Nucleic Acids Research, 2025, **53**, gkaf233
https://doi.org/10.1093/nar/gkaf233
Computational Biology

A deep learning model trained on expressed transcripts across different tissue types reveals cell-type codon-optimization preferences

Sandhiya Ravi[1,2,†], Tapan Sharma[1,2,†], Mitchell Yip[1], Huiya Yang[1], Jun Xie[1,2], Guangping Gao[1,2,3,*], Phillip W.L. Tai[1,2,3,*]

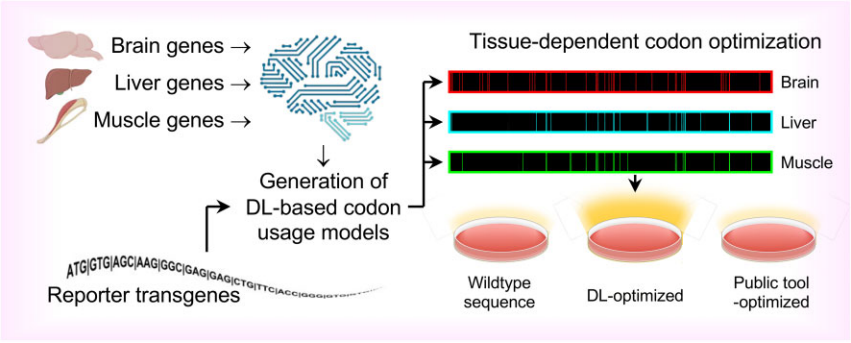
[1]Department of Genetic and Cellular Medicine, UMass Chan Medical School, Worcester, MA 01605, United States
[2]Department of Microbiology, UMass Chan Medical School, Worcester, MA 01605, United States
[3]Li Weibo Institute of Rare Diseases Research, UMass Chan Medical School, Worcester, MA 01605, United States

[*]To whom correspondence should be addressed. Email: guangping.gao@umassmed.edu
Correspondence may also be addressed to Phillip W.L. Tai. Email: phillip.tai2@umassmed.edu
[†]Equal contributors

Abstract

Species-specific differences in protein translation can affect the design of protein-based drugs. Consequently, efficient expression of recombinant proteins often requires codon optimization. Publicly available optimization tools do not always result in higher expression levels and can lead to protein misfolding and reduced expression. Here, we aimed to develop a novel deep learning (DL) tool using a recurrent neural network (RNN) to define cell type-dependent codon biases. Using gene expression data from three different tissue types (brain, liver, and muscle) and all secretory genes, we trained DL models to predict optimal codon usage. Codon-optimized sequences for test reporter genes exhibited enhanced protein expression compared to their original sequences and those optimized using a publicly available tool. Interestingly, DL models trained on genes expressed in liver cells (hepatocytes) resulted in the highest levels of expression when tested *in vitro*, irrespective of the cell type. Our findings also demonstrate that DL-based codon optimization algorithms can significantly enhance protein translation, particularly for secretory proteins, which are crucial for therapeutic applications. This research represents a novel approach to codon optimization with broader implications for protein-based pharmaceuticals, vaccine manufacturing, gene therapy, and other recombinant DNA products.

Graphical abstract



Introduction

The process of protein translation is dictated by the genetic code, a degenerate system of ribonucleic acid bases consisting of 64 unique codons that encode for the 20 standard amino acids and termination signals. Despite the inherent redundancy, selection of codons is nonrandom and codon biases exist [1]. This bias is thought to align with the cellular abundance of transfer (t)RNAs that carry the anti-codon sequences and are charged with specific amino acids [1, 2]. A gene's codon composition can significantly influence protein translation rates and abundances, and is determined by multiple factors [3–6]. There exists a direct link between the frequency of specific codons in a host organism's transcriptome and the expression levels of corresponding tRNAs [7]. Importantly, differences in codon use between organisms can influence the manufacturing of protein-based drugs, where bacterial, yeast, or mammalian cell expression systems are employed for large-scale production. There is also a growing need to opti-

Received: September 19, 2024. **Revised:** March 3, 2025. **Editorial Decision:** March 7, 2025. **Accepted:** March 28, 2025

mally express protein-coding genes in different model systems. Thus, codon optimization has been a key step for cross-species expression (i.e. protein expression in heterologous hosts) [7–9]. Publicly available algorithms exist that claim to optimize codons to achieve efficient protein translation. However, excessive dependence on high-frequency codons can result in protein misfolding and/or formation of insoluble proteins [10]. To avoid such outcomes, it is crucial to not just rely on synonymous codon frequency look-up tables. Furthermore, these publicly available tools do not seem to always generate the best optimized sequences. For example, many optimization tools will output variations of a queried sequence. These multiple "optimized" sequences often result in variable levels of expression and may even result in decreases in expression. Ideally, optimizing codon usage should yield consistent results with improvements to protein expression. These issues have raised the level of difficulty for designing optimized sequences. One such need exists in the gene therapy field, where research relies on multiple models to demonstrate safety and efficacy.

Gene therapy has been revolutionary for treating a wide array of genetic diseases [11]. Despite advancements in the field, the quest for optimizing transgene cassettes to ensure safe, effective, and efficient gene delivery is still ongoing. Subtherapeutic expression levels in target tissues remains a major barrier for gene therapy success. Therefore, many vector designs have human codon-optimized sequences that are tested in proof-of-concept studies using cell culture and animal models. Unfortunately, these may lead to suboptimal expression in test models, as a result of species-specific codon usage biases. Many publicly available websites have codon optimization tools that can produce sequences tailored to specific species. However, tissue- or cell-type differences adds another degree of complexity. Tissues in the human body play different roles and have different physiological demands. Not only do tissues express different sets of genes that differentiate them from others, but genes that are expressed in multiple tissues may have distinct relative abundances between cell types. Furthermore, the function of the protein can also dictate its demands for when and how they are expressed. For example, secreted proteins are synthesized in the rough endoplasmic reticulum and transported by the secretory network to the extracellular space [12, 13]. These proteins utilize a signal recognition peptide (SRP), which typically spans 15–70 amino acids at the N-terminus of the nascent polypeptide chain and facilitate their co-translational targeting to the ER lumen [14]. Secreted proteins are tightly regulated by cellular mechanisms [13], and irregularities in their expression have been observed in various pathological conditions [15].

Codon optimization can also unintentionally result in the increase of CpG dinucleotides. In mammals, CpG dinucleotides are typically methylated at the fifth carbon of the heterocyclic aromatic ring of cytosine. The presence of unmethylated CpG dinucleotides in foreign DNA, including those in viral vectors, is known to be deleterious. Unmethylated CpG dinucleotides can be recognized by toll-like receptors, thus activating innate immune responses and result in clearing of the delivered therapeutic gene from the target tissue [16–18].

Given that codon selection is one means of controlling protein expression, it can be hypothesized that proteins with high levels of expression utilize optimized codons for efficient translation. Thus, in order to understand codon biases, one would only need to explore the codons used in highly expressed proteins. Unfortunately, complete proteome profiles encompassing multiple tissue/cell types are incomplete. In addition, protein abundance is also controlled by cellular mechanisms that can cleave, degrade, or stabilize proteins in the cell. An alternative approach is to use messenger RNA (mRNA) abundance as a proxy to identify genes that utilize optimal codons. Whole-transcriptome profiles for many tissue/cell types gained from high-throughput RNA sequencing (RNA-seq) studies have been published and their data are publicly available and easy to access. Furthermore, RNA-seq data have been standardized in such a way that transcript abundances can be compared across studies.

Emerging machine-learning strategies, specifically deep learning (DL) systems [19], provide promising opportunities and advancements for studying complex systems in biology. Recurrent neural networks (RNNs), a subset of deep neural networks, excel in processing sequential data, making them an ideal choice for tasks demanding the comprehension of sequential information [20]. RNNs have the potential to facilitate improved synonymous codon selection by identifying the underlying patterns of synonymous codon usage [21]. Therefore, in order to address the complexity of codon biases, we had two goals: (i) leverage the ability of DL to reveal *de novo* patterns in codon usage and (ii) provide a proof-of-concept DL strategy based on mRNA expression to define tissue-dependent codon-optimization rules to increase protein expression in target cells. In this work, we have leveraged the power of RNNs to develop novel models that use a bidirectional long short-term memory (BiLSTM) architecture. Known for its ability to process information from both past and future sequences, this architecture has enabled us to train codon usage models on highly expressed genes found in three tissue groups: brain, liver, and muscle. To test the robustness of the approach, codon-optimized reporter gene sequences produced by DL models trained on tissue/cell transcripts were compared with sequences optimized by a commercially available and widely used tool in relevant mouse cell lines. Transgenes optimized by our DL-based models conferred higher expression in the appropriate cell lines as compared to those achieved by the wild-type transgene sequences. Interestingly, we found that the tissue-dependent codon-optimization models did not always yield transgenes that performed the best when tested in the corresponding cell types. Nonetheless, all of our DL models performed better than sequences optimized using the commercial tool. Our findings represent a novel approach to codon optimization and a significant advancement in the application of DL for sequence optimization, with potential implications in protein-based pharmaceuticals, vaccine manufacturing, gene therapy, and other recombinant DNA products.

Materials and methods

Reagents

The following reagents, instruments, and software packages were used for related cell culture experiments. Dulbecco's Modified Eagle Medium (DMEM) (Cat #11965, Thermo Fisher Scientific, Waltham, MA, USA), fetal bovine serum (FBS) (Ca t#A4736201, Thermo Fisher Scientific, Waltham, MA, USA), penicillin/streptomycin (Cat #15140122, Thermo Fisher Scientific, Waltham, MA, USA), insulin-transferrin-selenium (ITS) solution (Cat #41-400-045, Thermo Fisher Scientific, Waltham, MA, USA), dexamethasone (Cat #A13449,

Thermo Fisher Scientific, Waltham, MA, USA), horse serum (Cat #16050122, Thermo Fisher Scientific, Waltham, MA, USA), bovine insulin (Cat #I0516, Sigma–Aldrich, St. Louis, MO, USA), Lipofectamine 3000 (Cat #L3000015, Thermo Fisher Scientific, Waltham, MA, USA), 1× phosphate-buffered saline (PBS) (Cat #SH30028.LS, Cytiva, Marlborough, MA, USA), Trypsin (Cat #25200114, Gibco, Grand Island, NY, USA), Dual-Luciferase Reporter Assay (Cat #E1980, Promega Corporation, Madison, WI, USA), Leica DMI 6000B inverted microscope with a DFC7000 GT camera (Leica Camera AG, Wetzlar, Hesse, Germany), Attune™ NxT Acoustic Focusing Cytometer (Thermo Fisher Scientific, USA), BioTek Synergy HTX multimode plate reader with an automated dispenser (Agilent Technologies, Inc., Santa Clara, CA, USA), and FlowJo™ v10.8 Software (BD Life Sciences, USA).

Biological resources

Enhanced green-fluorescent protein (EGFP), firefly luciferase (*FLuc)*, Gaussia luciferase (*GLuc)*, and *Renilla luciferase* constructs were obtained from the UMass Chan Viral Vector Core. The following cells were used in this study: The C2C12 myoblasts (ATCC, Manassas, VA, USA), Neuro-2a neuroblasts (ATCC, Manassas, VA, USA), and AML12 hepatocytes (ATCC, Manassas, VA, USA).

Statistical analyses

Data generated from flow cytometry and dual-luciferase assays are shown as mean ± the standard deviation for at least three independent biological replicates. Statistical analyses were performed using either one-way or two-way ANOVA test in Prism9 (GraphPad Prism Software Inc., USA). *P*-values for waterfall plots were determined by Wilcoxon rank-sum tests.

Novel programs, software, and algorithms

All custom codes and scripts are hosted on deposited the code and data to Zenodo (https://doi.org/10.5281/zenodo.14991160).

Web sites/data base referencing

The brain datasets included bulk RNAseq of P20 mouse forebrain (GSE171765) [22], bulk RNAseq of eight-week-old mouse cortex (GSE188989) [23], and bulk RNAs from P0 mouse brains (GSE222464) [24]. Additionally, an *in vitro* dataset representing the transcriptome of Neuro2a cells (GSE206057) [25] was included. For liver, the datasets included transcripts from isolated hepatocytes (GSE216114) [26], bulk RNAs from P1 mouse livers (GSE201587) [27], and bulk RNAs from 11-week-old mouse livers (GSE140147) [28]. The *in vitro* dataset GSE209768, representing AML12 hepatocyte transcriptome [29], was also utilized. Muscle datasets included GSE156496, representing bulk RNAs detected in tibialis anterior (TA) muscle of 4-week-old mice [30], GSE68915, representing bulk RNAs from gastrocnemius muscle of 10-week-old mice [31], and GSE71679, representing bulk RNAs from TA muscle of 6-week-old mice [32]. The *in vitro* dataset GSE90175, representing transcripts present in C2C12 myocytes [33] after 60 h of differentiation was also included.

The following web-based analytical tools and visualization programs were used in this study: GenScript's Rare Codon Analysis Tool (https://www.genscript.com/tools/rare-codon-analysis), Clustal Omega (https://www.ebi.ac.uk/jdispatcher/msa/clustalo), Codon Optimization Tool (https://www.genscript.com/gensmart-free-gene-codon-optimization.html), Galaxy (version 3.12.0), ShinyGO (version 0.80), clusterProfiler (version 4.0.5) package in Rstudio, Keras API for TensorFlow (https://github.com/fchollet/keras), and Netron (https://netron.app/).

Selection of genes and data preprocessing

Transcript selection (nonmutually exclusive) in our study specifically focused on genes that are highly expressed in the brain [22–25], liver [26–29], and muscle [30–33]. "Highly expressed" genes were defined as transcripts from the top 5%–15% genes of the selected *in vivo* RNA-seq datasets (Supplementary File). Only the genes identified in all datasets for a particular tissue were considered for final training. Notably, expression of the selected genes in RNA-seq datasets generated in cell lines is weakly correlated with whole-tissue datasets. Gene names (ENSMUS IDs) and sequences were acquired and gene identifiers (Gene symbols and NCBI RefSeq accession numbers) were obtained using Galaxy [34]. From this set, protein-coding genes were extracted (refseq accession with NM prefixes) using a custom script. As a crucial precursor to model training, we implemented a rigorous validation method with data pre-processing. Initially, we verified the length of each coding sequence (CDS) region, ensuring it was divisible by three, a requisite of full-length protein-coding genes. Once validated, we translated these codons into corresponding amino acid sequences based on standard mammalian genetic code. Our second validation involved confirming that DNA sequences consisted solely of the standard DNA bases: adenine (A), cytosine (C), guanine (G), or thymine (T). We mandated that each sequence commenced with the "ATG" start codon, which encodes for methionine, and terminates with a stop codon (TAA, TAG, or TGA). This workflow resulted in 677 brain transcripts, 868 muscle transcripts, and 904 liver transcripts. Density plots were generated using ShinyGO [35]. Dot plots for GO term enrichment were generated using the clusterProfiler package in Rstudio [36].

Sequence training pipeline

For our analysis, we utilized a robust training dataset comprised of 2449 genes, divided in an ~85:10:5 ratio for training, evaluation, and prediction, respectively. A fixed split strategy is designed to ensure that the prediction set remains independent and is exclusively reserved for evaluating the model's generalization performance [37, 38]. The 85% subset of genes was further split into 80% for training the model and 20% for testing it. A set of 126 genes (35 brain genes, 45 liver genes, and 46 muscle genes) served as benchmark references and was used for validation. To ensure consistency, the data underwent a series of pre-processing steps, including tokenization and sequence padding. These steps were implemented to ensure that all sequences within our dataset had uniform lengths, facilitating effective training and analysis.

- Sequence tokenization

N-paired DNA and amino acid (AA) sequences defined the dataset $\{(DNA_i, AA_i)\}\ for\ i = 1\ to\ N$. Each DNA sequence DNA_i consists of a nucleotide sequence of length L_i [39].

Because a codon is a sequence of three nucleotides, we tokenized DNA_i into a sequence of $L_i/3$ nonoverlapping codons, where each codon c_j is mapped to a unique word in a DNA vocabulary V_{DNA}.

$$DNA_i = (c_1, c_2, \cdots, c_{L_i/3})$$

Each codon is an element of the DNA sequence: $c_j \in V_{DNA}$, where $|V_{DNA}|$ is the DNA vocabulary size. This allows us to represent DNA_i as a sequence of word indices: $DNA_i = (w_1, w_2, \cdots, w_{L_i/3})$, where $w_t \in \{1, 2, \cdots, |V_{DNA}|\}$.

We applied the same procedure to each AA sequence AA_i of length M_i, tokenizing it into a sequence of M_i residues, with each residue r_l mapped to a unique word in the AA vocabulary V_{AA}:

$$AA_i = (r_1, r_2, \cdots, r_{M_i})$$

Each $r_l \in V_{AA}$, where $|V_{AA}|$ is the AA vocabulary size.

This yields $AA_i = (z_1, z_2, \cdots, z_{M_i})$, where $z_t \in \{1, 2, \cdots, |V_{AA}|\}$.

The tokenizers used for sequence-token conversions were also stored for applying the models to predict codon-optimized sequences. Model performance evaluations were carried out by assessing their accuracy, a metric used to measure the model's predictive capabilities.

- *Sequence padding*

To enable batch processing, we padded each DNA and AA sequence with zero vectors to a maximum length N [40]:

If $L_i/3 < N$, the padded DNA sequence is: $DNA_i' = (w_1, w_2, \cdots, w_{L_i/3}, 0, 0, \cdots, 0)$

If $M_i < N$, the padded AA sequence is: $AA_i' = (z_1, z_2, \cdots, z_{M_i}, 0, 0, \cdots, 0)$ where, DNA_i' and AA_i are the padded sequences of length N.

- *Sequence embedding*

We mapped each discrete word index to a D-dimensional dense vector representation via an embedding matrices [41]. $E_{DNA} \in R^{|V_{DNA}| \times D}$ *and* $E_{AA} \in R^{|V_{AA}| \times D}$:

$$x_t = E_{DNA}[w_t], y_t = E_{AA}[z_t]$$

Yielding embedded inputs, $x \in R^{N \times D}$ *and* $y \in R^{N \times D}$.

In our model, sequence embedding is achieved through a matrix initialized in the first layer of the network. Each codon (or amino acid) is mapped to a dense vector representation in a D-dimensional space, where $D = 128$. The embedding matrix is randomly initialized at the start of training, with each row representing a codon (or amino acid) and each column corresponding to one of the embedding dimensions. Embeddings are refined during training through backpropagation, allowing the model to capture meaningful relationships between codons or amino acids based on the patterns in the training data. This approach facilitates the generation of optimized sequences that enhance translation efficiency.

- *Recurrent neural network*

The architecture of each RNN model was constructed using the Keras API for TensorFlow (https://github.com/fchollet/keras). Key components of this architecture included an embedding layer of input sequences x that are processed by a two-layer bidirectional gated recurrent unit (GRU) with H hidden units [42]. We assume that the output from the first layer is the input to the second layer.

- *Forward GRU*

The update and the reset gates are respectively computed as:

$$z_{ft} = \sigma\left(W_{zf}x_t + U_{zf}h_{t-1} + b_{zf}\right)$$

$$r_{ft} = \sigma\left(W_{rf}x_t + U_{rf}h_{t-1} + b_{rf}\right)$$

Then, the hidden state is updated:

$$h_{ft} = z_{ft} \odot h_{ft-1} + (1 - z_{ft}) \odot \widehat{h_{ft}}$$

where $\odot$ denotes element-wise multiplication, and :

$$\widehat{h_{ft}} = \tanh\left(W_{hf}x_t + U_{hf}\left(r_{ft} \odot h_{t-1}\right) + b_{hf}\right)$$

- *Backward GRU*

Similarly for the backward pass, the update and reset gates are respectively:

$$z_{bt} = \sigma\left(W_{zb}x_t + U_{zb}h_{t+1} + b_{zb}\right)$$

$$r_{bt} = \sigma\left(W_{rb}x_t + U_{rb}h_{t+1} + b_{rb}\right)$$

And the hidden state is updated:

$$h_{bt} = z_{bt} \odot h_{bt+1} + (1 - z_{bt}) \odot \widehat{h_{bt}}$$

where:

$$\widehat{h_{bt}} = \tanh\left(W_{hb}x_t + U_{hb}\left(r_{bt} \odot h_{t+1}\right) + b_{hb}\right)$$

We concatenated the forward and backward hidden states to obtain our final hidden representation at each timestep:

$$h_t = \left[h_{ft}, h_{bt}\right]$$

A dropout layer was incorporated to mitigate overfitting during the training phase if the validation loss failed to decrease for five consecutive epochs (Supplementary Fig. S3).

The training procedure was executed in batches of 16 samples per batch across ten epochs. To prevent overfitting, we terminated the training process. To determine the theoretical maximum accuracy (TMA) value for each training data set [38], we developed a customized script for the reported calculations. Briefly, the codon with the highest frequency in the dataset was identified for each amino acid. We then calculated the proportion of each amino acid relative to the total number of amino acids in the dataset. The frequency of the most common codon for each amino acid was multiplied by the frequency of the amino acid itself. These values were then summed for all amino acids to obtain the final TMA.

- *Prediction and optimization*

A time-distributed dense layer, which was applied at each timestep independently, predicted the DNA sequence from the hidden states [43]:

$$p\left(y_t|h_t\right) = \text{softmax}\left(Wh_t + b\right)$$

The model was trained to minimize the sparse categorical cross entropy between predicted and true DNA sequences across all timesteps.

Model architecture

We employed tissue-specific sequential neural network models for our analyses. Although the core architecture remains the same across all three tissues (brain, liver, and muscle), the models differ in the length of sequences they process. Each model consists of five key layers: (i) embedding layer, transforms integer-encoded vocabulary into fixed-size dense vectors; (ii) bidirectional LSTM layer, captures both forward and backward sequence dependencies; (iii) time-distributed dense layer (first), prepares the sequence for the final classification layer; (iv) dropout layer, prevents overfitting by randomly setting a fraction of input units to 0 during training; and (v) time-distributed dense layer (second), outputs a multi-class classification for each time step in the sequence.

Brain-specific model

Embedding layer: (none, 3054, 128), 2816 parameters
Bidirectional LSTM layer: (none, 3054, 32), 14 016 parameters
Time-distributed dense layer (first): (none, 3054, 128), 4224 parameters
Dropout layer: (none, 3054, 128), 0 parameters
Time-distributed dense layer (second): (None, 3054, 65), 8385 parameters

Liver-specific model

Embedding layer: (none, 4546, 128), 2816 parameters
Bidirectional LSTM layer: (none, 4546, 32), 14 016 parameters
Time-distributed dense layer (first): (none, 4546, 128), 4 224 parameters
Dropout layer: (none, 4546, 128), 0 parameters
Time-distributed dense layer (second): (none, 4546, 65), 8385 parameters

Muscle-specific model

Embedding Layer: (None, 8033, 128), 2816 parameters
Bidirectional LSTM Layer: (None, 8033, 32), 14 016 parameters
Time-distributed dense layer (First): (None, 8033, 128), 4224 parameters
Dropout layer: (None, 8033, 128), 0 parameters
Time-distributed dense layer (Second): (None, 8033, 65), 8385 parameters

All three tissue-specific models have the same number of trainable parameters (29 441). This is because they share an identical layer architecture and configurations, differing only in the sequence lengths they are designed to process. The uniformity in parameter count ensures a consistent computational complexity across models, facilitating a more straightforward comparison of their performances.

Uniform random choice and background frequency choice

For the uniform random choice (URC) approach [37], a simple lookup table was employed that mapped amino acids to their corresponding codons (Supplementary Table S1). This table contains a list of amino acids and their corresponding codons. The URC algorithm randomly selects a codon that corresponds to that amino acid from the lookup table without any consideration of frequency or context. The background frequency choice (BFC) approach uses a more complex lookup table that specifies the amino acids and their associated codons [37], along with probabilities indicating the frequency of occurrence for each codon in *Mus musculus* (Supplementary Table S2).

Metrics evaluations

Codon adaptation index (CAI), codon frequency distribution (CFD), directional codon bias score (DCBS), weighted sum of relative entropy (E_W), maximum-likelihood codon bias (MCB), frequency of optimal codons (FOP), and effective number of codon pairs (ENcp) were calculated using GenScript's Rare Codon Analysis tool. The CpG dinucleotide content was calculated using a custom script. Alignments were performed using Clustal Omega [44, 45].

Of note, CAI requires a reference codon usage table to calculate relative adaptiveness values for each codon. In this study, we used the GenScript Rare Codon Analysis Tool [46] to calculate CAI values, selecting *Mus musculus* as the reference organism. This approach ensured that all CAI calculations were based on a single codon usage table, and thus, a consistent basis for evaluating sequences optimized by our DL models, wild-type sequences, and those optimized by other platforms.

Plasmid constructs

The EGFP and *FLuc* constructs have been described previously [47, 48]. The GenScript-optimized sequences for EGFP and *FLuc* were optimized using the Codon Optimization tool [49]. Brain-, liver-, and muscle-optimized sequences were generated using the described DL models. The modified EGFP and *FLuc* sequences were cloned downstream of the chicken β-actin (*CB*) promoter in the pAAV-*CB6* plasmid using standard molecular cloning methods.

Cell culture and plasmid transfections

Mouse skeletal myoblasts (C2C12) and mouse neuroblasts (Neuro-2a) were maintained in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin in a humidified incubator at 37°C in 5% CO_2. Mouse hepatocytes (AML12) were cultured in DMEM:F12 medium supplemented with 10% FBS, 1× ITS solution, 40 ng/ml dexamethasone, and 1% penicillin/streptomycin in a humidified incubator at 37°C in 5% CO_2. For C2C12 differentiation, cells at >70% confluency were switched to DMEM medium supplemented with 2% horse serum and 2 μg/ml of bovine insulin. For transfections, 2.0E + 4 cells per well were seeded in 24-well plates and allowed to attach. Cells were transfected using Lipofectamine 3000 following the manufacturer's recommended protocol in triplicate or quadruplicate with plasmids containing the original or codon-optimized sequences. For FLuc plasmids, cells were co-transfected with a *Renilla* luciferase spike-in reference plasmid. For GLuc plasmids, cells were co-transfected with an Fluc plasmid spike-in reference.

Quantification of EGFP *expression by epifluorescence microscopy and flow cytometry analysis*

Cells transfected with the EGFP constructs were imaged 48 h post-transfection using Leica DMI 6000B inverted microscope with a DFC7000 GT camera. Cells were then washed with 1× PBS, trypsinized, and collected in 1 ml of ice-cold PBS. EGFP

expression was quantified by counting at least 10 000 events in biological quadruplicate using a Attune™ NxT Acoustic Focusing Cytometer. Acquired data was analyzed using FlowJo™ software. The percentage of EGFP-positive cells and mean fluorescence intensity (MFI) was calculated and reported.

Quantification of FLuc expression by dual-luciferase reporter assay

Culture media were removed from cells collected 48 h post-transfection. 100 µl of passive lysis buffer was added to all wells, and lysis was performed for 30 min on a rocker at room temperature. Lysates were homogenized and collected in 1.5 ml of Eppendorf tubes. 20 µl of lysate from each sample was used for analysis using the Dual-Luciferase Reporter Assay kit, according to manufacturer's recommended procedures, and read on a BioTek Synergy HTX multimode plate reader with an automated dispenser. Relative firefly luciferase activity was reported as percentage normalized to *Renilla* luciferase activity.

For dual-luciferase assays to evaluate expression of GLuc, cells were harvested 48 h post-transfection with either wild-type or different codon-optimized GLuc versions. 200 µl of conditioned media was collected from all wells, and 20 µl of sample was used directly to assay GLuc expression. Residual media was removed from these plates, and 100 µl passive lysis buffer was added after a gentle PBS wash. Twenty microliters of the lysate collected was used to assay for FLuc control. Protocol recommended by the manufacturer (Promega) was followed. Readouts were normalized and reported as relative luciferase activity.

Results

Selection of RNA-seq data and analyses of tissue-defined transcripts

Genome-wide expression at the transcript and protein levels has been shown to be influenced by a variety of factors at the post-transcriptional and post-translational levels [50]. These factors may affect overall protein abundances at the steady state. Numerous studies have explored the correlation between transcript levels and protein *in vivo* and *in vitro* [50–55]. While there is a lack of perfect correlation between mRNA expression and protein levels *in vitro*, there exists a relatively higher correlation for highly expressed genes [50, 54, 55]. Based on these observations, we reasoned that whole-transcriptomics data can be used as a proxy for highly expressed proteins. This rationale is also bolstered by the widespread availability and ease of generating transcript expression data as compared to proteomics-based expression studies. For this proof-of-concept study, we selected to focus our model training on transcriptomics data from three diverse tissue types: brain, liver, and muscle. In addition, we chose to focus our analysis on mouse transcripts, since immortalized mouse cell lines representing the three cell types are easy to work with and are readily available for validating our model-trained codon optimized transgenes.

Publicly deposited RNA-seq datasets were acquired from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. Brain, liver, and muscle datasets included both *in vivo* and *in vitro* samples from mice at various developmental stages and differentiated cell lines. In our selection of datasets, each tissue type was represented by three *in vivo* and one *in vitro* RNA-seq dataset. The *in vitro* datasets were derived from Neuro2a, AML12, and C2C12 cell lines, corresponding to neurons (brain), hepatocytes (liver), and myocytes (muscle), respectively. For each tissue set, the highest expressed genes common between the four datasets were considered (Supplementary File).

The brain dataset comprised of 507 unique genes, the liver dataset consisted of 593 unique genes, and the dataset for muscle comprised of 602 unique genes (Fig. 1A). Interestingly, 59 genes were common to all three tissue groups. Excluding these 59 genes, 174 were shared between liver and muscle transcripts, making up 29.34% and 28.91% of their respective gene lists. In contrast, 42 genes were shared between brain and liver, and 69 genes were shared between brain and muscle (Fig. 1A). These data suggested that the transcriptional profile of liver and muscle tissues had more in common with one another than with transcripts abundantly found in brain tissues.

We next characterized the CDSs of genes that were selected for model training from each tissue (brain, liver, and muscle) (Supplementary File). There exists a negative correlation between sequence length and expression levels (i.e. genes that are highly expressed tend to be shorter in length) [56]. Indeed, density plots revealed a shift toward shorter CDS lengths in the selected highly expressed brain, liver, and muscle genes compared to the lengths of all genes (Fig. 1B). We also observed an increased shift in the density plots of transcript guanine/cytosine (GC) content among the selected genes compared to those found in all genes. These data suggest that the nucleic acid sequences of genes highly abundant in the brain, liver, and muscle are shorter and GC-rich.

We next asked whether the highly expressed genes in the three tissue types had specific ontological functions. We performed independent gene ontology (GO) enrichment analyses with the brain genes, liver genes, and muscle genes. The top ontologies defined by the brain and muscle genes were those related to energy metabolism, including processes such as mitochondrial organization, electron transport chain, cellular respiration, and ATP metabolic processes (Supplementary Fig. S1A). In contrast, the top ontologies for the liver genes were related to various metabolic and catabolic processes, particularly involving small molecules, fatty acids, and organic acids (Supplementary Fig. S1A).

It was previously described that codons for high abundance genes tend to end in guanine or cytosine (G/C), whereas codons of low abundance genes tend to end in adenine or thymine (A/T) [57, 58]. This phenomenon was attributed to genes involved in cellular proliferation, which are usually enriched with A/T-ending codons, while genes involved in differentiation use G/C-ending codons [49]. To better understand codon usage patterns with regard to A/T- or G/C-ending codons, we examined the genes we selected for model training. We found that all three sets of selected genes predominantly use a higher frequency of codons ending with G/C (Fig. 1C). This trend was statistically significant, suggesting that the translation of tissue-specific genes favors G/C-ending codons.

DL model training

To construct a model capable of identifying the optimal DNA sequence for a given protein, we followed a structured pipeline (Supplementary Fig. S2). Separate RNNs were created and trained for the brain, liver, and muscle tissues using selected mouse transcripts. The training process began with data col-

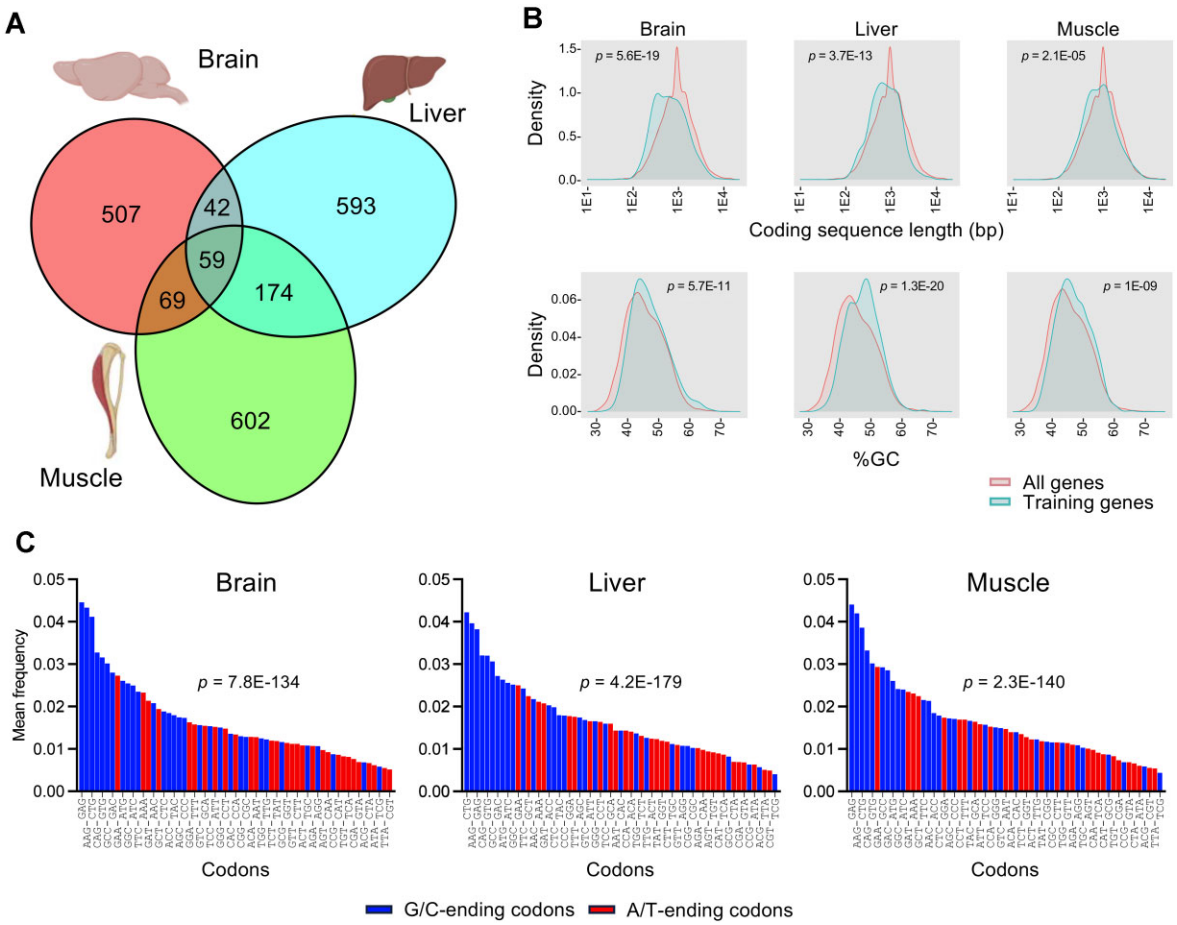


Figure 1. Comparative analyses of the selected highly expressed tissue-derived mouse genes for DL training. (**A**) A Venn diagram showing the overlap between brain, liver, and muscle genes. (**B**) Density plots representing CDS length (in base pairs, bp) and GC content (%). The *y*-axes represent the length of CDSs (bp) or %GC content. (**C**) Waterfall plots of the mean frequency of G/C-ending codons (blue bars) and A/T-ending codons (red bars) among the sets of training genes. *P*-values were determined by Wilcoxon rank sum test.

lection, where matched DNA and amino acid sequences were gathered from the top expressed genes in each tissue type. In the data preparation phase, the datasets were split into training, evaluation, and testing subsets to ensure a comprehensive and balanced approach for model development and validation (Supplementary File). For the brain dataset, a total of 677 sequences were divided into 575 for training, 67 for evaluation, and 35 for testing. Similarly, the liver dataset, comprising 868 sequences, was split into 737 for training, 86 for evaluation, and 45 for testing. The muscle dataset contained 904 sequences, divided into 768 for training, 90 for evaluation, and 46 for testing. During data preprocessing, these subsets were processed into FASTA sequences and prepared as .json files for analysis and training. Tokenization and format conversion were performed to create a machine-readable dataset. This systematic division ensured that each model was trained, evaluated, and tested on distinct yet representative sets of sequences, enabling robust performance assessment and validation across tissue types. Model development was centered on a DL architecture, represented by a trained bidirectional GRU (BiGRU) network, which serves as the computational backbone of the codon optimization tool. The model takes amino acid sequences as input and generates nucleotide sequences optimized for the codon biases of the host genome. To improve generalization and prevent overfitting, techniques such as early stopping and dropout were applied during the training phase. Metrics evaluation and benchmarking steps are described below.

Model evaluation and theoretical maximum accuracy

Upon completion of the training, each model showcased distinct characteristics. Due to redundancy within the genetic code, the training led to moderate levels of accuracy (Supplementary Fig. S3). Specifically, the brain, liver, and muscle models achieved accuracy scores of 0.517, 0.535, and 0.533 after nine epochs, respectively. This outcome can be explained by the fact that even when the model selects a suitable codon that accurately translates to the given amino acid, it is still considered a misclassification unless it matches the exact codon present in the original DNA sequence. To define the baseline for evaluating model performance, we calculated the TMA, which represents the highest achievable accuracy when selecting the most frequently used codon for each amino acid in the dataset. The TMA values for the complete datasets were 0.4808 for liver, 0.4725 for muscle, and 0.4804 for brain. Exceeding the TMA indicates that our models were not solely relying on high-frequency codons, but were able to capture nuanced relationships within codon sequences. For example, the

brain, liver, and muscle models demonstrated improvements over the TMA by 7.62%, 11.28%, and 12.81%, respectively. This finding highlights the models' ability to leverage codon–context relationships, validating the effectiveness of the training approach.

Evaluation of codon-optimized sequences with DL models

The effectiveness of the model-optimized genes were compared with four sequence sets: the wild-type gene sequences, sequences obtained by URC, BFC [37], and codon optimization with a publicly available tool from GenScript. To evaluate the optimized sequences, we employed several established indices. The first and primary metric for model evaluation was CAI [59], which assesses how closely the codon usage of the gene sequence matches those of other highly expressed genes in the training set. Our DL models outperformed both wild-type and GenScript-optimized sequences, with brain-optimized sequences showing a 3.6% CAI increase and muscle-optimized sequences achieving a 7.5% improvement (Fig. 2A), while those produced by the Genscript tool and BFC showed nonsignificant changes. URC performed the worst across tissues.

