Annual Review of Biomedical Data Science

Modeling the Language of Codons with Artificial Intelligence

Claudèle Lemay-St-Denis and Rachel Kolodny

Department of Computer Science, University of Haifa, Haifa, Israel;
email: claudele.lemaystdenis@gmail.com, trachel@cs.haifa.ac.il

Annu. Rev. Biomed. Data Sci. 2026. 9:475–500

First published as a Review in Advance on May 26, 2026

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-092624-105741

Keywords

deep learning, language models, codon usage, codon optimization, protein design

Abstract

The messenger RNA (mRNA) sequence plays a central role in expressing functional proteins from genes. Species-specific nonuniform codon choices influence key processes, including mRNA stability, translation accuracy and efficiency, and cotranslational folding. As evolutionarily selected codon sequences encode multiple overlapping, yet often subtle, signals, advanced deep learning models are needed to identify their statistical patterns. This review summarizes recent progress in artificial intelligence (AI) models that operate at the codon level, including both discriminative and generative approaches. We discuss codon language models and their application to downstream prediction tasks, as well as generative models used to design codon sequences for heterologous expression and to recode proteins for improved mRNA stability and expression. Finally, we highlight studies leveraging codon-based AI models to gain insights into molecular evolution and outline several open problems in this rapidly developing field.

1. INTRODUCTION

Artificial intelligence (AI) models, powerful computational tools that learn from data, have revolutionized protein science. Once trained, AI models capture the underlying distributions of functional proteins, and they can be used to predict protein properties and even sample, or generate, novel proteins from the learned distribution. Notable examples include AlphaFold (1) and ProteinMPNN (2). AlphaFold predicts a protein's structure from its amino acid sequence; ProteinMPNN designs amino acid sequences compatible with a given structure. An important subclass of AI models are language models (LMs), which learn meaningful representations of the data and can be used in downstream tasks. In protein science, protein LMs (pLMs) learn from the databases of protein sequences and even incorporate richer protein modalities, such as structure (3–5; for recent reviews, see 6–9). Nonetheless, pLMs exclusively represent protein sequences at the abstraction level of amino acids (10).

Genomic LMs (gLMs) model nucleotide sequences (11–13) and can be applied to genomic, downstream prediction, and generative tasks (for recent reviews, see 14, 15). Genomes encode sequence types encompassing regulatory elements, noncoding RNA genes, and extensive intergenic DNA. Therefore, although coding sequences are part of the training sets, they are only a fraction of the data modeled by gLMs (16). A biologically meaningful and compact representation for coding regions is a sequence of codons, the language of the ribosome. As outlined in the biological background below, substantial evidence shows that codon-level signals are critical for functional protein synthesis. Here, we discuss recent AI-based studies that model proteins at the level of species-dependent codon sequences.

AI models are often classified as discriminative or generative. Discriminative models learn to predict, classify, or score inputs with discrete or continuous labels. More formally, they learn the likelihood $p(y|x)$ of an output y (e.g., codon label or expression level), given an input codon sequence x. Generative models aim to model the underlying data distribution $p(x)$, allowing them to sample, or generate, new, realistic codon sequences from this distribution. Codon-based AI models do not necessarily adhere to these boundaries, in that both model types are used to predict properties and generate codon sequences.

This review is organized as follows: After a biological background, we highlight codon LMs (cLMs) developed to represent coding sequences, along with the downstream tasks to which they have been applied. Then, we describe generative models for codon sequences. Because codon sequences are species dependent, we distinguish between generative tasks aimed at heterologous expression and those focused on optimized endogenous expression. Finally, we discuss evolutionary insights derived from analyzing AI models of codon sequences and conclude with future directions.

2. BIOLOGICAL BACKGROUND

Proteins are macromolecules built from 20 amino acids, encoded by 61 codons. The codon vocabulary is degenerate: Except for methionine and tryptophan, each amino acid can be specified by two to six codons, dubbed synonymous. Consequently, there are exponentially many theoretical codon encodings for each protein sequence. Early analysis of different genomes revealed that there are preferred codons within a genome, a phenomenon termed "codon bias" (17–19). Some codons are optimal, in that they can be decoded rapidly and accurately by efficient recruitment of cognate charged transfer RNAs (tRNAs) (20–22). Efficient recruitment depends on both the supply of charged tRNAs and the codon demand in expressed genes (23). Multiple computational indices were developed to quantify this (for reviews, see 24, 25). Producing sufficient amounts of functional proteins requires passing through several steps: synthesis of stable messenger RNA

(mRNA) through transcription and maturation, efficient and accurate translation after proper initiation, correct folding of the nascent polypeptide, and appropriate posttranslational modification and localization. We outline how codon encodings impact these processes.

The codon sequence influences mRNA transcription (26). Nonoptimal codons significantly suppress transcription through chromatin modification, as shown in *Neurospora* (27), in yeast, and in mammals (28–30). Radrizzani et al. (31) proposed that in humans, negative selection on codon use can reduce unwanted transcripts. The codon sequence also affects mRNA maturation, as synonymous mutations can disrupt splicing and cause disease (32, 33).

Translated and untranslated regions of the mRNA influence its stability (34–38). The quantitative measure of stability, mRNA half-life, varies widely among species—from ