 corresponds to the region 1:10. B: Correlations between the opening energies for the sub-sequences of GFP transcripts and protein abundances. The opening energy at the region −30 to 18 nt (sub-sequence l = 48 at position i = 18, green crosshair) shows the strongest correlation with

protein abundance [R_s = −0.65; N = 14,425, GFP expression dataset of Cambray et al. (2018)]. For this dataset, the reporter plasmid used is pGC4750, in which the promoter and ribosomal binding site are oFAB1806 inducible promoter and oFAB1173/BCD7, respectively. C: Prediction accuracy of the expression outcomes of the PSI:Biology targets using opening energy (N = 11,430). The opening energy at the region −23:24 (sub-sequence l = 47 at position i = 24, green crosshair) shows the highest prediction accuracy score (AUC = 0.70). For this dataset, the expression vector used is pET21_NESG, in which the promoter and fusion tag are T7lac and C-terminal His tag, respectively. D: Comparison between the correlations and AUC scores by sub-sequence region taken from the above analyses. The sub-sequence regions that have strong correlations are likely to have high AUC scores, whereas the sub-sequence regions that have no correlations are likely not useful in prediction of the expression outcomes. E: Correlations between the opening energies for the sub-sequences of *E. coli* transcripts and protein abundances. The transcripts used for this analysis are protein-coding sequences concatenated with 50 and 10 nt located upstream and downstream, respectively. The opening energy at the region −25:16 (sub-sequence l = 41 at position i = 16, green crosshair) shows the strongest correlation with protein abundance (R_s = −0.17; N = 3,725, PaxDb integrated proteomics dataset). See also S1 File. R_s, Spearman's rho.

https://doi.org/10.1371/journal.pcbi.1009461.g003

predicted using a neural network model trained with ribosome profiling data (S3 Fig). These are the global features associated with translation elongation rate. We built a random forest model to rank the Gini importance scores of these local and global features. The local features ranked higher than the global features (Fig 4B). We then calculated and compared the prediction accuracy of these features. The AUC scores for the local features were 0.70, 0.67 and 0.62 for the opening energy, MFE and avoidance, respectively, whereas the global features were 0.58, 0.57, 0.54, 0.54 and 0.51 for Iχnos, G+C content, CAI, CC and tAI, respectively (Fig 4C). The local features outperform the global features, suggesting that effects on translation initiation are a major predictor of the outcome of heterologous protein expression. We further examined the local G+C contents corresponding to the local features (S4 Fig). The G+C contents in the regions −24:24 and −30:30 weakly correlate with opening energy and MFE, respectively. The AUC scores for these local G+C contents are also lower than the corresponding local features, suggesting that these local G+C contents are not good proxies for the corresponding local features. Overall, our findings support previous reports that the effects on translation initiation are rate-limiting [15, 20] which, interestingly, correlate with the binary outcome of recombinant protein expression (Fig 4D). Importantly, accessibility significantly outperformed all other features (Fig 4C, see confidence intervals of AUC scores).

To identify a good opening energy threshold, we calculated positive likelihood ratios for different opening energy thresholds using the cumulative frequencies of true negative, false negative, true positive and false positive derived from the above receiver operating characteristic (ROC) analysis (top panel in S5 Fig). Meanwhile, we calculated the 95% confidence intervals of these positive likelihood ratios using 10,000 bootstrap replicates. We reasoned that there is an upper and lower bound on translation initiation rate, therefore the relationship between translation initiation rate and accessibility is likely to follow a sigmoidal pattern. We fit the positive likelihood ratios into a four-parametric logistic regression model (S5 Fig). As a result, we are 95% confident that an opening energy of 10 kcal/mol or below at the region −24:24 is about two times more likely to belong to the sequences which are successfully expressed than those that failed. To allow easy interpretation of results, we derived 'Expression Score' from this logistic regression curve with a scale from (see Methods).

Accessibility captures the full ensemble average energy of a sequence

To illustrate the advantage of accessibility over MFE, we analysed a proteomic dataset of *E. coli* cells infected with bacteriophage T7 [53]. The major capsid gene was codon-deoptimised to generate a mutant T7 strain [53]. Specifically, the first and the last 14 codons of the major

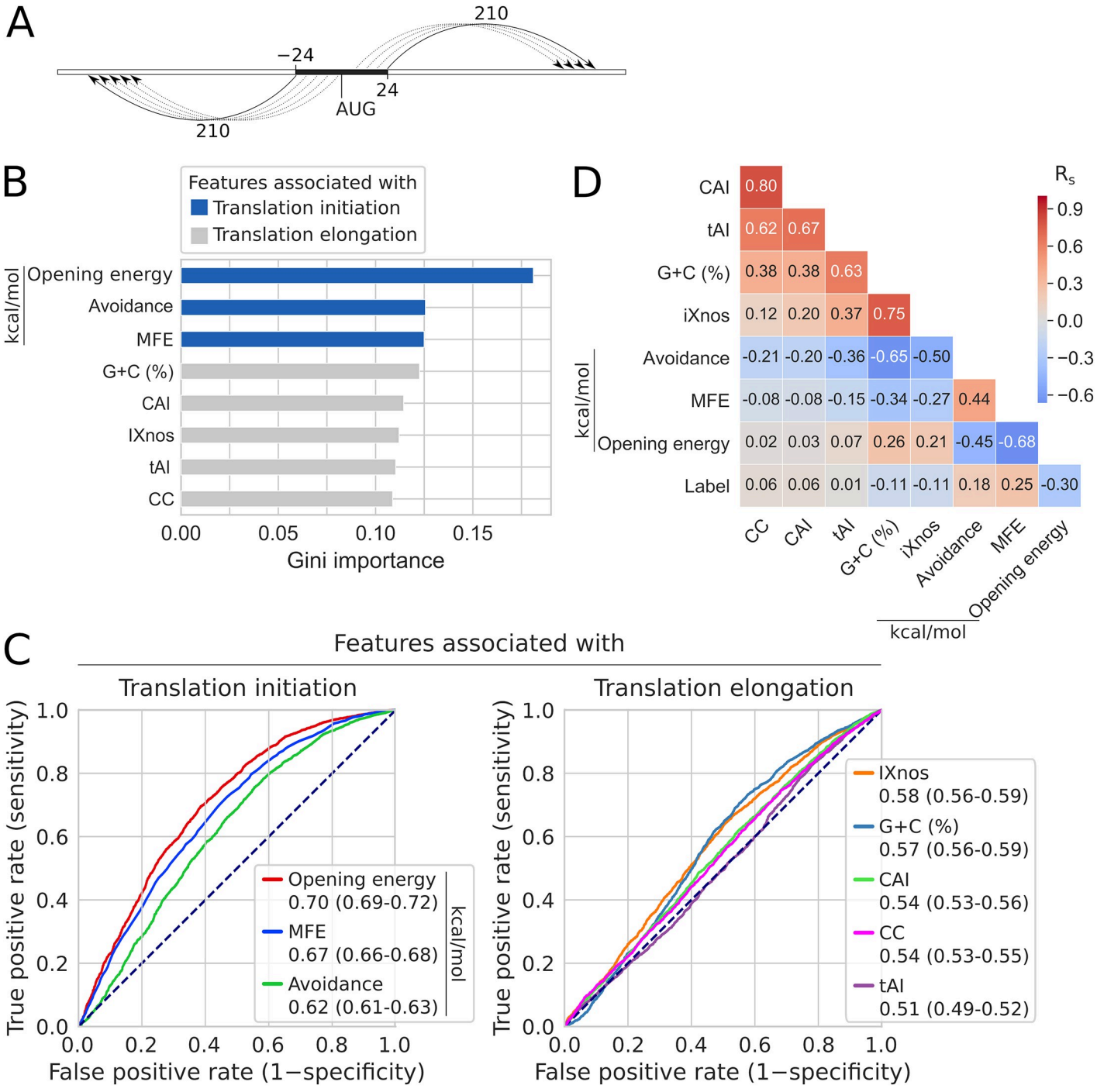

Fig 4. Accessibility of translation initiation sites is the strongest predictor of heterologous protein expression in *E. coli*. A: A partition function approach to compute the opening energy of the region −24:24 (solid black) in this analysis. Each window (arrow) is of the length of 210 nucleotides. The solid arrows represent the flanking windows. The dotted arrows represent other windows in between. Thus, the computation of opening energy for the region −24:24 captures the unpairing propensities of the surrounding region. It should be noted that the sliding window is constrained by the lengths of the flanking sequences. This partition function approach can be customised and executed using the algorithm implemented in RNAplfold. B: mRNA features ranked by Gini importance for random forest classification of the expression outcomes of the PSI:Biology targets (N = 8,780 and 2,650, 'success' and 'failure' groups, respectively). The features associated with translation initiation rate (blue; opening energy −24:24, minimum free energy (MFE) −30:30, and mRNA:ncRNA avoidance 1:30) have higher scores than the feature associated with translation elongation rate [grey; tRNA adaptation index (tAI), codon context (CC), codon adaptation index (CAI), G+C content (%), and Iχnos]. The Iχnos scores are translation elongation rates predicted using a neural network model trained with ribosome profiling data (S3 Fig). C: ROC analysis shows that accessibility (opening energy −24:24) has the highest classification accuracy. The AUC scores with 95% confidence intervals are shown. See also S1 File. D: Accessibility (opening energy −24:24) is the best feature in explaining the expression outcomes. Outcomes are represented with 'Label'. For this dataset, these labels were binary. MFE, Minimum Free Energy; R_s, Spearman's rho.

https://doi.org/10.1371/journal.pcbi.1009461.g004

capsid gene remained unchanged, making this a good case study to compare accessibility, MFE, and CAI.

We calculated the opening energies, MFEs and CAIs of the wild-type and mutant sequences (S6 Fig, orange and blue colours, respectively). We also scored the wild-type and mutant sequences using the above logistic regression curve, and obtained the approximated 'Expression Scores' of 89 and 38, respectively (opening energies of 9.05 kcal/mol and 13.76 kcal/mol, respectively). The MFEs of the wild-type and mutant sequences are the same because the local regions −30:30 are identical.

In contrast to MFE, accessibility was able to make an accurate prediction by capturing the full ensemble average energy of the mutant sequence. Importantly, although we use the region −24:24 for the accessibility computation, we are able to capture the key propensity beyond this region. This is because the computation of the partition function, thus the opening energy, for the region −24:24, also utilises the surrounding region (here 210 nucleotides around −24:24, see Methods and Fig 4A). This unique approach makes opening energy more robust than the traditional MFE approach. These results also highlight that mRNA features are interrelated (accessibility and CAI in this case), as such a careful factorial design is necessary to identify the causal features [16, 54].

Accessibility can be improved using a simulated annealing algorithm

The above results suggest that accessibility can, in part, explain the low expression problem of heterologous protein expression. Therefore, we sought to exploit this idea for optimising gene expression. Due to the lack of open source libraries/packages specialised for sequence optimisation, we developed a simulated annealing (Metropolis-Hastings) based algorithm to maximise the accessibility at the region −24:24 using synonymous codon substitution (see Code and data availability, our custom JavaScript and Python modules for the web server and command line tool, respectively). Previous studies have found that full-length synonymous codon-substituted transgenes may produce unexpected results, such as a reduction in mRNA abundance, RNA toxicity, and/or protein misfolding [21, 52, 55, 56]. Therefore, we sought to determine the minimum number of codons required for synonymous substitutions in order to achieve near-optimum accessibility. For this purpose, we used the PSI:Biology targets that failed to be expressed. We applied our simulated annealing algorithm such that synonymous substitutions can happen at any codon of the sequences except the start and stop codons (S7 Fig), although the changes may not necessarily happen to all codons due to the stochastic nature of our optimisation algorithm (see Methods). Next, we constrained synonymous codon substitution to the first 14 codons and applied the same procedure (Fig A in S7 Fig). Therefore, the changes may only occur at any or all of the first 14 codons. We repeated the same procedure for the first nine and also the first four codons. Thus a total of four series of codon-substituted sequences were generated. We then compared the distributions of opening energy −24:24 for these series using the Kolmogorov-Smirnov statistic (D_{KS}; see Fig B in S7 Fig). The distance between the distributions of the nine and full-length codon-substituted series was significantly different yet sufficiently close ($D_{KS} = 0.087$, $P = 3.3 \times 10^{-8}$), suggesting that optimisation of the first nine codons is sufficient in most cases to achieve an optimum accessibility of translation initiation sites. We named our software Translation Initiation coding region designer (TIsigner), which by default, allows synonymous substitutions in the first nine codons.

We asked to what extent the existing gene optimisation tools modify the accessibility of translation initiation sites. For this purpose, we first submitted the PSI:Biology targets that failed to be expressed to the ExpOptimizer web server from NovoPro Bioscience (see

Methods). We also optimised the PSI:Biology targets using the standalone version of Codon Optimisation OnLine (COOL) [26]. We found that both tools increase accessibility indirectly even though their algorithms are not specifically designed to do so. In fact, a purely random synonymous codon substitution on these PSI:Biology targets using our own script resulted in similar increases in accessibility (Fig C in S7 Fig). These results may explain some indirect benefits from the existing gene optimisation tools (i.e. any change from suboptimal is likely to be an improvement, see below).

Low protein yields can be improved by synonymous codon changes in the vicinity of translation initiation sites

To demonstrate that heterologous protein expression is tunable with minimum effort, we designed and tested a series of GFP reporter gene constructs. We tested 29 plasmids harbouring GFP reporter genes with synonymous changes within the first nine codons (opening energies of 5.56–21.68 kcal/mol; Tables A-C in S2 File; and S3 File). GFP expression is controlled by an IPTG (isopropyl-β-D thiogalactopyranoside) inducible T7lac promoter. In addition, all plasmids harbour a second reporter gene (mScarlet-I), which is controlled by the constitutive promoter from the nptII gene for aminoglycoside-3′-O-phosphotransferase of *E. coli* transposon Tn5 [57, 58]. mScarlet-I expression was measured to correct for plasmid copy number and as a proxy for bacterial growth [59].

Consistent with the above results, the GFP level significantly correlates with accessibility (i.e., anti-correlates with opening energy, $R_s = -0.53$, $P = 3.4 \times 10^{-3}$; Fig 5A). This correlation was also the strongest compared to other features, which independently supports our observations on multiple large-scale datasets. Curiously, we observed a diminishing return with opening energies lower than that of the wild-type sequence (11.68 kcal/mol). To investigate this, we simulated a protein production experiment by modelling cell growth, transcription, translation, and turnovers (see Methods). We assumed that opening energy of 12 kcal/mol or below is favourable in this model, based on our analysis of 8,780 PSI:Biology 'success' group (S7 Fig). Interestingly, this stochastic model shows a similar protein production trend as the actual experiment (Fig 5B). Surprisingly, this in silico model also shows that an efficient protein production leads to slower cell growth (Fig 5B). This phenomenon, also known as protein cost, is observed in vivo, in which overexpression slows down cell growth due to the cost-benefit trade-offs of protein circuits [60–66].

Additionally, we tested this finding using the luciferase reporter from *Renilla reniformis* (RLuc). We designed and tested a series of nine RLuc variants with higher accessibility than the wild-type sequence, in which sequence optimisations were performed within the first 9 codons (opening energies of 5.77–10.38 kcal/mol; S8 Fig and S3 File). We also tested five commercially designed sequences, which incorporated sequence optimisations across the entire RLuc gene. The TIsigner optimal sequence (5.77 kcal/mol) was expressed at a higher level compared to a TIsigner suboptimal sequence (10.38 kcal/mol) and the wild type (13.15 kcal/mol) in the BL21Star(DE3) *E. coli* host. However, it is worth noting that a high proportion of the expressed RLuc protein is insoluble, as analysed by SDS-PAGE (Fig A in S8 Fig). We also carried out luciferase reporter assay to compare the levels of active soluble RLuc but detected no significant differences between the luminescence levels of these three sequences. The poor correlation between RLuc protein abundance with luciferase activity could be partly attributed to the observed aggregation problems in Fig A in S8 Fig.

Nevertheless, a TIsigner suboptimal sequence (9.90 kcal/mol) and the commercially optimised sequences did produced approximately 1.5 times higher luminescence than the wild-type (Fig B in S8 Fig), an increase which is still insufficient for purification and recombinant

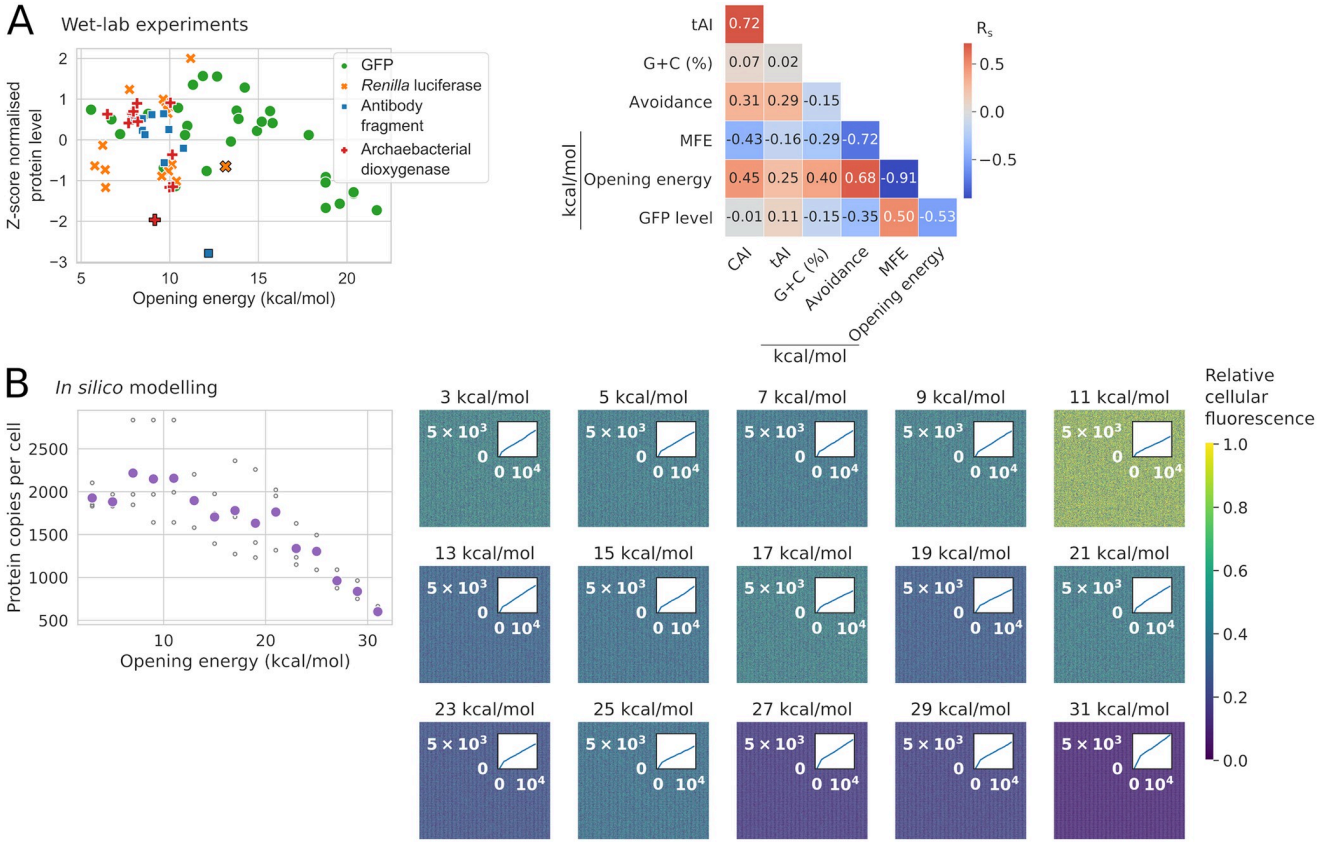

Fig 5. The yields of heterologous protein productions are tunable by synonymous codon changes in the first nine codons. A: GFP level strongly correlates with accessibility, i.e., anti-correlates with opening energy ($R_s = -0.53$, $P = 3.4 \times 10^{-3}$; N = 29). This correlation is the strongest compared to other features (right), which independently supports the observations on multiple large-scale datasets (Figs 2–4). The protein levels of GFP, *Renilla* luciferase (RLuc), an antibody fragment and an archaebacterial dioxygenase were transformed using a z-score method. The GFP and RLuc levels were derived from the average values of at least two and three independent biological replicates, respectively. Black outlines denote wild-type sequences. See also S8 and S9 Figs and S3 File. CAI, codon adaptation index; CC, codon context; R_s, Spearman's rho; tAI, tRNA adaptation index. B: Stochastic simulation of a protein production experiment (e.g., GFP) by modelling cell growth, transcription, translation, and turnovers, given that translation initiation sites with opening energies less than or equal to 12 kcal/mol is optimum. The in silico model shows a similar trend of protein production as the wet-lab experimental results. Unfilled and filled (purple) circles denote the in silico replicates and their corresponding average values, respectively ($R_s = -0.75$, $P = 2.8 \times 10^{-9}$). Similar to an actual recombinant production experiment, the in silico model also shows that efficient protein production (higher relative cellular fluorescence) leads to slower cell growth and vice versa (right, see insets for the opening energies of 11 kcal/mol versus 31 kcal/mol). Insets represent the number of cells with iterations.

https://doi.org/10.1371/journal.pcbi.1009461.g005

protein production standard. It is worth noting that this TIsigner sequence (9.90 kcal/mol) harbours only two nucleotide changes compared to 187 to 241 nucleotide changes for the commercially designer sequences (0.2% versus 20.0%-25.7% of the full-length sequence). Due to the persisting aggregation issues, further testing in RLuc reporter for the full spectrum of opening energies is no longer warranted.

As both wild-type GFP and RLuc proteins were expressed at high levels in *E. coli*, we posited whether poorly expressed proteins can also be improved by increasing accessibility of translation initiation sites. We performed densitometric analysis of previously published Western blots using imageJ [67], which include the results of a cell-free expression system using constructs harbouring a wild-type antibody fragment or archaebacterial dioxygenase and its synonymous variants (within the first six codons) [30]. Indeed, we observed variants with opening energies lower than the wild-type sequences were expressed at higher levels (S9 Fig).

A recent study also showed that an increase in accessibility of a 30 bp region from the Shine-Dalgarno sequence enhances the expression level of human voltage dependent anion channel in an *E. coli* cell-free system, which further supports our findings [68]. Overall, the findings show that optimising accessibility is useful for tuning protein expression in both cellular and cell-free expression systems.

Discussion

Accessibility is a single sequence feature that explains most of the variation in protein abundance

Our data-driven approach shows that the accessibility of translation initiation sites is the strongest predictor of heterologous protein expression in *E. coli*. However, protein expression is inherently noisy due to the interplay of many cellular processes. At the transcript level, many mRNA features are not truly independent (Fig 4, e.g., accessibility, MFE, and G+C content), which aggravates the problem of identifying the key features. As such, a careful design of experiments such as using factorial methods for generating mRNA sequences is crucial for a complete traversal of the feature landscape. Due to the large-scale nature of such factorial designs, to-date few attempts have been made, e.g., 244,000 GFP and 86 firefly luciferase synonymous variants tested in *E. coli* and HeLa cells, respectively [16, 54]. These fluorescence reporter studies concluded that MFE was the best predictor but with modest correlations (e.g., Fig 2, Spearman's correlations of 0.51 in E. coli). These modest correlations reflect the noisiness of the system which further poses a problem for obtaining a better predictor. Furthermore, MFE estimation involves identifying the thermodynamically most probable structure from Boltzmann's ensemble, which is often inaccurate in a biological system where different constraints may prevent a mRNA from attaining the most probable conformation.

With this in mind, we used opening energy, an accessibility-based approach that takes the full ensemble average energy into account. This includes all possible RNA structures, including suboptimal structures that are not reported by MFE models by default [22, 69]. Indeed, our approach gave us a better correlation from multiple datasets where MFE was previously concluded to be a better predictor. We have shown that accessibility is superior to MFE even for the datasets without factorial designs such as the PSI:Biology dataset (Fig 4), where the feature space is sampled irregularly, and the expression levels of recombinant proteins were categorised into 'Tested_Not_Found' and 'Protein_Confirmed' with SDS-PAGE analysis [42, 44] (S2 Fig, 11,430 proteins from over 189 diverse species).

Moreover, the correlation between endogenous mRNA and protein levels is limited in both bacteria and eukaryotes (0.4–0.7) [10, 70–75], where theoretically mRNA levels should provide an upper-bound on correlation statistics of mRNA features. Besides mRNA level, accessibility is a sequence feature that explains most of the variation in protein abundance (Fig 2, R_s of 0.28–0.65). Any further improvements in correlations are likely to be hindered by the noise and encountered diminishing returns.

Adoption of accessibility for tuning protein expression

The accessibility of a region of mRNA can be understood as the ability of that region to base-pair with other nucleotides, including the flanking region. There are two distinct ways to define the accessibility of a region. (i) The first way is to consider the minimum Gibbs free energy (MFE) of the region. A lower value of MFE implies a stronger folding at that region. Hence, the region is less likely to be available for pairing. The region is then said to be less accessible. Thus, several authors have defined accessibility using the Gibbs free energy, which

reflects the strength of mRNA folding around the region of interest [14, 24, 28, 29, 31, 32]. Since, the MFE can be calculated through the computation of a partition function for base pairing, a more rigorous way to define accessibility is to use the partition function to compute the basepair probabilities [30, 76]. Despite the improvements over MFE, this partition function-based approach has not been widely adopted. (ii) Another way that accessibility has been defined is to use the partition function to compute the probabilities of bases being unpaired, or an equivalent pseudo energy called the opening energy [22]. In this work, we use this approach to define accessibility as it is mathematically well-defined and provides greater contextual information (See Figs 1 and 4A). Furthermore, an efficient algorithm to compute opening energy is implemented in the RNAplfold of the ViennaRNA suite [77]. A short comparison of these methods is given in Additional notes in S2 File.

Terai and Asai (2020) and ourselves have independently discovered that modelling the accessibility of translation initiation sites using the base-unpairing approach is superior to the simplistic MFE estimation [33, 78] (see Fig 2A for example). A key difference between our approaches is that we used RNAplfold which is based on a biophysical model of RNA with Turner's nearest-neighbor parameters [79], whereas Terai and Asai (2020) used Raccess, which is based on a probabilistic model and uses a further optimised set of Turner's parameters [80, 81].

Besides, very few applications of accessibility have been developed, for example, as implemented in RNAup and IntaRNA for the prediction of RNA-RNA intermolecular interactions [69, 77, 82]. We have advanced our findings by developing a web service and a command line tool for tuning protein expression called TIsigner. The underlying JavaScript and Python modules that we developed for the web service and command line tool, respectively, are open source and freely available (see Code and data availability).

Our method requires only a modest number of synonymous codon changes to the 5′ coding region of a given sequence (Fig 5). In contrast, other gene optimisers incorporate several features that require synonymous codon changes of almost the entire sequence.

Implementations of TIsigner for improving recombinant protein production

Our TIsigner web service offers several unique features and supports recombinant protein expression in *E. coli*, *S. cerevisiae*, and *M. musculus* (optimisation regions −24:24, −7:89 and −-8:11, respectively; see Fig 2). Users can easily change the optimisation region to accommodate other expression hosts. Our TIsigner web service also allows full-length sequence optimisation in addition to the first nine codons. For *E. coli* hosts, users are warned when terminators or any custom sequence motifs are detected.

Furthermore, we provide a holistic solution to design experiments for recombinant protein production [83]. Importantly, TIsigner is integrated with SoDoPE and Razor web services that allow solubility optimisation and signal peptide prediction, respectively [84, 85]. Such integration allows a seamless transition between these three services. For example, a protein sequence of interest can be first submitted to Razor. If a signal peptide is detected, users have an option to check for solubility using SoDoPE and select the mature region using the interactive interface. Otherwise, users can also check the solubility of the full-length sequence. If protein domains are detected, users can consider optimising solubility by selecting any subregions. Regions with optimised solubility are instantly returned. Users can then redirect any selected region to TIsigner for accessibility optimisation. In contrast to the existing gene optimisers, their features are very limited [68, 86–88].

Concluding remarks

The strengths of our approaches are five-fold. Firstly, the likelihood of success or failure can be assessed prior to running an experiment. Users can compare the opening energies calculated for the input and optimised sequences and the distributions of the 'success' and 'failure' of the PSI:Biology targets. We also introduced a scoring scheme in TIsigner to score the input and optimised sequences based upon how likely they are to be expressed (S5 Fig; see also Methods). Secondly, optimised sequences can have up to the first nine codons substituted (by default), meaning that gene optimisation can be done using PCR. For cloning, we propose a nested PCR approach, in which the final PCR reaction utilises a forward primer designed according to the optimised sequence [89] (Fig D in S7 Fig). Thirdly, the cost of gene optimisation can be reduced dramatically as gene synthesis is replaced with PCR using our approach. This enables high-throughput protein expression screening using the optimised sequences, generated at a low cost. Fourthly, tunable expression is possible, i.e. high, intermediate or even low expression 5′ codon sequences can be designed, allowing for more control over heterologous protein production, as demonstrated by our experiments (Fig 5). Finally, our fast, lightweight, stochastic simulation approach has opened up new avenues to study several aspects of gene expression, such as transcription, translation, cellular growth, and turnovers, which give good proxies to how cellular systems behave.

Methods

Plasmids

Plasmids were constructed using the MIDAS Golden Gate cloning system [90] (see Additional Methods, Figs A-E and Tables A-C in S2 File).

Data

Datasets used in this study are listed in S1 File. These include fluorescence reporter expression datasets previously generated using *E. coli*, *Saccharomyces cerevisiae*, and *Mus musculus* cultured cells (S1 Fig), and recombinant protein production dataset from the Protein Structure Initiative: Biology (PSI:Biology; S2 Fig). Representative sequences were chosen from the *E. coli* green fluorescence protein (GFP) reporter dataset [16] using CD-HIT-EST [91, 92]. Two ribosome profiling libraries previously generated using *E. coli* were retrieved from the Sequence Read Archive (SRR7759806 and SRR7759807) [93].

Sequence features analysis

CAI, tAI and Codon Context (CC) were calculated using the reference weights from Sharp and Li [11], Tuller et al. [50] and Ang et al. [51], respectively. Translation elongation rate was predicted using Iχnos [52] that was trained using an *E. coli* ribosome profiling dataset (S3 Fig). Local G+C content was also examined (S4 Fig).

We use the minimum free energy (MFE), opening energy and avoidance as thermodynamic features of a mRNA. The MFE of a sequence is the energetically most optimal sequence in a Boltzmann's ensemble. Consequently this feature is based on a single structure. In contrast, the opening energy of a stretch *i..j* of nucleotides is the pseudo energy required to unpair that region considering all sub-optimal structures and is given by:

$$Opening\ \ energy = -kT\ \ ln\frac{Z_{unpaired}}{Z}, \tag{1}$$

where the where k is the Boltzmann's constant and T is the absolute temperature, $Z_{unpaired}$ is

the canonical partition function of the region *i..j* and *Z* is the total canonical partition function. The ratio $Z_{unpaired}/Z$ is the probability that the nucleotides *i..j* are unpaired, note that this ratio incorporates base-unpairing probabilities for the larger region $(i - W)..(j + W)$ (Fig 1), where *W* is the window size used to compute partition functions. See also the expanded equation in Supporting Information (S2 File).

A crucial difference between opening energy and the widely used MFE [14–16] is that $Z_{unpaired}$ captures all possible optimal and suboptimal structures in the Boltzmann's ensemble, where the nucleotides *i..j* are unpaired, including possible pairings from the surrounding region (*W*). For a sufficiently large window (*W*), this approach captures the bulk of RNA contacts, even for very large RNAs [94]. In other words, all possible optimal or suboptimal structures beyond the target region are accounted for, resulting in a full ensemble average energy for a region (see a case study in the Results section, Accessibility captures the full ensemble average energy of a sequence). This is distinct from the traditional MFE approach that returns a single solution (optimal or near-optimal structure) for a given region.

The avoidance metric measures the number of potential intermolecular misinteractions, and is calculated by computing the possible interactions among RNAs. The conventional way to model interactions is to assume a two step process where nucleotides unpair and then hybridise. Thus the total interaction energy ΔG_{int} is given by:

$$\Delta G_{int} = \Delta G_u + \Delta G_h, \tag{2}$$

where ΔG_u is the opening energy and ΔG_h is the hybridisation energy. In the previous study, ΔG_u was inadvertently upweighted. In this study we correctly address this by using only the ΔG_h component, where $\Delta G_h < \Delta G_u$.

The computation of the partition function and related energetic quantities involve dynamic programming. The implementation of these dynamic programming algorithms is available on several tools such as ViennaRNA and IntaRNA [77, 82].

We used RNAfold, RNAplfold and RNAup from the ViennaRNA package (version 2.4.11) to calculate MFE, opening energy and avoidance, respectively [22, 69, 77, 95–98]. RNAfold was run with default parameters. Based on previous studies, we calculated MFE using the mRNA region -30:30 [16, 37–40]. For RNAplfold, sub-sequences were generated from the input sequences to calculate opening energies (using the parameters -W 210 -u 210), in practise there is a subtle difference between the u and W parameters, but for this work we set these to be equal (https://www.tbi.univie.ac.at/RNA/RNAplfold.1.html) [22]. For RNAup, we examined the stochastic interactions between the region 1:30 of each mRNA and 54 non-coding RNAs (using the parameters -b -o). RNAup reports the total interaction between two RNAs as the sum of energy required to open accessible sites in the interacting molecules ΔG_u and the energy gained by subsequent hybridisation ΔG_h [69]. For the interactions between each mRNA and 54 non-coding RNAs, we chose the most stable mRNA:ncRNA pair to report an inappropriate mRNA:ncRNA interaction, i.e. the pair with the strongest hybridisation energy, $(\Delta G_h)_{min}$.

Simulation

To better understand the dynamics between accessibility and protein production, we performed a stochastic simulation using constructs with increasing opening energy on a simulated cellular system.

To set the simulation, we binned the opening energies between 2 and 32 kcal/mol in intervals of two, with each bin representing a 'reporter plasmid construct' whose opening energy is the mean of the bin. For each construct, 'technical replicates' were generated by allowing slight

variations on the mean opening energy of the bin. This is to model variation between replicates, and the discrepancies between the estimated and the actual opening energies in vivo. For each round of transcription, mRNA copies were randomly generated from 30 to 60 plasmid DNA copies [3, 99, 100]. Based upon our analysis of targets from PSI:Biology (S7 Fig), we chose an optimum opening energy of 12 kcal/mol or less for translation. However, this is probabilistic which occasionally allows protein production from higher opening energy transcripts. We allowed mRNA to decay probabilistically when a mRNA molecule is translated for more than 10 times.

To simulate protein toxicity, we set a threshold of protein to be 1,000,000 copies where the copy number of endogenous proteins is usually less than 10,000 [10]. Beyond this limit, a sporadic death of cells is simulated. However, in this simulation, the chance of staying viable and reproducing is higher than death, and cells grow steadily. This threshold also simulated random but low cell deaths in the experiment, without setting an extra variable.

To limit the computational complexity of the simulation, we use smaller constants and iterations. Initialising with 100 cells, the algorithm was set to terminate either after 10,000 iterations or when the total number of cells is zero. After termination, the total number of proteins and cells for each construct were taken from the endpoints. To imitate 'biological replicates', we repeated the above simulation three times with different random numbers, which provides slightly different initial conditions for each experiment.

Development of Translation Initiation coding region designer (TIsigner)

Finding a synonymous sequence with a maximum accessibility is a combinatorial problem that spans a vast search space. For example, for a protein-coding sequence of nine codons, assuming an average of 3 synonymous codons per amino acid, we can expect a total of 19,682 unique synonymous coding sequences. This number increases rapidly with increasing numbers of codons. Heuristic optimisation approaches are preferred in such situations because the search space can be explored more efficiently to obtain nearly optimal solutions.

To optimise the accessibility of a given sequence, TIsigner uses a simulated annealing algorithm [101–104], a heuristic optimisation technique based on the thermodynamics of a system settling into a low energy state after cooling. Simulated annealing algorithms have been used to solve many combinatorial optimisation problems in bioinformatics. For example, we previously applied this algorithm to align and predict non-coding RNAs from multiple sequences [105]. Other studies use this algorithm to find consensus sequences [103], optimise ribosome binding sites [24] and predict mRNA foldings [106] using MFE models.

According to statistical mechanics, the probability p_i of a system occupying energy state E_i, with temperature T, follows a Boltzmann distribution of the form $e^{E_i/T}$, which gives a set of probability mass functions along every point i in the solution space. Using a Markov chain sampling, these probabilities are sampled such that each point has a lower temperature than the previous one. As the system is cooled from high to low temperatures ($T \rightarrow 0$), the samples converge to a minimum of E, which in many cases will be the global minimum [103]. A frequently used Markov chain sampling technique is Metropolis-Hastings algorithm in which a 'bad' move E_2 from initial state E_1 such that $E_2 > E_1$, is accepted if $R(0, 1) \geq p_2/p_1$, where $R(0, 1)$ is a uniformly random number between 0 and 1.

In our implementation, each iteration consists of a move that may involve multiple synonymous codon substitutions. The algorithm begins at a high temperature where the first move is drastic, synonymous substitutions occur in all replaceable codons. At the end of the first iteration, a new sequence is accepted if the opening energy is smaller than that of the input sequence. However, if the opening energy of a new sequence is greater than that of the input

sequence, acceptance depends on the Metropolis-Hastings criteria. The accepted sequence is used for the next iteration, which repeats the above process. As the temperature cools (exponentially decreasing), the moves get milder with fewer synonymous codon changes (S7 Fig). Simulated annealing stops upon reaching a near-optimum solution.

For the web version of TIsigner, the default number of replaceable codons is restricted to the first nine codons. However, this default setting can be reset to range from the first four to nine codons, or the full length of the coding sequence. Since the accessibility of a fixed region is optimised, this process only takes $\mathcal{O}(1)$ time (S10 Fig). Furthermore, TIsigner runs multiple simulated annealing instances, in parallel, to obtain multiple possible sequence solutions.

When users select T7lac promoter as the 5′UTR, they can adjust 'Expression Score', that is calculated based on the PSI:Biology dataset (see below). This allows them to tune the expression level of a target gene. In contrast, when users input a custom 5′UTR sequence, they only have the option to either maximise or minimise expression.