Our study goals emphasized the impact of low-frequency codons (defined as those with <30% usage in the host genome) on translation efficiency [37]. Codons below this threshold are classified as rare and can adversely affect translation by causing ribosomal stalling or slower translation rates. Reducing the percentage of rare CFD is therefore critical. Using CFD as an additional evaluation metric [37], we found that our DL models significantly reduced CFD compared to other approaches (Fig. 2B). For example, the brain-optimized model achieved a 50% reduction ($P < 0.001$), while the muscle-optimized model showed a 60% reduction. In contrast, sequences generated by GenScript showed no significant changes. These findings highlight the superior performance of our tissue-optimization models in enhancing codon usage and potentially improving protein expression.

CpG dinucleotide and overall GC content

A critical, yet often overlooked facet of codon optimization is the propensity to increase the G/C base content within an optimized sequence. High GC content can contribute to mRNA stability and efficient protein translation [60, 61]. The optimal GC-content for recombinant genes is commonly recognized to fall within the 30%–70% range. Any deviation from this interval can negatively impact the rate of transcription and translation. However, the GC content of a sequence can also impact DNA stability and the abundance of CpG dinucleotides. For therapeutic applications, CpG content is an important parameter to consider when designing DNA-based biomedicines [62, 63]. Our DL models yielded sequences with a small but significant increase in GC content within this range (Fig. 2C). In contrast, sequences obtained from the GenScript-optimized and BFC method did not result in any significant changes in GC content. The URC model led to a reduction in GC percentage across all benchmark genes. Unexpectedly, our DL models consistently decreased CpG dinucleotide frequencies (Fig. 2D).

Evaluation of nonuniform codon usage of predicted benchmark gene sequences

We next evaluated the performance of the DL models for nonuniform codon usage by analyzing the DCBSs [64]. This metric evaluates codon representation in mRNA sequences and is beneficial for understanding how specific codons are favored or disfavored in a gene. DCBS values can be 1 or greater. Higher values indicate preferential use of certain synonymous codons rather than an equal distribution. We also quantified the deviation of observed codon frequencies from expected uniform distributions using weighted sum of relative entropy (E_W) [65]. E_W can be useful for identifying patterns of nonuniform codons in a gene, highlighting regions where certain codons are preferred or avoided. We showed that the E_W values obtained by the DL models reduced entropy significantly compared to other methods, demonstrating a preference for nonuniform codon usage (Fig. 2F).

Additionally, we evaluated the performance of the DL models by MCB [66]. MCB tests codon bias by calculating the probability of observing the specific frequencies of codons in a gene, considering variations in how often different codons are used and the overall base composition. High MCB values indicate specific codon preferences, while low values suggest a more even distribution of codon usage. We found that MCB values were four to five times higher in DL-optimized sequences than wild-type or GenScript-optimized sequences (Fig. 2G).

To further assess improvements in codon usage frequencies, we measured the FOP used as measured against an organism's tRNA availability [67, 68]; in this case, mouse tRNAs. Large FOP values indicate a higher usage of codons that are attributed to abundantly expressed tRNAs, which benefits translation efficiency and high protein expression levels. FOP values were highest for DL models, with the brain-optimized model achieving a mean FOP of 0.73 compared to 0.47 for wild-type and 0.44 for GenScript (Fig. 2H).

Another important aspect of codon usage for efficient protein translation are tandemly linked codons. It has been shown previously that codon optimization and protein translation can be substantially improved by considering optimization of codons as pairs [69]. One metric that evaluates this feature is known as the ENcp [70]. ENcp quantifies the bias in codon pair usage by comparing the observed frequency of codon pairs to the expected frequency by random pairing. A lower ENcp value indicates a higher bias toward specific codon pairs. ENcp values range from 20 (representing maximum bias) to 61 (indicating no bias of synonymous codon pair usage). ENcp values for DL models were significantly lower (~21–22), and close to the ideal value of 20 (Fig. 2I). In contrast, GenScript and URC sequences showed increased ENcp values. These results demonstrate improved codon pair bias by DL-based codon optimization with the potential of enhanced translational efficiency.

DL models reveal distinct codon optimization patterns across all tissues

Visualization with a t-distributed stochastic neighbor embedding (t-SNE) plot illustrates the clustering of sequences between the wild-type benchmark genes and their sequences after codon optimization (Fig. 2J). Of note, these genes were specifically selected to test the performance of the DL models

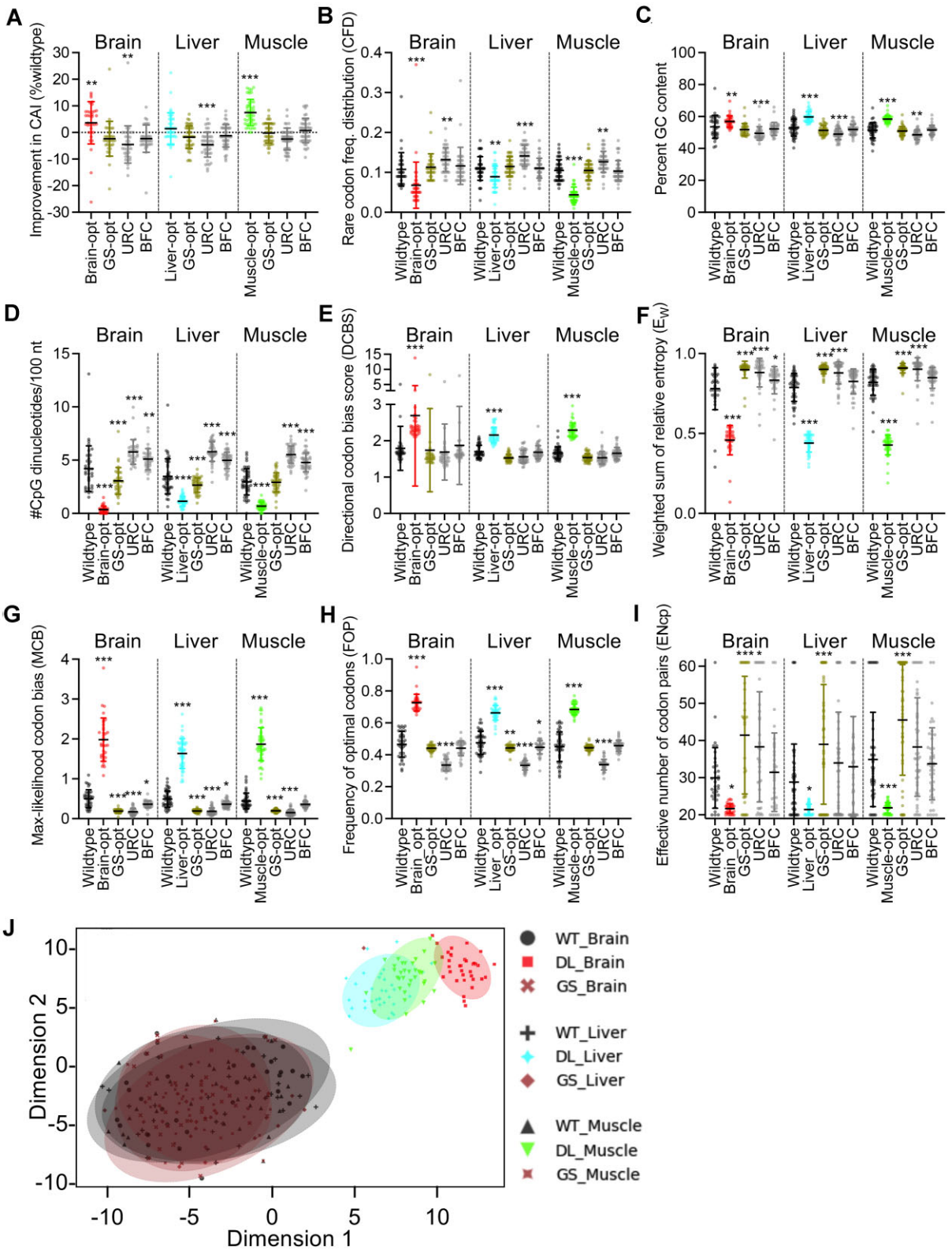


Figure 2. Performance evaluation of wild-type and codon-optimized sequences. Tissue-related genes were specifically selected for evaluating model performance (benchmark genes) and are exclusive from the training dataset. (**A–I**) Scatter plots of wild-type sequences; codon-optimized sequences based on DL models trained on brain transcripts (Brain-opt, red), liver transcripts (Liver-opt, cyan), and muscle transcripts (Muscle-opt, green); GenScript optimization (GS-opt, brown); URC; and BFC for the percentage improvement (+) or loss (-) in CAI scores (A), scatter plots of rare CFD (B), percent GC content for wild-type and codon-optimized sequences (C), the number of CpG dinucleotides (D), DCBS (E), weighted sum of relative entropy (E_W) (F), MCB scores (G), FOP (H), and ENcp (I). Mean values ± SD are also shown. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. (**J**) t-SNE visualization of sequences from different optimization strategies: tissue-dependent DL and GenScript optimized (GS). Wild-type sequences (WT), are also included. Axes labeled as "Dimension 1" and "Dimension 2," represent the two principal components resulting from the dimensionality reduction of high-dimensional codon usage data. These dimensions do not correspond to specific biological variables but are instead derived from the algorithm to best separate the data points in two dimensions.

for generating tissue-dependent codon optimization patterns and were not included in the training dataset, ensuring an unbiased evaluation of model performance (Supplementary File). Notably, DL model-based optimization resulted in distinct clustering, which suggests that our DL models employ a unique optimization strategy and is markedly different from traditional methods like those used by GenScript. This distinct clustering implies that the DL model may be capturing specific codon usage patterns tailored to each tissue, potentially leading to more efficient protein expression or increased stability. The observed separation in the t-SNE plot highlights the ability of our DL models to specify codon optimization rules while considering tissue-dependent nuances.

DL models identify G/C-ending codon biases across tissue types

As discussed before, the selected tissue-related gene sets exhibited a prominent preference for G/C-ending codons across the three tissue types (Fig. 1C). We wanted to evaluate if a similar codon usage preference existed for the benchmark gene sequences after DL optimization. It is worth noting that the wild-type sequences for the training and benchmark genes used all 61 codons (Figs. 1C and 3A–C). We observed that the GenScript-optimized benchmark genes continued to use all 61 available codons. In contrast, sequences predicted by our DL models showed an overall decrease in the number of codons used. To elaborate, the brain- and liver-optimized DL models used only 45 codons, while the muscle-optimized DL model used only 44 codons. The decreased preference for codon usage is not equally represented by A/T- and G/C-ending codons. Excluding the termination codons, there are a total of 31 G/C-ending and 30 A/T-ending codons. While most G/C-ending codons were retained and used at a higher mean frequency, this was not the case for A/T-ending codons (Fig. 3A–C). We observe that the preferred codon for all degenerate amino acids ended in G/C. A key finding from this analysis showed that our DL models tended to avoid certain codons entirely. For example, leucine is known to have six degenerate codons, four of which were completely excluded from sequences predicted by our DL models. The remaining two codons were CTC and CTG, both of which end in G/C.

Valine also showed a similar trend, where only two of its four codons were preferred by our DL models. These two codons end in G/C while the unused two end in A/T. This result indicates a novel and previously unknown pattern learned by our DL models from highly expressed transcripts.

Evaluation of DL models using reporter genes

To test the robustness of our DL-based models, we subjected two standard cellular reporter genes that are frequently used in basic and preclinical research to codon optimization: *FLuc* and *Egfp*. These reporter genes were subjected to the three tissue-trained DL models to produce codon-optimized sequences. The GenScript-optimized versions of both reporter genes were also generated for comparison. As mentioned above, we observed that the GenScript codon optimization tool produces multiple "optimized" sequences for the same gene. Thus, using GenScript's optimizer tool, we generated two codon-optimized sequences for each of the said reporter genes and used them as controls to evaluate translation efficiency as compared to wild-type sequences and those generated by our DL models.

To validate our approach, we first ascertained the similarities between the wild-type (or original) sequences and the different optimized versions (Fig. 4A and B, and Supplementary Figs. S4 and S5). We found that the *FLuc* sequences predicted using our DL models and those generated by GenScript optimization were highly divergent from wild-type *FLuc*, with only 75%–77% similarity (Fig. 4A and Supplementary Fig. S4). In contrast, for *Egfp*, the DL-optimized sequences showed 87% similarity to the original sequence, while the GenScript-optimized *Egfp* sequences displayed only 80% similarity (Fig. 4B and Supplementary Fig. S5). The brain-, liver-, and muscle-optimized sequences for both *FLuc* and *Egfp* genes had similarity values exceeding 92%, when compared with each other. Interestingly, the two GenScript-optimized *FLuc* sequences only shared a similarity of 77% (Fig. 4A). Similarly, the two GenScript-optimized *Egfp* sequences had a similarity of 82% (Fig. 4B). These results demonstrate the variability of output sequences by the GenScript tool.

The CAI values for the DL-optimized *FLuc* sequences were substantially greater than those of the wild-type and GenScript-optimized sequences (Fig. 4C). Similarly, their GC contents were also higher. For *Egfp*, the DL models produced sequences with CAI values that were similar to the original sequence (Fig. 4D). Importantly, our DL models generated sequences with superior CAIs compared to the GenScript-optimized sequences. The GC contents of the DL-optimized *Egfp* sequences were slightly lower than the original sequence and marginally higher than those of the GenScript-optimized sequences (Fig. 4D). The GC contents for all of the optimized *FLuc* and *Egfp* sequences were within the desired 30%–70% range. As mentioned previously, a surprising finding was that the DL-trained models produced sequences with a marked decrease in CpG dinucleotide content when tested on the benchmark genes (Fig. 2D). We found that the DL-optimized versions of the *FLuc* and *Egfp* reporter genes also exhibited a substantial reduction in CpG dinucleotides as compared to their wild-type/original counterparts (Fig. 4C and D). Notably, our algorithms were able to better decrease the number of CpG dinucleotides than the GenScript codon-optimization tool was able to achieve. The reduction in *FLuc* CpGs ranged from ∼7- to 16-fold, while the reduction for *Egfp*, was between 15- and 60-fold (Fig. 4C and D). Despite causing only marginal changes to the overall GC content of the reporter genes, DL optimization resulted in a substantial reduction in CpG dinucleotides.

We next performed a comparison of codon preferences among the *FLuc* and *Egfp* sequences generated by our DL models, the wild-type/original sequences, and the GenScript-optimized sequences (Fig. 4E and F). The most striking observation was that the GenScript-optimized sequences showed less bias for specific codons for many amino acids. For example, in the DL-optimized sequences, the amino acids leucine, lysine, and valine, which are typically represented by 4–6 codons, are dominated by a single codon (CTG, AAG, and GTG, respectively), while the GenScript-optimized sequences showed a more evenly distributed usage of the codons encoding for these three amino acids. There were also codons that were consistently favored across all DL models. This observation was similar to what was documented for the benchmark genes (Fig. 3A–C). Moreover, our analysis also revealed differences in codon preference for certain amino acids among the three DL models, supporting the notion that codon usage biases may exist between tissues. For instance, there was a

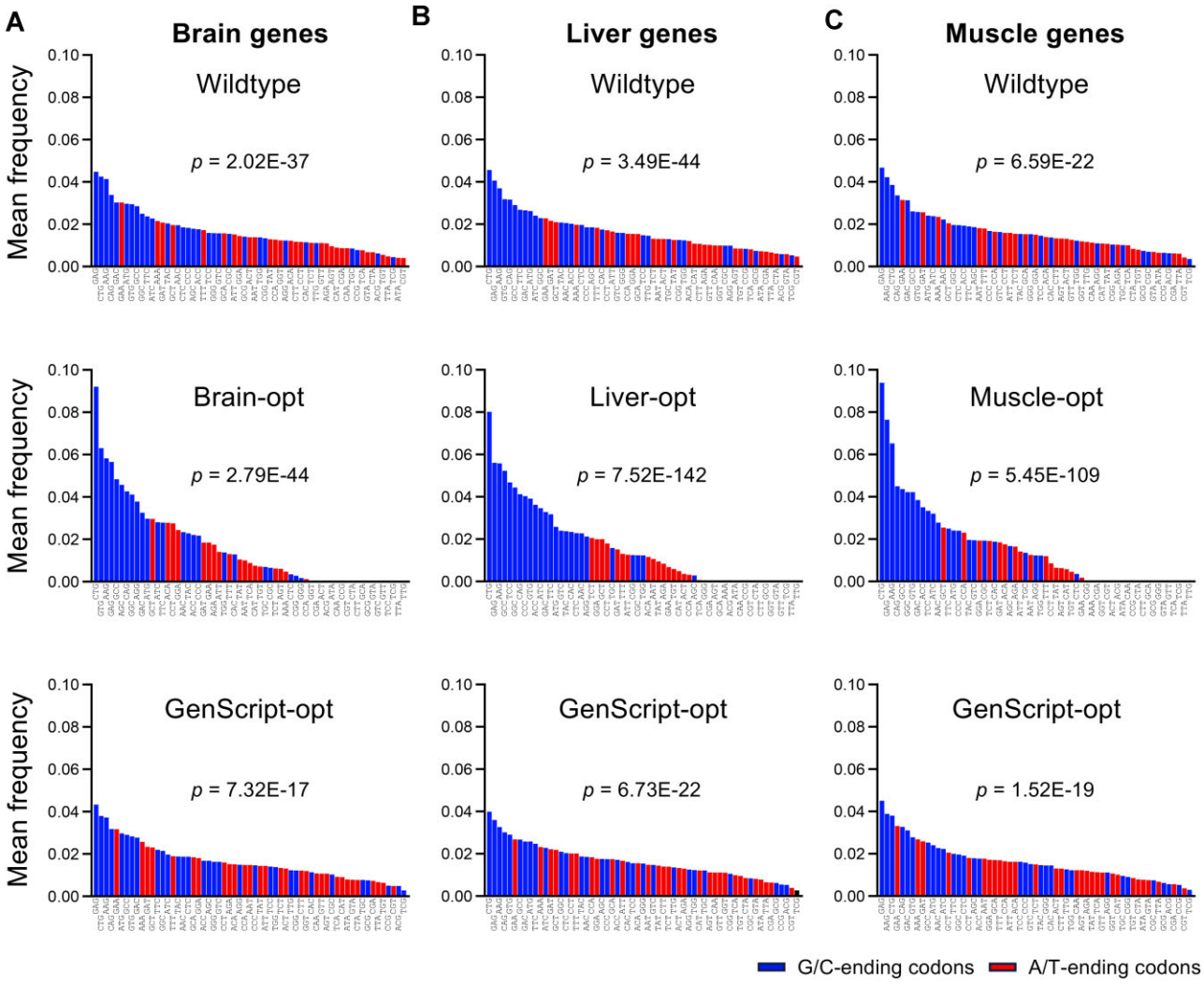


Figure 3. Comparative analysis of G/C- or A/T-ending codons used between wild-type and codon-optimized benchmark gene sequences. Waterfall plots of the mean frequencies of G/C-ending codons (blue bars) and A/T-ending codons (red bars) among the brain (**A**), liver (**B**), and muscle (**C**) benchmark gene sequences obtained before and after codon-optimization using tissue-dependent DL models or GenScript tool. *P*-values were determined by Wilcoxon rank sum tests.

stronger preference for ACA to encode threonine in the brain-trained model, while ACC was the codon of preference in liver- and muscle-trained DL models (Fig. 4E and F). This finding indicates that not only can the DL models optimize for codon usage but can also identify tissue-related codon preferences.

Evaluation of DL-optimized *rep*orter genes in mouse cell lines

The elevated CAI values and increased GC content of the DL-optimized *Egfp* and *FLuc* genes suggested that these sequences should be expressed at higher levels. We therefore set out to test whether the optimizations conferred tissue-dependent improvements to protein expression, since the DL-optimizations were trained on transcripts that were highly expressed in three separate tissue types (brain, liver, and muscle). Constructs expressing wild-type *FLuc* or its various codon-optimized versions were transfected into mouse cell lines representing brain (Neuro-2a, neuroblasts), liver (AML12, hepatocytes), and muscle (C2C12, myoblasts) tissues. Relative luciferase activities conferred by the brain-optimized sequence were >3-fold higher than those achieved by the original sequence in neuroblasts (Fig. 5A); the liver-optimized sequence exhibited 8-fold higher activity in hepatocytes (Fig. 5B); and muscle-optimized sequence displayed a 7-fold higher activity in undifferentiated myoblasts (Fig. 5C), and in differentiated myotubes (Fig. 5D). Interestingly, while all DL-optimized sequences displayed increased expression as compared to wild-type; the liver-optimized sequence produced the highest increase in luciferase expression across all cell types. The muscle-optimized sequence also showed an ∼5-fold increase in Neuro-2a and AML12 cells (Fig. 5A and B). Strikingly, the GenScript-optimized sequence did not increase expression over the original sequence in any cell lines (Fig. 5A–D).

We also performed transfections of DL-optimized *Egfp* constructs in Neuro2a, AML12, and C2C12 cells, and analyzed EGFP expression by epifluorescence and flow cytometry (Fig. 5E–H). It should be noted that the transfection efficiencies were high in Neuro-2A cells (60%–70%) and lower in AML12 and C2C12 cells (10%–20%) (Fig. 5E and Supplementary Fig. S6A–C). The brain-optimized *Egfp*

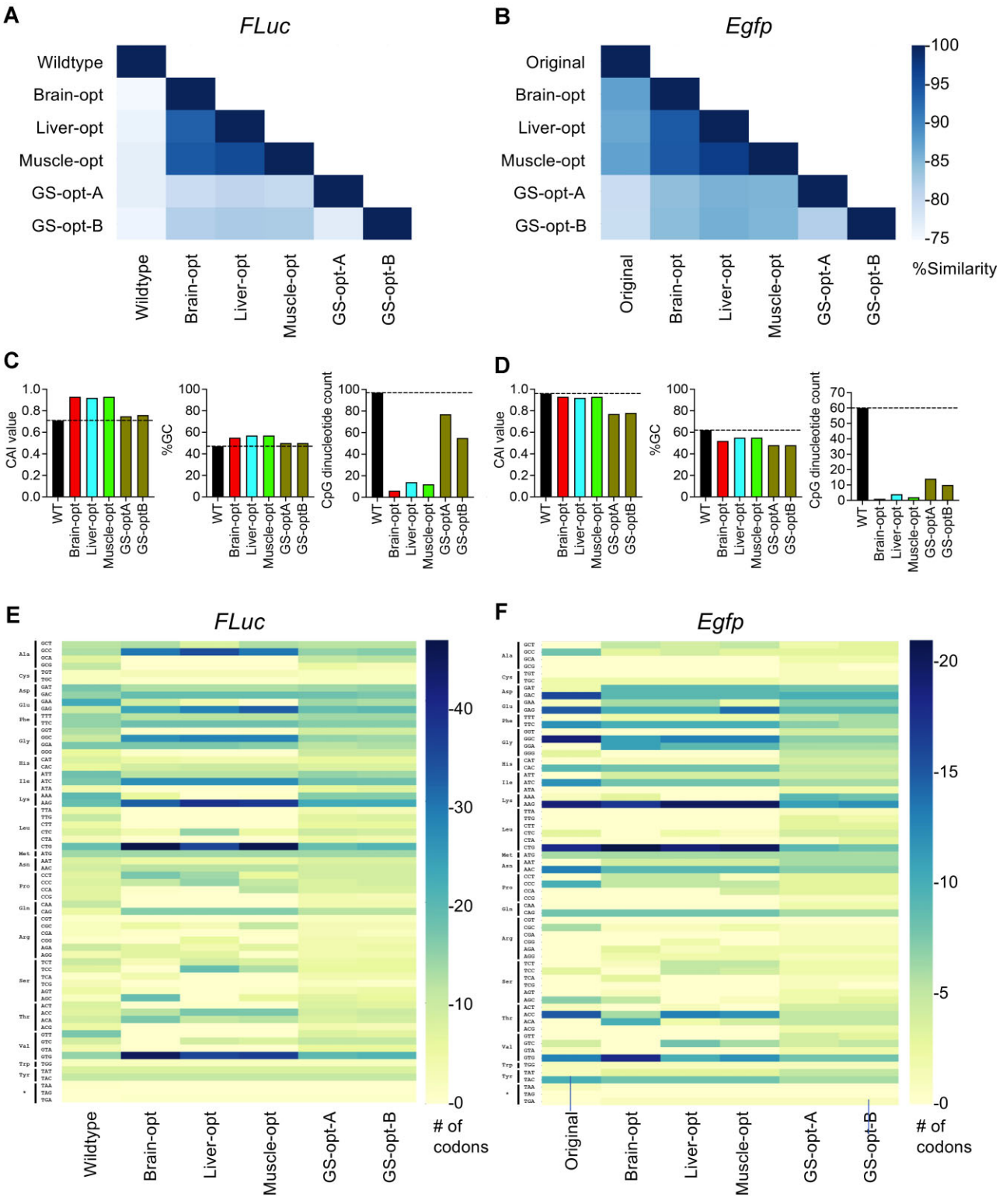


Figure 4. Comparative analyses of reporter gene sequences following codon optimization. (**A** and **B**) Heatmap display of DNA sequence similarity between codon-optimized *Egfp* (A) or *FLuc* (B) reporter gene sequences through DL models, wild-type/original sequences, and codon-optimized by the GenScript (GS) tool. (**C** and **D**) Histograms of CAI values, GC content, and CpG dinucleotides for codon-optimized *FLuc* (C) or *Egfp* (D) sequences. (**E** and **F**) Heatmap of codon preferences between optimized *FLuc* (E) or *Egfp* (F).