To implement 'Expression Score', the posterior probabilities of success for input and optimised sequences are evaluated using the following equations from Bayesian statistics:

$$positive\ \ posterior\ \ odds = prior\ \ odds \times fitted\ \ positive\ \ likelihood\ \ ratio, \tag{3}$$

$$positive\ \ posterior\ \ probability = \frac{positive\ \ posterior\ \ odds}{1 + positiveposteriorodds}, \tag{4}$$

The fitted positive likelihood ratios in Eq (3) were obtained from the following 4-parametric logistic regression equation:

$$fitted\ \ positive\ \ likelihood\ \ ratio = d + \frac{a - d}{1 + \left(\frac{positive\ \ likelihood\ \ ratio}{c}\right)^{b}}, \tag{5}$$

with parameters a, b, c, and d. The prior probability was set to 0.49, which is the proportion of 'Expressed' (N = 21,046) divided by 'Cloned' (N = 42,774) of the PSI:Biology targets reported as of 28 June 2017 (http://targetdb.rcsb.org/metrics/). Posterior probabilities were scaled as percentages to score the input and optimised sequences (S5 Fig).

The presence of terminator-like elements [107] in the protein-coding region may result in expression of truncated mRNAs due to early transcription termination. Therefore, we implemented an optional check for putative terminators in the input and optimised sequences by cmsearch (INFERNAL version 1.1.2) [108] using the covariance models of terminators from RMfam [109, 110]. We also allow users to filter the output sequences for the presence of restriction sites. Restriction modification sites (AarI, BsaI, and BsmBI) are avoided by default.

Besides *E. coli*, users can choose *S. cerevisiae*, *M. musculus* or 'Other' as the expression host. The regions for optimising accessibility are −7:89, −8:11 and −24:89 for *S. cerevisiae*, *M. musculus* and 'Other', respectively (Fig 2 and S1 Fig). When users choose 'Custom' for expression host, the region for optimising accessibility becomes customisable.

Sequence optimisation

To compare accessibility between sequences optimised using TIsigner and other tools, we submitted the PSI:Biology targets that failed to be expressed (N = 2,650) to the ExpOptimizer web server from NovoPro Bioscience (https://www.novoprolabs.com/tools/codon-optimization). A total of 2,573 sequences were optimised. The target sequences were also optimised using a local version of COOL [26] and TIsigner using default settings. We also ran a random synonymous codon substitution as a control for these 2,573 sequences.

GFP assay

BL21(DE3)pLysS competent *E. coli* cells (Invitrogen) were transformed with plasmids and grown overnight on Luria-Bertani (LB) agar plates containing spectinomycin (50 μg/ml) and chloramphenicol (25 μg/ml). Single colonies were picked and inoculated into 3 ml LB broth containing the same antibiotics, and cultures were grown for 18 hours at 37˚*C*, 200 rpm. Cultures were diluted with fresh media at 1:20 and grown at 37˚*C*, 200 rpm, until reaching the mid-logarithmic growth phase (optical densities at 600 nm (OD600) of 0.3). Of each culture, 20 μl was seeded into 96-well plates containing 180 μl LB broth supplemented with antibiotics and isopropyl-β-D thiogalactopyranoside (IPTG) (1 mM final concentration) per well. Fluorescence intensities and ODs were measured in a black, flat, clear bottom 96-well plate with lid (CELLSTAR, Greiner) using a FLUOstar Omega plate reader (BMG Labtech) equipped with an excitation filter (band pass 485–12) and an emission filter (band pass 520) for GFP and excitation filter (band pass 484) and an emission filter (band pass 610–10) for mScarlet-I. The plate was incubated at 37˚*C* with "meander corner well shaking" at 300 rpm for 7 hours measuring fluorescence and ODs every 10 minutes. Fluorescence was measured in a 2 mm circle recording the average of 8 measurements per well. Average values of technical replicates were calculated and normalised to the mScarlet-I second reporter, and then to the normalised value of the GFP variant with the highest opening energy (21.68 kcal/mol). Normalised fluorescence values were obtained from the average values of biological replicates (S8 Fig and S3 File).

Luciferase assay

BL21Star(DE3) competent cells (Invitrogen) were transformed with plasmids and grown overnight at 37˚C on LB agar plates containing 50 μg/ml spectinomycin. Single colonies were picked and inoculated into 5 ml LB broth (50 μg/ml spectinomycin) and grown for 18 hours at 37˚*C*, 200 rpm. Bacterial cultures were diluted with fresh media at 1:20 and grown at 37˚*C*, 200 rpm, up to a mid-logarithmic phase (OD600 of 0.4). The cultures were split and induced with IPTG at a final concentration of 0.25 mM (or uninduced as controls), and seeded into a white, flat, clear bottom 96-well white plate with lid (Costar, Corning), 150 μl per well, in triplicates. Cells were incubated in a FLUOstar Omega Microplate Reader (BMG LABTECH) for 90 minutes at 25˚*C*, 200 rpm, and OD600 was measured every 15 minutes (over 7 cycles). Cells were harvested by centrifugation at 3000 ×g, for 10 minutes, at 20˚*C*. Supernatants were removed. As the substrate can penetrate into cells, 50 μl of coelenterazine h (Promega) was added to living cells to minimise sample processing steps and variability [111, 112]. Luminescence was measured (λ_{em} = 475 nm) in a Clariostar microplate reader (BMG LABTECH) at 25˚*C* every 2 minutes (over 11 cycles). Average values of technical replicates were calculated and normalised to the wild-type. Normalised luminescence values were obtained from the average values of biological replicates (S9 Fig and S3 File).

Statistical analysis

AUC and Gini importance scores were calculated using scikit-learn (version 0.20.2) [113]. The 95% confidence intervals for AUC scores were calculated using DeLong's method [114]. Spearman's correlation coefficients and Kolmogorov-Smirnov statistics were calculated using Pandas (version 0.23.4) [115] and scipy (version 1.2.1) [116, 117], respectively. Positive likelihood ratios with 95% confidence intervals were calculated using the bootLR package [118, 119]. The P-values of multiple testing were adjusted using Bonferroni's correction and reported to machine precision. Plots were generated using Matplotlib (version 3.0.2) [120] and Seaborn (version 0.9.0) [121].

Supporting information

S1 Fig. Heatmaps of correlations between opening energies and protein abundances for each of the sub-sequence regions (related to Fig 2). Green unfilled triangles indicate the regions before and after scaling (left and right panels, respectively). A: For *E. coli*, we used a representative GFP expression dataset from Cambray et al. (2018) [16]. The reporter library consists of GFP fused in-frame with a library of 96-nt upstream sequences (N = 14,425). B: For *S. cerevisiae*, we used a YFP expression dataset from Dvir et al. (2013) [19]. The YFP reporter library consists of 2,041 random decameric nucleotides inserted at the upstream of YFP start codon. C: For *M. musculus*, we used the GFP expression dataset from Noderer et al. (2014) [34]. The GFP reporter library consists of 65,536 random hexameric and dimeric nucleotides inserted at the upstream and downstream of GFP start codon, respectively. R_s, Spearman's rho.
(PDF)

S2 Fig. Expression outcomes of the PSI:Biology targets in *E. coli* (related to Figs 3C and 4). A total of 11,430 PSI:Biology targets from over 189 species were analysed in this study (N = 8,780 and 2,650, 'success' and 'failure' groups, respectively). Genera with at least 20 target genes are shown and the remaining as 'Others'. The top three PSI:Biology targets are from four *Pseudomonas*, five *Bacillus* and six *Clostridium* species. Red asterisk, obelisk and diesis indicate *Homo sapiens*, *S. cerevisiae* and *E. coli*, respectively. These target genes were inserted into the pET21_NESG expression vector, in which the promoter and fusion tag are T7lac and C-terminal His tag, respectively.
(PDF)

S3 Fig. Ribosome footprints in 25-nt fragments show a strong triplet periodicity, indicating translation (related to Fig 4). These 25-nt footprints (green unfilled rectangle) were used to train a neural network model [52] in order to predict the translation elongation rates of the PSI:Biology targets. Ribosome profiling data (SRR7759806 and SRR7759807 [93]) were first aligned to *S. cerevisiae* transcriptome. SAM alignment files were merged, and ribosome footprints which were mapped to each frame were enumerated. See https://github.com/Gardner-BinfLab/TIsigner_paper_2019. FP, footprints.
(PDF)

S4 Fig. Analysis of the local G+C contents in the PSI:Biology target genes (related to Fig 4). A: The G+C contents in the regions −24:24 and −30:30 weakly correlate with opening energy and MFE, respectively. Green unfilled squares indicate Spearman's correlations (R_s) between the local G+C contents and the corresponding local features. B: The local G+C contents show a similar prediction accuracy (AUC scores shown in parentheses). AUC, Area Under the receiver operating characteristic Curve; MFE, Minimum Free Energy.
(PDF)

S5 Fig. Opening energy of 10 kcal/mol or below at the region −24:24 is about two times more likely to come from the target genes that are successfully expressed than those that failed. Cumulative frequency distributions of the true positive and false positive (less than type), and true negative and false negative (more than type) derived from the ROC analysis in Fig 4C (left panel, opening energy −24:24). These values were used to estimate positive likelihood ratios with 95% confidence intervals using 10,000 bootstrap replicates. The estimated ratios and/or confidence intervals are inaccurate at low numbers of true positives or true negatives. Therefore, a four-parameter logistic curve was fitted to the positive likelihood ratios. Fitted values are useful to estimate the posterior probability of protein expression.
(PDF)

S6 Fig. Accessibility is able to capture the full ensemble average energy of a sequence. The levels of major capsid protein expressed by the wild-type (orange) and mutant (blue) strains of bacteriophage T7 [53]. The mutant major capsid gene was codon-deoptimised such that the first and the last 14 codons remained unchanged. Proteomic analyses were carried out at 1, 5 and 9 min post-infection in four biological replicates. Opening energy −24:24, MFE −30:30, and CAI of the wild-type and mutant sequences were compared. The approximated 'Expression Scores' of the wild-type and mutant sequences are 89 and 38, respectively (opening energies of 9.05 kcal/mol and 13.76 kcal/mol, respectively). MFE, Minimum Free Energy; CAI, Codon Adaptation Index.
(PDF)

S7 Fig. Accessibility of translation initiation sites can be increased by synonymous codon substitution within the first nine codons using simulated annealing. A: Schedules in simulated annealing. The ratio of temperature to the number of the first N substitutable codons decreases exponentially with increasing number of iterations. B: Accessibility of translation initiation sites increases with increasing number of the first N replaceable codons. The PSI:Biology targets that failed to be expressed were optimised using simulated annealing (N = 2,650). The Kolmogorov-Smirnov distance between the distributions of '9' and 'full-length' was significantly different but sufficiently close ($D_{KS} = 0.09$, $P < 10^{-7}$), indicating that optimisation of the first nine codons can achieve nearly optimum accessibility. For comparison, the distribution of the PSI:Biology targets that were successfully expressed are shown (N = 8,780). See also S1 File. C: Accessibility of translation initiation sites can be increased indirectly using the existing gene optimisation tools and random synonymous codon substitution. 'TIsigner (9)' refers to the default settings of our tool, which allows synonymous substitutions up to the first nine codons (as above). D: Accessibility of translation initiation sites can be optimised using PCR. The forward primer should be designed according to TIsiger optimised sequences. For example, using a nested PCR approach, the optimised sequence can be produced using the forward primer designed with appropriate mismatches (gold bulges) to amplify the amplicon from the initial PCR reaction.
(PDF)

S8 Fig. Luciferase reporter assay (related to Fig 5A). A: SDS-PAGE gel shows the protein bands of *Renilla* luciferase (RLuc) in the soluble and insoluble fractions of BL21Star(DE3) lysates. The expression of RLuc can be improved, despite its poor solubility in *E. coli*. Selected bacterial clones were grown at 25˚*C*, 200 RPM. The solubilities of wildtype (WT) RLuc and designed variants were compared after 4-hour IPTG induction. The blue and red arrows (about 36kDa) indicate that RLuc was poorly soluble. No RLuc protein bands were detected from the uninduced cultures and IPTG-induced negative control (empty vector control that lacks Rluc gene and T7lac promoter). B: The luciferase activities of commercially designed RLuc reporter genes (full-length sequence optimisation) and a TIsigner optimised sequence (9.9 kcal/mol) are significantly higher than the wild-type luciferase (Mann-Whitney U tests, $P = 9.1 \times 10^{-3}$). Opening energies are shown next to labels. IPTG, isopropyl-β-D thiogalactopyranoside.
(PDF)

S9 Fig. The yields of an antibody fragment and an archaebacterial dioxygenase can be improved by synonymous codon changes within the first six codons (related to Fig 5A). A RTS *E. coli* cell-free expression system was previously used to express these recombinant proteins [30]. The expression levels are shown in arbitrary units (AU) based on the densitometric analysis of previously published Western blots (S3 File). WT, wild-type.
(PDF)

S10 Fig. Sequence length does not affect software performance because only a fixed region is taken into account during optimisation ($\mathcal{O}(1)$ time).
(PDF)

S1 File. Datasets used in this study and results. Results for Fig 4D, and Figs B and C in S7 Fig.
(XLSX)

S2 File. Additional notes and methods. Figs A-E, and Tables A-C.
(PDF)

S3 File. Experimental results. GFP and RLuc reporter assay results and densitometric data for Fig 1A, Fig B in S8 and S9 Figs.
(XLSX)

Acknowledgments

We thank Professor Ivo Hofacker for fruitful discussions at the Benasque RNA Meeting, and Dr Ronny Lorenz for helpful discussions about RNAplfold. We are grateful to the members of the Biomolecular Interaction Centre at the University of Canterbury for supporting this research. We thank New Zealand eScience Infrastructure for providing high performance computing resources.

Author Contributions

Conceptualization: Bikash K. Bhandari, Chun Shen Lim, Craig van Dolleweerd, Paul P. Gardner.

Data curation: Chun Shen Lim.

Formal analysis: Bikash K. Bhandari, Chun Shen Lim, Daniela M. Remus, Augustine Chen, Craig van Dolleweerd.

Funding acquisition: Craig van Dolleweerd, Paul P. Gardner.

Investigation: Bikash K. Bhandari, Chun Shen Lim, Daniela M. Remus, Augustine Chen, Craig van Dolleweerd, Paul P. Gardner.

Methodology: Bikash K. Bhandari, Chun Shen Lim, Daniela M. Remus, Augustine Chen, Craig van Dolleweerd, Paul P. Gardner.

Project administration: Chun Shen Lim, Daniela M. Remus, Augustine Chen, Craig van Dolleweerd, Paul P. Gardner.

Resources: Craig van Dolleweerd, Paul P. Gardner.

Software: Bikash K. Bhandari, Chun Shen Lim.

Supervision: Chun Shen Lim, Craig van Dolleweerd, Paul P. Gardner.

Validation: Bikash K. Bhandari, Chun Shen Lim, Daniela M. Remus, Augustine Chen.

Visualization: Bikash K. Bhandari, Chun Shen Lim, Daniela M. Remus, Augustine Chen.

Writing – original draft: Bikash K. Bhandari, Chun Shen Lim, Daniela M. Remus, Augustine Chen, Craig van Dolleweerd, Paul P. Gardner.

Writing – review & editing: Bikash K. Bhandari, Chun Shen Lim, Daniela M. Remus, Augustine Chen, Craig van Dolleweerd, Paul P. Gardner.

References

1. Kimelman A, Levy A, Sberro H, Kidron S, Leavitt A, Amitai G, et al. A vast collection of microbial genes that are toxic to bacteria. Genome Res. 2012; 22(4):802–809. https://doi.org/10.1101/gr.133850.111 PMID: 22300632
2. Berlec A, Strukelj B. Current state and recent advances in biopharmaceutical production in *Escherichia coli*, yeasts and mammalian cells. J Ind Microbiol Biotechnol. 2013; 40(3-4):257–274. https://doi.org/10.1007/s10295-013-1235-0 PMID: 23385853
3. Rosano GL, Ceccarelli EA. Recombinant protein expression in *Escherichia coli*: advances and challenges. Front Microbiol. 2014; 5:172. https://doi.org/10.3389/fmicb.2014.00172 PMID: 24860555
4. Abreu RdS, de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Molecular BioSystems. 2009; 5(12):1512–26. https://doi.org/10.1039/b908315d PMID: 20023718
5. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol. 2018; 19(1):20–30. https://doi.org/10.1038/nrm.2017.91 PMID: 29018283
6. Lim CS, Wardell SJT, Kleffmann T, Brown CM. The exon–intron gene structure upstream of the initiation codon predicts translation efficiency. Nucleic Acids Res. 2018; 46(9):4575–4591. https://doi.org/10.1093/nar/gky282 PMID: 29684192
7. Stevens SG, Brown CM. In silico estimation of translation efficiency in human cell lines: potential evidence for widespread translational control. PLoS One. 2013; 8(2):e57625. https://doi.org/10.1371/journal.pone.0057625 PMID: 23460887
8. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473(7347):337–342. https://doi.org/10.1038/nature10098 PMID: 21593866
9. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc Natl Acad Sci U S A. 2002; 99(15):9697–9702. https://doi.org/10.1073/pnas.112318199 PMID: 12119387
10. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010; 329(5991):533–538. https://doi.org/10.1126/science.1188308 PMID: 20671182
11. 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):1281–1295. https://doi.org/10.1093/nar/15.3.1281 PMID: 3547335
12. Reis Md, d Reis M. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004; 32(17):5036–5044. https://doi.org/10.1093/nar/gkh834 PMID: 15448185
13. Sabi R, Tuller T. Modelling the Efficiency of Codon–tRNA Interactions Based on Codon Usage Bias. DNA Res. 2014; 21(5):511–526. https://doi.org/10.1093/dnares/dsu017 PMID: 24906480
14. Pelletier J, Sonenberg N. The involvement of mRNA secondary structure in protein synthesis. Biochem Cell Biol. 1987; 65(6):576–581. https://doi.org/10.1139/o87-074 PMID: 3322328
15. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. Science. 2009; 324(5924):255–258. https://doi.org/10.1126/science.1170160 PMID: 19359587
16. Cambray G, Guimaraes JC, Arkin AP. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. Nat Biotechnol. 2018; 36(10):1005–1015. https://doi.org/10.1038/nbt.4238 PMID: 30247489
17. de Smit MH, van Duin J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc Natl Acad Sci U S A. 1990; 87(19):7668–7672. https://doi.org/10.1073/pnas.87.19.7668 PMID: 2217199
18. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011; 12(1):32–42. https://doi.org/10.1038/nrg2899 PMID: 21102527
19. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, et al. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci U S A. 2013; 110(30):E2792–801. https://doi.org/10.1073/pnas.1222534110 PMID: 23832786
20. Tuller T, Zur H. Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. 2015; 43(1):13–28. https://doi.org/10.1093/nar/gku1313 PMID: 25505165
21. Umu SU, Poole AM, Dobson RC, Gardner PP. Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. Elife. 2016; 5:e13479. https://doi.org/10.7554/eLife.13479 PMID: 27642845

22. Bernhart SH, Mückstein U, Hofacker IL. RNA Accessibility in cubic time. Algorithms Mol Biol. 2011; 6 (1):3. https://doi.org/10.1186/1748-7188-6-3 PMID: 21388531

23. Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinformatics. 2006; 7:285. https://doi.org/10.1186/1471-2105-7-285 PMID: 16756672

24. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol. 2009; 27(10):946–950. https://doi.org/10.1038/nbt.1568 PMID: 19801975

25. Raab D, Graf M, Notka F, Schödl T, Wagner R. The GeneOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization. Syst Synth Biol. 2010; 4(3):215–225. https://doi.org/10.1007/s11693-010-9062-3 PMID: 21189842