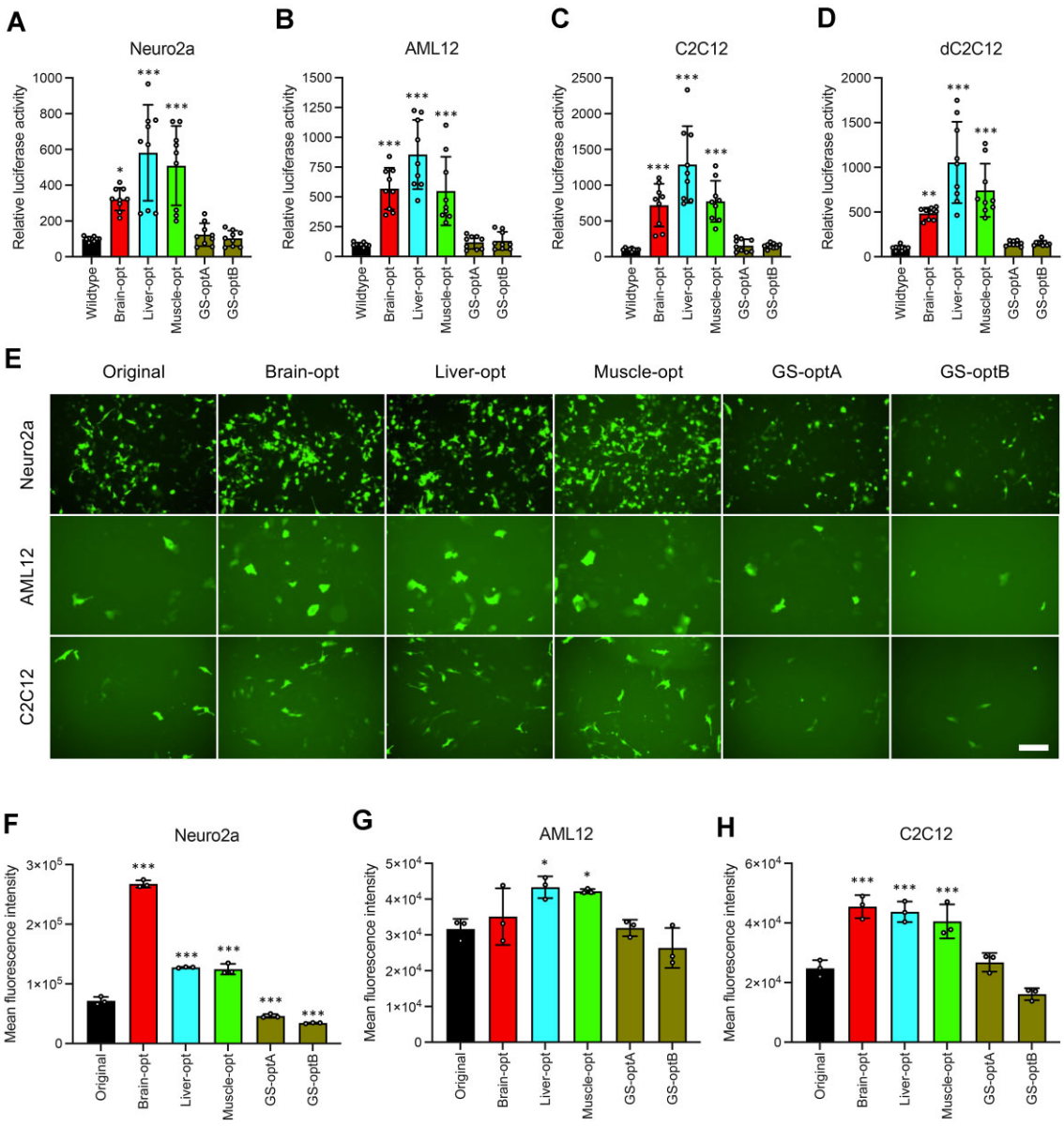


Figure 5. Quantification of codon-optimized reporter genes. (**A–D**) Histograms showing relative luciferase activity from Neuro-2a (neuroblasts) (A), AML12 (hepatocytes) (B), C2C12 (myoblasts) (C), and differentiated C2C12 (dC2C12, myotubes) (D) transfected with the wild-type *FLuc* construct, codon-optimized constructs obtained from DL models trained with brain, liver, and muscle genes, or optimized by the GenScript (GS) algorithm. Values represent normalized luciferase activities scaled to wild-type levels set to 100. (**E**) Representative epifluorescence microscopy images of transfected cells (scale bar: 10 μm). Neuro-2a (top row), AML12 (middle row), or C2C12 (bottom row) cells were transfected with the original *Egfp* construct, DL-based codon-optimized constructs, or optimized by the GenScript algorithm. (**F–H**) Cells were subjected to flow cytometry analyses and evaluated for MFI. Histograms showing the detection of EGFP fluorescence for Neuro2a (F), AML12 (G), and C2C12 (H). $n = 3$; >10 000 events counted. All histograms show mean values ± SD; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

construct conferred the highest level of fluorescence in neuroblasts (Fig. 5E and F). Similarly, the liver-optimized *Egfp* construct showed the highest expression in AML12 hepatocytes (Fig. 5E and G). In C2C12 cells, all versions of the DL-optimized *Egfp* constructs performed comparably to each other but were significantly higher than the original sequence or GenScript-optimized *Egfp* versions (Fig. 5E and H). Flow cytometry analyses showed that in neuroblasts, the fluorescence intensity from cells transfected with the brain-optimized *Egfp* construct was >3.5-fold greater than cells transfected with the original *Egfp* sequence and 6-fold greater than what was achieved by either of the GenScript-optimized sequences (Fig. 5F). Similarly, the expression conferred by the liver-optimized *Egfp* sequence in AML12s was higher than the expression levels generated by the original or GenScript-optimized sequences (Fig. 5G). Quantification from biological replicates revealed a 30% increase in expression by liver-optimized *Egfp* over that of the original *Egfp* sequence (Fig. 5G). In the case of C2C12 myoblasts, expression of the muscle-optimized *Egfp* construct was comparable among all of the DL-optimized constructs but ∼ 2.5-fold higher than the original or GenScript-optimized sequences (Fig. 5H). The overall percentage of GFP+ cells were not drastically affected in any of the transfection groups, except in one condition (brain-optimized and Supplementary Fig. S6C), indicating that the increases in fluorescence was largely due to

higher amounts of EGFP protein in the transfected cells and not necessarily due to differences in transfection efficiencies across test groups. A higher percentage of GFP+ cells in the one outlier condition for C2C12 cells transfected with the brain-opt EGFP did not translate to higher MFI values and may be attributed to higher viability among those samples (Supplementary Fig. S6C).

Model training using secretome genes

One useful application of our methodology is the optimization of genes encoding secreted proteins. Since secreted proteins do not necessarily require cell-type specific targeting, like blood clotting factors (factor VIII or factor IX), insulin, and α-antitrypsin 1, many therapeutic approaches can enlist organs or cell types that are easily transduced by gene therapy vehicles [71–74]. We therefore aimed to determine whether generating a DL-codon optimization model to specifically improve codon usage for secreted protein products would yield better expression. In order to achieve this goal, we trained an algorithm using curated gene lists from The Human Protein Atlas [75–77], incorporating data from the human secretome (Supplementary File). Orthologous mouse genes were also identified, and their extracellular localization was confirmed using the Uniprot database [78]. A total of 1555 human and 1319 mouse genes were selected for DL model training and validation as described before. The gene lists were randomly split into training, testing, and validation lists. Standard loss and accuracy metrics for the training process were generated (Supplementary Fig. S7). This evaluation showed that the human and mouse secretome-defined models achieved an accuracy of 0.5523 and 0.5464, respectively. The human model achieved a training loss of 0.0803 (14.75% improvement, TMA = 0.4813), while the mouse model reached a training loss of 0.0893 (15.91% improvement, TMA = 0.4714), indicating that the models are learning fine nuanced features during training.

Evaluation of codon optimization indices generated by human and mouse secretome DL models

We evaluated the human and mouse secretome DL models on a set of benchmark genes (Supplementary File), comparing wild-type sequences and those optimized by URC, BFC, and GenScript. Both models significantly improved CAI values (17% and 19%, respectively) compared to wild-type. Other models, including GenScript's, showed no improvement or loss in CAI (Fig. 6A). Similarly, CFD values were significantly reduced in sequences obtained from our DL models, while others showed no reduction or increase (Fig. 6B). Our models slightly increased %GC content (Fig. 6C). Notably, CpG dinucleotide frequencies significantly decreased in DL-optimized sequences. In contrast, the URC and BFC model, yielded sequences with increased CpG content, and the GenScript model yielded no significant change in mean values (Fig. 6D). These outcomes reflect our models' ability to mimic natural transcripts by limiting CpG content. Other metrics also confirmed superior optimization by the DL models. DCBS values were enhanced and E_W scores reduced, indicating improved codon usage patterns (Fig. 6E and F). The MCB and FOP metrics were significantly higher for DL models, demonstrating better codon adaptation (Fig. 6G and H). ENcp values for DL models indicated greater codon pair specificity, while GenScript and URC sequences showed less optimal codon pair usage (Fig. 6I). These results highlight the superior performance of our DL models in optimizing codon usage for human and mouse secretome genes.

The DL-optimized sequences also formed a separate cluster away from the wild-type and GenScript-optimized sequences by t-SNE plots (Fig. 6J). This result indicated that DL optimization effectively captured codon usage patterns that were unique from those generated by the GenScript tool and those found in the wild-type sequences.

Genes encoding secreted proteins from the human and mouse also displayed a bias toward G/C-ending codons (Supplementary Fig. S8A and B). Resulting sequences obtained from the DL models showed an even stronger bias, while those obtained from the GenScript tool showed a drop in G/C-ending codon bias, as indicated by the increase in *P*-values (Supplementary Fig. S8C and D). We also observed an overall increase in mean frequencies of G/C-ending codons for the human and mouse DL models. The overall number of codons used in human and mouse DL model-predicted sequences decreased to 45 and 44, respectively, while the wild-type and GenScript-generated sequences continued to utilize all 61 codons (Supplementary Fig. S8C and D).

Model evaluation using a secreted reporter gene

In order to test the functionality of our secretome DL models, we selected to use the secreted reporter *Gaussia* luciferase (*GLuc*). We generated wild-type and codon-optimized sequences for *GLuc* using our human and mouse secretome DL models. Codon-optimized sequences generated from GenScript's tool were again used for comparative evaluation. To test the secretome-optimized *GLuc* sequences, we first compared the similarities between the original sequences with the optimized versions (Supplementary Figs S9A and S10). The DL models produced *GLuc* sequences with 87%–89% similarity to the wild-type sequence, while the GenScript-optimized version showed ∼80% similarity. The CAI values for the DL-optimized *GLuc* sequences were higher than those of the original and GenScript-optimized sequences (Supplementary Fig. S9B). The GC contents of the DL-optimized sequences were also higher, but they exhibited a significant reduction in CpG dinucleotides compared to the wild-type sequence (Supplementary Fig. S9B). In contrast, the GenScript-optimized *GLuc* sequence showed an ∼2-fold increase in CpG dinucleotide content (Supplementary Fig. S9B).

We next performed a codon preference analysis on the *GLuc* sequences (Supplementary Fig. S9C). The DL-optimized sequences showed strong biases for specific codons, whereas the GenScript-optimized sequences displayed a more generalized usage of synonymous codons. We also observed that the two secretory DL models (human versus mouse) displayed distinct codon usage for certain amino acids such as glutamic acid (Glu) and isoleucine (Ile). This result shows that the DL-based models were able to learn differences from the mRNA sequences encoding orthologous genes from two divergent species.

In vitro evaluation of DL-optimized secretory reporter gene

Mouse and human cell lines were transfected to evaluate the performance of wild-type and various codon-optimized versions of the *GLuc* transgene. We also generated codon-optimized *GLuc* sequences using our tissue-dependent DL

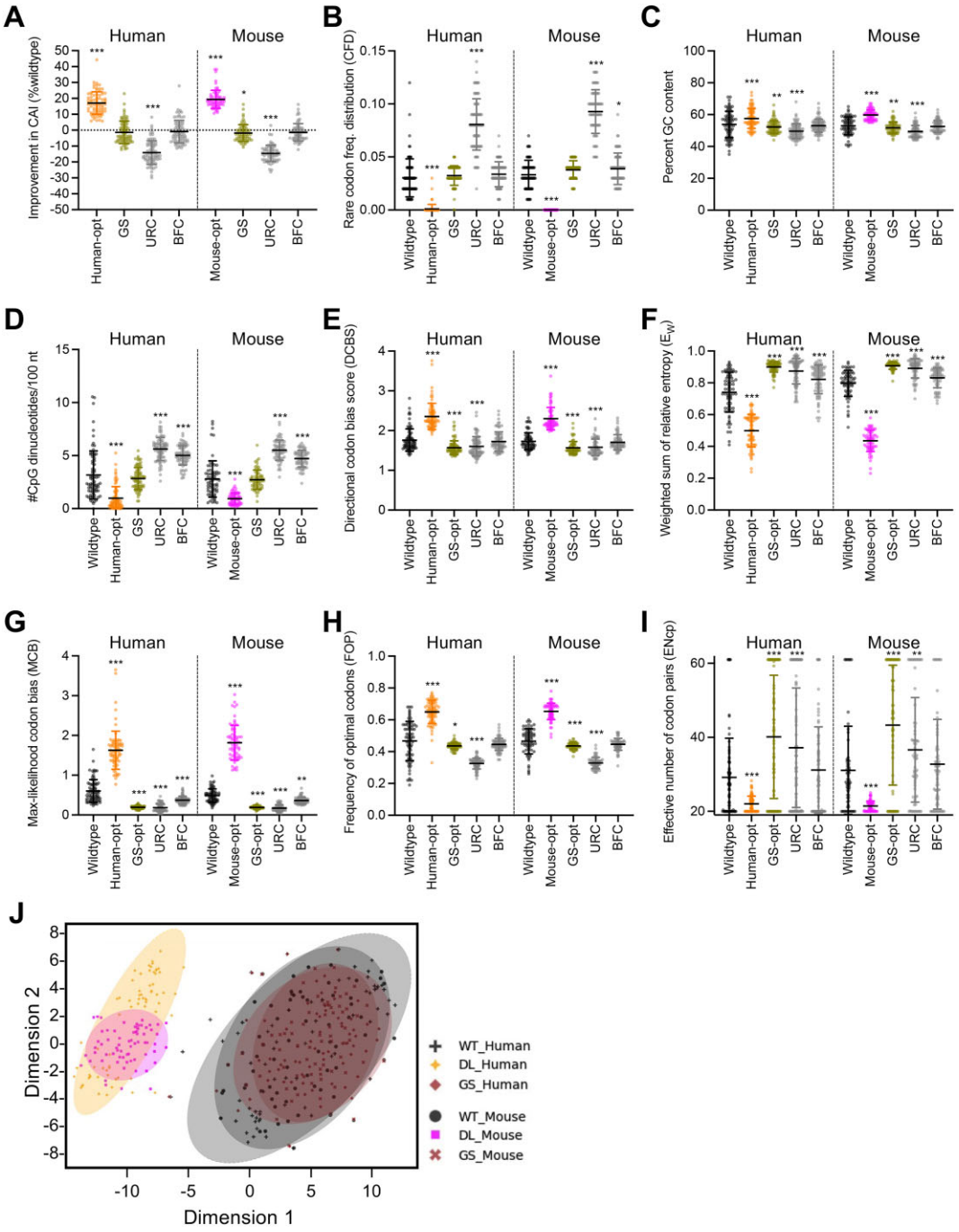


Figure 6. Performance evaluation of wild-type and codon-optimized genes encoding secreted proteins. Gene sequences representing secretory proteins were specifically selected for evaluating model performance and are exclusive from the training dataset. (**A–J**) Scatter plots of wild-type sequences; codon-optimized sequences based on DL models trained on human (orange) and mouse (magenta) genes encoding secreted proteins; GenScript optimization (GS-opt, brown); URC; and BFC for the percentage improvement (+) or loss (−) in CAI scores (A), scatter plots of rare CFD (B), %GC content for wild-type and codon-optimized sequences (C)**,** the number of CpG dinucleotides (D), DCBSs (E), weighted sum of relative entropy (E_W) (F), MCB scores (G), FOP (H), and ENcp (I). Mean values ± SD are also shown. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. (J) t-SNE visualization of sequences from different optimization strategies: DL models trained on human and mouse secretome genes (DL) and GenScript optimized (GS). Wild-type sequences (WT) are also included. Axes labeled as "Dimension 1" and "Dimension 2" represent the two principal components resulting from the dimensionality reduction of high-dimensional codon usage data. These dimensions do not correspond to specific biological variables but are instead derived from the algorithm to best separate the data points in two dimensions.

models to determine whether codon optimization based on the secretome versus tissue enrichment yields differences in protein expression. Dual-luciferase assays across multiple cell lines show that sequences generated by both general and secretome-specific DL models were able to achieve significant improvement in GLuc expression as compared to the wild-type construct (Supplementary Fig. S9D). Interestingly, we also observed improvements from constructs encoding *GLuc* sequences optimized by GenScript in HeLa cells. However, the performance of human and mouse secretome-optimized, and brain- and liver-optimized *GLuc* versions improved across several cell lines (Supplementary Fig. S9D). In fact, mouse secretome- and liver-optimized *GLuc* showed a 30- and 40-fold increase in GLuc expression in HeLa cells as compared to their wild-type counterpart. All optimized *GLuc* sequences obtained by DL models, with the exception of the muscle-optimized model showed expression levels that were higher than those of the GenScript-optimized sequence, which again demonstrates that DL-based algorithms trained on mRNA expression data alone are capable of generating codon-optimized transgenes with increased protein expression.

Discussion

Deciphering codon usage patterns holds immense value. Precisely modulating codon usage facilitates improved protein translation rates, minimizes the occurrence of translation errors, and maximizes the expression of therapeutic genes. In this study, DL models were formulated to examine transcripts that are highly expressed in different tissues. The strategy was specifically chosen based on the hypothesis that the highest expressing transcripts are evolutionarily optimized for protein translation, and thus, would select the most optimal codons. The aim was to decipher complex sequential patterns hidden in highly expressed transcripts across three tissue types: brain, liver, and muscle. Unlike traditional algorithms that explicitly utilize known parameters, DL can acquire knowledge implicitly through the analysis of sequence patterns within a training dataset to harness hidden biological phenomena. This proof-of-concept evaluation encompassed sufficient data and an effective architecture of neural layers. It is worth noting that other research groups have employed neural networks to understand and forecast codon usage in relation to distributions of ribosome density or protein expression [37, 79]. However, the availability for this type of data is scant for diverse tissue types, whereas whole transcriptome data are widely available.

The sequences optimized using DL models displayed significant differences from sequences produced by GenScript's codon optimization algorithm, which served as a proxy for commercially available codon optimization tools. There are several that are available, some of which produce multiple sequences when prompted to optimize the same sequence. We observed that the codon distributions were more dispersed with the GenScript algorithm, which may partially explain the variety in sequence outputs that are generated from a single gene. Nevertheless, the substantial enhancements in CAI values suggest at the effectiveness of our approach. By surpassing the performance of a commercially available algorithm, our study highlights the potential of DL models to enhance gene expression and optimize codon usage; thus, contributing to the advancement of gene design and synthetic biology applications.

There are other DL methodologies that have resulted in models that can encapsulate the intricate sequential arrangements found within DNA sequences. Nevertheless, the distinguishing characteristic of our DL methodology implemented here lies in its unsupervised framework. These models do not rely on prior encoding of sequences into codon boxes or specifying a particular range for pattern observations [79–81]. Instead, it gains implicit understanding from sequence patterns spread across entire gene lists, which can be performed on genes that are expressed in specified tissues.

In this study, we explicitly restricted our study to training our DL models on highly expressed genes of *Mus musculus*. However, the DL methodology employed here can be extended to other species, organs, tissues, and cell types. One of our goals was to determine whether training a DL model on different tissues or cell types could reveal codon usage rules that were optimized for corresponding tissues. However, we observed that the liver-optimized *FLuc* sequence turned out to be the most robust, even in neuronal and muscle cells (Fig. 5A–D). The robustness of the liver-optimized model could be attributed to certain genes. In the case of *Egfp*, the optimizations did not result in consistently high expression across all cell types. Nevertheless, our findings hint at a methodology for developing superior codon-optimized genes that are far superior to conventional tools. Furthermore, the study hints at tissue differences in synonymous codon biases. Still, this work did not support our hypothesis that the DL models could design codon-optimized sequences that are tissue-dependent. The reason for these results is unclear. One possible explanation is that there is a potential overlap in codon-usage rules for proteins expressed in the different tissues. The differences and overlaps in training gene sets (Fig. 1A) and their gene ontologies (Supplementary Fig. S1A) may also confer overlapping tissue-independent expression in nonrepresentative cell lines. Further investigation of these features is warranted in order to explain differences in tissue-specific codon usage rules. Exploration into additional tissue types may reveal further distinctions and unknowns in mechanisms that underpin efficient protein translation.

Some limitations to our work are related to the relatively few genes used for training. The training algorithm is likely to be influenced by the cutoff thresholds applied to the gene lists during training. In addition, this investigation only used a single-split method (random selection of genes used for training, evaluation, and testing at a relative 85:10:5 split) [37, 38]. Given the low number of training genes used, this approach will undoubtedly lead to inherent biases that are defined by the selection of the random genes. We did not perform multiple independent training runs with different random splits. Incorporating K-fold cross-validation and hyperparameter tuning would also lead to improved robustness and confidence of the approach. Nevertheless, each resulting model trained on three different datasets defined by brain, liver, and muscle transcripts yielded better expression than the original sequences and those generated by publicly available codon-optimization tools. This achievement gives us confidence that our approach can be repeated with improved outcomes.

An unintended, but important, finding was that our DL models were able to improve CAI values, while reducing CpG dinucleotides. This finding is significant because unmethylated CpG dinucleotides can trigger innate immune responses in the host cell. This is a critical consideration in DNA-based gene therapy platforms. The ability to reduce immunogenic-

ity, while maintaining optimal codon usage demonstrates the superior performance of the DL models compared to commercial tools. Therefore, our study provides an invaluable tool for researchers in the gene therapy field, offering essential guidance for the design and development of optimized gene delivery systems and therapeutic interventions.

Our methodology can also be potentially applied to improving transgene cassettes with specific cellular functions. We have shown that DL models trained on a smaller subset of transcripts that encode secreted proteins were able to identify subtle and nuanced patterns in their cDNAs. Our analysis also revealed differences in orthologous genes between human and mouse, supporting the notion of species-specific codon usage preferences. These findings highlight the potential for using DL approaches to help design optimal transgene cassettes for explicit cellular functions and can potentially offer insights into the basic biology of protein translation and mRNA stability.

Acknowledgements

We would like to extend our thanks to Dan Wang for thoughtful discussions and support. We would also like to thank the UMass Chan Viral Vector Core for reagent support.

Author contributions: S.R., T.S., J.X., G.G., and P.W.L.T. conceived the study, designed experiments, and interpreted the data. S.R. executed and analyzed all of the AI/DL-related experiments. T.S. and M.Y. performed all of the construct cloning and cell culture experiments, and analyzed the data. H.Y. generated the liver RNA-seq data reported in the study. G.G. and P.W.L.T. supervised the overall study and secured funding for the project. S.R. and T.S. wrote the initial draft of the manuscript. S.R., T.S., G.G., and P.W.L.T. revised and finalized the manuscript.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

G.G. is a scientific co-founder of Voyager Therapeutics and Aspa Therapeutics and holds equity in these companies. G.G. and P.W.L.T. are inventors on patents with royalties licensed to biopharmaceutical companies. The remaining authors declare no competing interests.

Funding

This work was supported by grants from the UMass Chan Medical School (an internal grant) and by the National Institutes of Health (R01NS076991-01, P01HL131471-05, R01AI121135, UG3HL147367-01, R01HL097088, R01HL152723-02, U19AI149646-01, and UH3HL147367-04) to GG. P.W.L.T. is supported by the National Institutes of Health (1R21AI183080-01A1), The Bassick Family Foundation, and a BRIDGE Fund Award (a UMass Chan Medical School internal grant). Funding to pay the Open Access publication charges for this article was provided by the NIH.

Data availability

All custom codes and scripts are hosted on Zenodo (https://doi.org/10.5281/zenodo.14991160).

References

1. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 1985;**2**:13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335
2. Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res* 2014;**42**:9171–81. https://doi.org/10.1093/nar/gku646
3. Quax TE, Claassens NJ, Soll D *et al.* Codon bias as a means to fine-tune gene expression. *Mol Cell* 2015;**59**:149–61. https://doi.org/10.1016/j.molcel.2015.05.035
4. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* 2018;**19**:20–30. https://doi.org/10.1038/nrm.2017.91
5. Gustafsson C, Minshull J, Govindarajan S *et al.* Engineering genes for predictable protein expression. *Protein Expr Purif* 2012;**83**:37–46. https://doi.org/10.1016/j.pep.2012.02.013
6. Singh R, Sophiarani Y. A report on DNA sequence determinants in gene expression. *Bioinformation* 2020;**16**:422–31. https://doi.org/10.6026/97320630016422
7. Lipinszki Z, Vernyik V, Farago N *et al.* Enhancing the translational capacity of *E. coli* by resolving the codon bias. *ACS Synth Biol* 2018;**7**:2656–64. https://doi.org/10.1021/acssynbio.8b00332
8. Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. *Trends Biotechnol* 2004;**22**:346–53. https://doi.org/10.1016/j.tibtech.2004.04.006
9. Zhou Z, Dang Y, Zhou M *et al.* Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci USA* 2016;**113**:E6117–25. https://doi.org/10.1073/pnas.1606724113
10. Chaney JL, Steele A, Carmichael R *et al.* Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput Biol* 2017;**13**:e1005531. https://doi.org/10.1371/journal.pcbi.1005531
11. Goncalves GAR, Paiva RMA. Gene therapy: advances, challenges and perspectives. *Einstein (São Paulo)* 2017;**15**:369–75. https://doi.org/10.1590/s1679-45082017rb4024
12. Benham AM. Protein secretion and the endoplasmic reticulum. *Cold Spring Harb Perspect Biol* 2012;**4**:a012872. https://doi.org/10.1101/cshperspect.a012872
13. Barlowe CK, Miller EA. Secretory protein biogenesis and traffic in the early secretory pathway. *Genetics* 2013;**193**:383–410. https://doi.org/10.1534/genetics.112.142810
14. Crowley KS, Liao S, Worrell VE *et al.* Secretory proteins move through the endoplasmic reticulum membrane via an aqueous, gated pore. *Cell* 1994;**78**:461–71. https://doi.org/10.1016/0092-8674(94)90424-3
15. Karamyshev AL, Tikhonova EB, Karamysheva ZN. Translational control of secretory proteins in health and disease. *Int J Mol Sci* 2020;**21**:2538. https://doi.org/10.3390/ijms21072538
16. Hamilton BA, Wright JF. Challenges posed by immune responses to AAV vectors: addressing root causes. *Front Immunol* 2021;**12**:675897. https://doi.org/10.3389/fimmu.2021.675897
17. Li J, Fu L, Wang G *et al.* Unmethylated CpG motif-containing genomic DNA fragment of Bacillus calmette-guerin promotes macrophage functions through TLR9-mediated activation of NF-kappaB and MAPKs signaling pathways. *Innate Immun* 2020;**26**:183–203. https://doi.org/10.1177/1753425919879997
18. Zhou J, Deng GM. The role of bacterial DNA containing CpG motifs in diseases. *J Leukoc Biol* 2021;**109**:991–8. https://doi.org/10.1002/JLB.3MR1220-748RRRRR
19. Tang B, Pan Z, Yin K *et al.* Recent advances of deep learning in bioinformatics and computational biology. *Front Genet* 2019;**10**:214. https://doi.org/10.3389/fgene.2019.00214

20. Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun Signal* 2020;**18**:145. https://doi.org/10.1186/s12964-020-00642-6
21. Emmert-Streib F, Yang Z, Feng H *et al.* An introductory review of deep learning for prediction models with big data. *Front Artif Intell* 2020;**3**:4. https://doi.org/10.3389/frai.2020.00004
22. Harvey TJ, Davila RA, Vidovic D *et al.* Genome-wide transcriptomic analysis of the forebrain of postnatal Slc13a4(±) mice. *BMC Res Notes* 2021;**14**:269. https://doi.org/10.1186/s13104-021-05687-5
23. Karaayvaz M, Cristea S, Gillespie SM *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat Commun* 2018;**9**:3588. https://doi.org/10.1038/s41467-018-06052-0
24. Zhang B, Zhao C, Shen W *et al.* KDM2B regulates hippocampal morphogenesis by transcriptionally silencing Wnt signaling in neural progenitors. *Nat Commun* 2023;**14**:6489. https://doi.org/10.1038/s41467-023-42322-2
25. Mencio CP, Tilve SM, Suzuki M *et al.* A novel cytoskeletal action of xylosides. *PLoS One* 2022;**17**:e0269972. https://doi.org/10.1371/journal.pone.0269972
26. Bennett H, Troutman TD, Zhou E *et al.* Discrimination of cell-intrinsic and environment-dependent effects of natural genetic variation on Kupffer cell epigenomes and transcriptomes. *Nat Immunol* 2023;**24**:1825–38. https://doi.org/10.1038/s41590-023-01631-w
27. Yang H, Brown RH Jr, Wang D *et al.* Rescue of GM3 synthase deficiency by spatially controlled, rAAV-mediated ST3GAL5 delivery. *JCI Insight* 2023;**8**:e168688. https://doi.org/10.1172/jci.insight.168688
28. Perez VM, Gabell J, Behrens M *et al.* Deletion of fatty acid transport protein 2 (FATP2) in the mouse liver changes the metabolic landscape by increasing the expression of PPARalpha-regulated genes. *J Biol Chem* 2020;**295**:5737–50. https://doi.org/10.1074/jbc.RA120.012730
29. Slabber CF, Bachofner M, Speicher T *et al.* The ubiquitin ligase Uhrf2 is a master regulator of cholesterol biosynthesis and is essential for liver regeneration. *Sci Signal* 2023;**16**:eade8029. https://doi.org/10.1126/scisignal.ade8029
30. Chemello F, Wang Z, Li H *et al.* Degenerative and regenerative pathways underlying Duchenne muscular dystrophy revealed by single-nucleus RNA sequencing. *Proc Natl Acad Sci USA* 2020;**117**:29691–701. https://doi.org/10.1073/pnas.2018391117
31. Lynch CJ, Kimball SR, Xu Y *et al.* Global deletion of BCATm increases expression of skeletal muscle genes associated with protein turnover. *Physiol Genomics* 2015;**47**:569–80. https://doi.org/10.1152/physiolgenomics.00055.2015
32. Stewart MD, Lopez S, Nagandla H *et al.* Mouse myofibers lacking the SMYD1 methyltransferase are susceptible to atrophy, internalization of nuclei and myofibrillar disarray. *Dis Model Mech* 2016;**9**:347–59. https://doi.org/10.1242/dmm.022491
33. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247
34. Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res* 2024;**52**:W83–94. https://doi.org/10.1093/nar/gkae410
35. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 2020;**36**:2628–9. https://doi.org/10.1093/bioinformatics/btz931
36. Wu T, Hu E, Xu S *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021;**2**:100141.
37. Jain R, Jain A, Mauro E *et al.* ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinf* 2023;**24**:132. https://doi.org/10.1186/s12859-023-05246-8
38. Goulet DR, Yan Y, Agrawal P *et al.* Codon optimization using a recurrent neural network. *J Comput Biol* 2023;**30**:70–81. https://doi.org/10.1089/cmb.2021.0458
39. Jurafsky DM JH. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. 2024. https://web.stanford.edu/~jurafsky/slp3/
40. Goodfellow I, Bengio Y, Courville A. *Deep Feedforward Networks* MIT Press. 2016, https://www.deeplearningbook.org/
41. Mikolov T, Sutskever I, Chen K *et al.* Distributed representations of words and phrases and their compositionality. *Adv Neural Inform Process Syst* 2013;**26**.
42. Cho KvM B, Gulcehre C, Bougares F *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014.
43. Bishop CM. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
44. McWilliam H, Li W, Uludag M *et al.* Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res* 2013;**41**:W597–600. https://doi.org/10.1093/nar/gkt376
45. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 2000;**302**:205–17. https://doi.org/10.1006/jmbi.2000.4042
46. Fan K, Li Y, Chen Z *et al.* GenRCA: a user-friendly rare codon analysis tool for comprehensive evaluation of codon usage preferences based on coding sequences in genomes. *BMC Bioinf* 2024;**25**:309. https://doi.org/10.1186/s12859-024-05934-z
47. Xiao W, Chirmule N, Berta SC *et al.* Gene therapy vectors based on adeno-associated virus type 1. *J Virol* 1999;**73**:3994–4003. https://doi.org/10.1128/JVI.73.5.3994-4003.1999
48. Bell CL, Vandenberghe LH, Bell P *et al.* The AAV9 receptor and its modification to improve *in vivo* lung gene transfer in mice. *J Clin Invest* 2011;**121**:2427–35. https://doi.org/10.1172/JCI57367
49. Gingold H, Tehler D, Christoffersen NR *et al.* A dual program for translation regulation in cellular proliferation and differentiation. *Cell* 2014;**158**:1281–92. https://doi.org/10.1016/j.cell.2014.08.011
50. Edfors F, Danielsson F, Hallstrom BM *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* 2016;**12**:883. https://doi.org/10.15252/msb.20167144
51. Gygi SP, Rochon Y, Franza BR *et al.* Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999;**19**:1720–30. https://doi.org/10.1128/MCB.19.3.1720
52. Maier T, Guell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett* 2009;**583**:3966–73. https://doi.org/10.1016/j.febslet.2009.10.036
53. Koussounadis A, Langdon SP, Um IH *et al.* Relationship between differentially expressed mRNA and mRNA–protein correlations in a xenograft model system. *Sci Rep* 2015;**5**:10775. https://doi.org/10.1038/srep10775
54. Greenbaum D, Colangelo C, Williams K *et al.* Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 2003;**4**:117. https://doi.org/10.1186/gb-2003-4-9-117
55. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 1997;**18**:533–7. https://doi.org/10.1002/elps.1150180333
56. Caldwell R, Lin YX, Zhang R. Comparisons between *Arabidopsis thaliana* and *Drosophila melanogaster* in relation to coding and noncoding sequence length and gene expression. *Int J Genomics* 2015;**2015**:269127. https://doi.org/10.1155/2015/269127
57. Khandia R, Gurjar P, Kamal MA *et al.* Relative synonymous codon usage and codon pair analysis of depression associated genes. *Sci Rep* 2024;**14**:3502. https://doi.org/10.1038/s41598-024-51909-8
58. Hernandez-Alias X, Benisty H, Radusky LG *et al.* Using protein-per-mRNA differences among human tissues in codon

optimization. *Genome Biol* 2023;**24**:34. https://doi.org/10.1186/s13059-023-02868-2
59. Sharp PM, Li WH. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucl Acids Res* 1987;**15**:1281–95. https://doi.org/10.1093/nar/15.3.1281
60. Kudla G, Lipinski L, Caffin F *et al.* High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* 2006;**4**:e180. https://doi.org/10.1371/journal.pbio.0040180
61. Mordstein C, Savisaar R, Young RS *et al.* Codon usage and splicing jointly influence mRNA localization. *Cell Syst* 2020;**10**:351–62. https://doi.org/10.1016/j.cels.2020.03.001
62. Wright JF. Codon modification and PAMPs in clinical AAV vectors: the tortoise or the hare? *Mol Ther* 2020;**28**:701–3. https://doi.org/10.1016/j.ymthe.2020.01.026
63. Wright JF. Quantification of CpG motifs in rAAV genomes: avoiding the toll. *Mol Ther* 2020;**28**:1756–8. https://doi.org/10.1016/j.ymthe.2020.07.006
64. Sabi R, Tuller T. Modelling the efficiency of codon–tRNA interactions based on codon usage bias. *DNA Res* 2014;**21**:511–26. https://doi.org/10.1093/dnares/dsu017
65. Suzuki H, Saito R, Tomita M. The 'weighted sum of relative entropy': a new index for synonymous codon usage bias. *Gene* 2004;**335**:19–23. https://doi.org/10.1016/j.gene.2004.03.001
66. Urrutia AO, Hurst LD. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 2001;**159**:1191–9. https://doi.org/10.1093/genetics/159.3.1191
67. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 1981;**151**:389–409. https://doi.org/10.1016/0022-2836(81)90003-6
68. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* 1982;**158**:573–97.
69. Irwin B, Heck JD, Hatfield GW. Codon pair utilization biases influence translational elongation step times. *J Biol Chem* 1995;**270**:22801–6. https://doi.org/10.1074/jbc.270.39.22801
70. Alexaki A, Hettiarachchi GK, Athey JC *et al.* Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies. *Sci Rep* 2019;**9**:15449. https://doi.org/10.1038/s41598-019-51984-2
71. Lorincz R, Curiel DT. Advances in alpha-1 antitrypsin gene therapy. *Am J Respir Cell Mol Biol* 2020;**63**:560–70. https://doi.org/10.1165/rcmb.2020-0159PS
72. Pipe SW, Leebeek FWG, Recht M *et al.* Gene therapy with etranacogene dezaparvovec for hemophilia B. *N Engl J Med* 2023;**388**:706–18. https://doi.org/10.1056/NEJMoa2211644
73. Colella P, Sellier P, Gomez MJ *et al.* Gene therapy with secreted acid alpha-glucosidase rescues Pompe disease in a novel mouse model with early-onset spinal cord and respiratory defects. *E Bio Med* 2020;**61**:103052. https://doi.org/10.1016/j.ebiom.2020.103052
74. Casazza JP, Cale EM, Narpala S *et al.* Safety and tolerability of AAV8 delivery of a broadly neutralizing antibody in adults living with HIV: A phase 1, dose-escalation trial. *Nat Med* 2022;**28**:1022–30. https://doi.org/10.1038/s41591-022-01762-x
75. Uhlen M, Karlsson MJ, Hober A *et al.* The human secretome. *Sci Signal* 2019;**12**:1260419. https://doi.org/10.1126/scisignal.aaz0274
76. Thul PJ, Akesson L, Wiking M *et al.* A subcellular map of the human proteome. *Science* 2017;**356**:eaal3321. https://doi.org/10.1126/science.aal3321
77. Uhlen M, Fagerberg L, Hallstrom BM *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 2015;**347**:1260419. https://doi.org/10.1126/science.1260419
78. Apweiler R, Bairoch A, Wu CH *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004;**32**:115D–119. https://doi.org/10.1093/nar/gkh131
79. Fu H, Liang Y, Zhong X *et al.* Codon optimization with deep learning to enhance protein expression. *Sci Rep* 2020;**10**:17617. https://doi.org/10.1038/s41598-020-74091-z
80. Tian T, Li S, Lang P *et al.* Full-length ribosome density prediction by a multi-input and multi-output model. *PLoS Comput Biol* 2021;**17**:e1008842. https://doi.org/10.1371/journal.pcbi.1008842
81. Tunney R, McGlincy NJ, Graham ME *et al.* Accurate design of translational output by a neural network model of ribosome distribution. *Nat Struct Mol Biol* 2018;**25**:577–82. https://doi.org/10.1038/s41594-018-0080-2

Received: September 19, 2024. **Revised:** March 3, 2025. **Editorial Decision:** March 7, 2025. **Accepted:** March 28, 2025