26. Chung BKS, Lee DY. Computational codon optimization of synthetic gene for protein expression. BMC Syst Biol. 2012; 6:134. https://doi.org/10.1186/1752-0509-6-134 PMID: 23083100

27. Terai G, Kamegai S, Asai K. CDSfold: an algorithm for designing a protein-coding sequence with the most stable secondary structure. Bioinformatics. 2016; 32(6):828–834. https://doi.org/10.1093/bioinformatics/btv678 PMID: 26589279

28. Bhattacharyya S, Jacobs WM, Adkar BV, Yan J, Zhang W, Shakhnovich EI. Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in *E. coli*. Mol Cell. 2018; 70(5):894–905.e5. https://doi.org/10.1016/j.molcel.2018.05.008 PMID: 29883608

29. Nieuwkoop T, Claassens NJ, van der Oost J. Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design. Microb Biotechnol. 2019; 12(1):173–179. https://doi.org/10.1111/1751-7915.13332 PMID: 30484964

30. Voges D, Watzele M, Nemetz C, Wizemann S, Buchberger B. Analyzing and enhancing mRNA translational efficiency in an *Escherichia coli* in vitro expression system. Biochem Biophys Res Commun. 2004; 318(2):601–614. https://doi.org/10.1016/j.bbrc.2004.04.064 PMID: 15120642

31. Scherr M, Rossi JJ, Sczakiel G, Patzel V. RNA accessibility prediction: a theoretical approach is consistent with experimental studies in cell extracts. Nucleic Acids Res. 2000; 28(13):2455–2461. https://doi.org/10.1093/nar/28.13.2455 PMID: 10871393

32. Espah Borujeni A, Channarasappa AS, Salis HM. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. Nucleic Acids Res. 2014; 42(4):2646–2659. https://doi.org/10.1093/nar/gkt1139 PMID: 24234441

33. Terai G, Asai K. Improving the prediction accuracy of protein abundance in *Escherichia coli* using mRNA accessibility. Nucleic Acids Res. 2020; 48(14):e81–e81. https://doi.org/10.1093/nar/gkaa481 PMID: 32504488

34. Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, et al. Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol Syst Biol. 2014; 10:748. https://doi.org/10.15252/msb.20145136 PMID: 25170020

35. Shine J, Dalgarno L. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci U S A. 1974; 71(4):1342–1346. https://doi.org/10.1073/pnas.71.4.1342 PMID: 4598299

36. Hinnebusch AG. Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. Trends Biochem Sci. 2017; 42(8):589–611. https://doi.org/10.1016/j.tibs.2017.03.004 PMID: 28442192

37. Del Campo C, Bartholomäus A, Fedyunin I, Ignatova Z. Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. PLoS Genet. 2015; 11(10): e1005613. https://doi.org/10.1371/journal.pgen.1005613 PMID: 26495981

38. Burkhardt DH, Rouskin S, Zhang Y, Li GW, Weissman JS, Gross CA. Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. Elife. 2017; 6:e22037. https://doi.org/10.7554/eLife.22037 PMID: 28139975

39. Saito K, Green R, Buskirk AR. Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. Elife. 2020; 9:e55002. https://doi.org/10.7554/eLife.55002 PMID: 32065583

40. Mustoe AM, Busan S, Rice GM, Hajdin CE, Peterson BK, Ruda VM, et al. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. Cell. 2018; 173(1):181–195. e18. https://doi.org/10.1016/j.cell.2018.02.034 PMID: 29551268

41. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. Bioinformatics. 2004; 20(16):2860–2862. https://doi.org/10.1093/bioinformatics/bth300 PMID: 15130928

42. Seiler CY, Park JG, Sharma A, Hunter P, Surapaneni P, Sedillo C, et al. DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research. Nucleic Acids Res. 2014; 42 (Database issue):D1253–60. https://doi.org/10.1093/nar/gkt1060 PMID: 24225319

43. Acton TB, Gunsalus KC, Xiao R, Ma LC, Aramini J, Baran MC, et al. Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. Methods Enzymol. 2005; 394:210–243. https://doi.org/10.1016/S0076-6879(05)94008-1 PMID: 15808222

44. Xiao R, Anderson S, Aramini J, Belote R, Buchwald WA, Ciccosanti C, et al. The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. J Struct Biol. 2010; 172(1):21–33. https://doi.org/10.1016/j.jsb.2010.07.011 PMID: 20688167

45. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015; 15 (18):3163–3168. https://doi.org/10.1002/pmic.201400441 PMID: 25656970

46. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJL, Bunk DM, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography- tandem mass spectrometry. J Proteome Res. 2009; 9(2):761–776. https://doi.org/10.1021/pr9006365 PMID: 19921851

47. Nilsson T, Mann M, Aebersold R, Yates JR 3rd, Bairoch A, Bergeron JJM. Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods. 2010; 7(9):681–685. https://doi.org/10.1038/nmeth0910-681 PMID: 20805795

48. Deuschle U, Kammerer W, Gentz R, Bujard H. Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. EMBO J. 1986; 5(11):2987–2994. https://doi.org/10.1002/j.1460-2075.1986.tb04596.x PMID: 3539589

49. Delvigne F, Baert J, Sassi H, Fickers P, Grünberger A, Dusny C. Taking control over microbial populations: Current approaches for exploiting biological noise in bioprocesses. Biotechnol J. 2017; 12 (7):1600549. https://doi.org/10.1002/biot.201600549 PMID: 28544731

50. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 2010; 107(8):3645–3650. https://doi.org/10.1073/pnas.0909910107 PMID: 20133581

51. Ang KS, Kyriakopoulos S, Li W, Lee DY. Multi-omics data driven analysis establishes reference codon biases for synthetic gene design in microbial and mammalian cells. Methods. 2016; 102:26–35. https://doi.org/10.1016/j.ymeth.2016.01.016 PMID: 26850284

52. Tunney R, McGlincy NJ, Graham ME, Naddaf N, Pachter L, Lareau LF. Accurate design of translational output by a neural network model of ribosome distribution. Nat Struct Mol Biol. 2018; 25(7):577–582. https://doi.org/10.1038/s41594-018-0080-2 PMID: 29967537

53. Jack BR, Boutz DR, Paff ML, Smith BL, Bull JJ, Wilke CO. Reduced Protein Expression in a Virus Attenuated by Codon Deoptimization. G3. 2017; 7(9):2957–2968. https://doi.org/10.1534/g3.117.041020 PMID: 28698233

54. Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, et al. mRNA structure regulates protein expression through changes in functional half-life. Proc Natl Acad Sci U S A. 2019; 116 (48):24075–24083. https://doi.org/10.1073/pnas.1908052116 PMID: 31712433

55. Ben-Yehezkel T, Atar S, Zur H, Diament A, Goz E, Marx T, et al. Rationally designed, heterologous *S. cerevisiae* transcripts expose novel expression determinants. RNA Biol. 2015; 12(9):972–984. https://doi.org/10.1080/15476286.2015.1071762 PMID: 26176266

56. Mittal P, Brindle J, Stephen J, Plotkin JB, Kudla G. Codon usage influences fitness through RNA toxicity. Proc Natl Acad Sci U S A. 2018; 115(34):8639–8644. https://doi.org/10.1073/pnas.1810022115 PMID: 30082392

57. Bindels DS, Haarbosch L, van Weeren L, Postma M, Wiese KE, Mastop M, et al. mScarlet: a bright monomeric red fluorescent protein for cellular imaging. Nat Methods. 2017; 14(1):53–56. https://doi.org/10.1038/nmeth.4074 PMID: 27869816

58. Schlechter RO, Jun H, Bernach M, Oso S, Boyd E, Muñoz-Lintz DA, et al. Chromatic Bacteria—A Broad Host-Range Plasmid and Chromosomal Insertion Toolbox for Fluorescent Protein Expression in Bacteria. Front Microbiol. 2018; 9:3052. https://doi.org/10.3389/fmicb.2018.03052 PMID: 30631309

59. Schlechter RO, Kear EJ, Remus DM, Remus-Emsermann MNP. Fluorescent Protein Expression as a Proxy for Bacterial Fitness in a High-Throughput Assay. Appl Environ Microbiol. 2021; 87(18): e00982–21. https://doi.org/10.1128/AEM.00982-21 PMID: 34260309

60. Shachrai I, Zaslaver A, Alon U, Dekel E. Cost of unneeded proteins in *E. coli* is reduced after several generations in exponential growth. Mol Cell. 2010; 38(5):758–767. https://doi.org/10.1016/j.molcel.2010.04.015 PMID: 20434381

61. Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005; 436(7050):588–592. https://doi.org/10.1038/nature03842 PMID: 16049495

62. Alon U. An Introduction to Systems Biology: Design Principles of Biological Circuits. CRC Press; 2006.

63. Babu MM, Aravind L. Adaptive evolution by optimizing expression levels in different environments. Trends Microbiol. 2006; 14(1):11–14. https://doi.org/10.1016/j.tim.2005.11.005 PMID: 16356718

64. Zaslaver A, Mayo A, Ronen M, Alon U. Optimal gene partition into operons correlates with gene functional order. Phys Biol. 2006; 3(3):183–189. https://doi.org/10.1088/1478-3975/3/3/003 PMID: 17021382

65. Kalisky T, Dekel E, Alon U. Cost-benefit theory and optimal design of gene regulation functions. Phys Biol. 2007; 4(4):229–245. https://doi.org/10.1088/1478-3975/4/4/001 PMID: 17991990

66. Tănase-Nicola S, ten Wolde PR. Regulatory control and the costs and benefits of biochemical noise. PLoS Comput Biol. 2008; 4(8):e1000125. https://doi.org/10.1371/journal.pcbi.1000125 PMID: 18716677

67. Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET, et al. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics. 2017; 18(1):529. https://doi.org/10.1186/s12859-017-1934-z PMID: 29187165

68. Zayni S, Damiati S, Moreno-Flores S, Amman F, Hofacker I, Ehmoser EK. Enhancing the cell-free expression of native membrane proteins by in-silico optimization of the coding sequence—an experimental study of the human voltage-dependent anion channel. Membranes. 2021; 11(10):741. https://doi.org/10.3390/membranes11100741

69. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA–RNA binding. Bioinformatics. 2006; 22(10):1177–1182. https://doi.org/10.1093/bioinformatics/btl024 PMID: 16446276

70. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol. 2007; 25(1):117–124. https://doi.org/10.1038/nbt1270 PMID: 17187058

71. Maier T, Schmidt A, Güell M, Kühner S, Gavin AC, Aebersold R, et al. Quantification of mRNA and protein and integration with protein turnover in a bacterium. Mol Syst Biol. 2011; 7:511. https://doi.org/10.1038/msb.2011.38 PMID: 21772259

72. Masuda T, Saito N, Tomita M, Ishihama Y. Unbiased quantitation of *Escherichia coli* membrane proteome using phase transfer surfactants. Mol Cell Proteomics. 2009; 8(12):2770–2777. https://doi.org/10.1074/mcp.M900240-MCP200 PMID: 19767571

73. Nie L, Wu G, Zhang W. Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis. Genetics. 2006; 174(4):2229–2243. https://doi.org/10.1534/genetics.106.065862 PMID: 17028312

74. Guimaraes JC, Rocha M, Arkin AP. Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. Nucleic Acids Res. 2014; 42(8):4791–4799. https://doi.org/10.1093/nar/gku126 PMID: 24510099

75. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020; 21(10):630–644. https://doi.org/10.1038/s41576-020-0258-4 PMID: 32709985

76. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990; 29(6-7):1105–1119. https://doi.org/10.1002/bip.360290621 PMID: 1695107

77. 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 PMID: 22115189

78. Bhandari BK, Lim CS, Gardner PP. Highly accessible translation initiation sites are predictive of successful heterologous protein expression. bioRxiv. 2019; https://doi.org/10.1101/726752

79. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A. 2004; 101(19):7287–7292. https://doi.org/10.1073/pnas.0401799101 PMID: 15123812

80. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006; 22(14):e90–8. https://doi.org/10.1093/bioinformatics/btl246 PMID: 16873527

81. Kiryu H, Terai G, Imamura O, Yoneyama H, Suzuki K, Asai K. A detailed investigation of accessibilities around target sites of siRNAs and miRNAs. Bioinformatics. 2011; 27(13):1788–1797. https://doi.org/10.1093/bioinformatics/btr276 PMID: 21531769

82. Mann M, Wright PR, Backofen R. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. Nucleic Acids Res. 2017; 45(W1):W435–W439. https://doi.org/10.1093/nar/gkx279 PMID: 28472523

83. Bhandari BK, Lim CS, Gardner PP. TISIGNER.com: web services for improving recombinant protein production. Nucleic Acids Res. 2021; 49(W1):W654–W661. https://doi.org/10.1093/nar/gkab175 PMID: 33744969

84. Bhandari BK, Gardner PP, Lim CS. Solubility-Weighted Index: fast and accurate prediction of protein solubility. Bioinformatics. 2020; 36(18):4691–4698. https://doi.org/10.1093/bioinformatics/btaa578 PMID: 32559287

85. Bhandari BK, Gardner PP, Lim CS. Razor: annotation of signal peptides from toxins. bioRxiv. 2020; https://doi.org/10.1101/2020.11.30.405613

86. Chin JX, Chung BKS, Lee DY. Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. Bioinformatics. 2014; 30(15):2210–2212. https://doi.org/10.1093/bioinformatics/btu192 PMID: 24728853

87. 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):W526–31. https://doi.org/10.1093/nar/gki376 PMID: 15980527

88. 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 PMID: 17439967

89. Sambrook J, Russell DW. Molecular cloning: a laboratory manual. Vol. 3. CSHL Press; 2001.

90. van Dolleweerd CJ, Kessans SA, Van de Bittner KC, Bustamante LY, Bundela R, Scott B, et al. MIDAS: A Modular DNA Assembly System for Synthetic Biology. ACS Synth Biol. 2018; 7(4):1018–1029. https://doi.org/10.1021/acssynbio.7b00363 PMID: 29620866

91. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences; 2006; 22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

92. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

93. Mohammad F, Green R, Buskirk AR. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. Elife. 2019; 8:e42591. https://doi.org/10.7554/eLife.42591 PMID: 30724162

94. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics. 2004; 5:105. https://doi.org/10.1186/1471-2105-5-105 PMID: 15296519

95. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie / Chemical Monthly. 1994; 125(2):167–188. https://doi.org/10.1007/BF00818163

96. Bernhart S, Hofacker IL, Stadler PF. Local Base Pairing Probabilities in Large RNAs. Bioinformatics. 2006; 22(5):614–615. https://doi.org/10.1093/bioinformatics/btk014 PMID: 16368769

97. Bompfünewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, et al. Variations on RNA folding and alignment: lessons from Benasque. J Math Biol. 2008; 56(1-2):129–144. https://doi.org/10.1007/s00285-007-0107-5 PMID: 17611759

98. Lorenz R, Hofacker IL, Stadler PF. RNA folding with hard and soft constraints. Algorithms Mol Biol. 2016; 11:8. https://doi.org/10.1186/s13015-016-0070-z PMID: 27110276

99. Held D, Yaeger K, Novy R. New coexpression vectors for expanded compatibilities in *E. coli*. Novagen; 2003. 18.

100. Gomes L, Monteiro G, Mergulhão F. The Impact of IPTG Induction on Plasmid Stability and Heterologous Protein Expression by Biofilms. Int J Mol Sci. 2020; 21(2). https://doi.org/10.3390/ijms21020576 PMID: 31963160

101. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by Simulated Annealing. Science. 1983; 220 (4598):671–680. https://doi.org/10.1126/science.220.4598.671 PMID: 17813860

102. Ingber L. Adaptive simulated annealing (ASA): Lessons learned. arXiv. 2000; https://arxiv.org/abs/cs/0001018

103. Keith JM, Adams P, Bryant D, Kroese DP, Mitchelson KR, Cochran DAE, et al. A simulated annealing algorithm for finding consensus sequences. Bioinformatics. 2002; 18(11):1494–1499. https://doi.org/10.1093/bioinformatics/18.11.1494 PMID: 12424121

104. Brownlee J. Clever Algorithms: Nature-inspired Programming Recipes. Jason Brownlee; 2011.

105. Lindgreen S, Gardner PP, Krogh A. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. Bioinformatics. 2007; 23(24):3304–3311. https://doi.org/10.1093/bioinformatics/btm525 PMID: 18006551

106. Gaspar P, Moura G, Santos MAS, Oliveira JL. mRNA secondary structure optimization using a correlated stem-loop prediction. Nucleic Acids Res. 2013; 41(6):e73. https://doi.org/10.1093/nar/gks1473 PMID: 23325845

107. Chen YJ, Liu P, Nielsen AAK, Brophy JAN, Clancy K, Peterson T, et al. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat Methods. 2013; 10 (7):659–664. https://doi.org/10.1038/nmeth.2515 PMID: 23727987

108. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013; 29(22):2933–2935. https://doi.org/10.1093/bioinformatics/btt509 PMID: 24008419

109. Gardner PP, Eldai H. Annotating RNA motifs in sequences and alignments. Nucleic Acids Res. 2015; 43(2):691–698. https://doi.org/10.1093/nar/gku1327 PMID: 25520192

110. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018; 46(D1):D335–D342. https://doi.org/10.1093/nar/gkx1038 PMID: 29112718

111. Lorenz WW, Cormier MJ, O'Kane DJ, Hua D, Escher AA, Szalay AA. Expression of the *Renilla reniformis* luciferase gene in mammalian cells. J Biolumin Chemilumin. 1996; 11(1):31–37. https://doi.org/10.1002/(SICI)1099-1271(199601)11:1%3C31::AID-BIO398%3E3.0.CO;2-M PMID: 8686494

112. Fuhrmann M, Hausherr A, Ferbitz L, Schödl T, Heitzer M, Hegemann P. Monitoring dynamic expression of nuclear genes in *Chlamydomonas reinhardtii* by using a synthetic luciferase reporter gene. Plant Mol Biol. 2004; 55(6):869–881. https://doi.org/10.1007/s11103-005-2150-1 PMID: 15604722

113. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12:2825–2830.

114. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44(3):837–845. https://doi.org/10.2307/2531595 PMID: 3203132

115. McKinney W. Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference; 2010. p. 51–56.

116. Oliphant TE. Python for Scientific Computing. Computing in Science Engineering. 2007; 9(3):10–20. https://doi.org/10.1109/MCSE.2007.58

117. Millman KJ, Aivazis M. Python for Scientists and Engineers. Computing in Science Engineering. 2011; 13(2):9–12. https://doi.org/10.1109/MCSE.2011.36

118. Marill KA, Chang Y, Wong KF, Friedman AB. Estimating negative likelihood ratio confidence when test sensitivity is 100%: A bootstrapping approach. Stat Methods Med Res. 2017; 26(4):1936–1948. https://doi.org/10.1177/0962280215592907 PMID: 26152746

119. R Core Team. R: A Language and Environment for Statistical Computing; 2019.

120. Hunter JD. Matplotlib: A 2D Graphics Environment Comput Sci Eng. 2007; 9(03):90–95. https://doi.org/10.1109/MCSE.2007.55

121. Waskom M, Botvinnik O, O'Kane D, Hobson P, Ostblom J, Lukauskas S, et al. mwaskom/seaborn: v0.9.0 (July 2018). 2018. https://doi.org/10.5281/zenodo.1313201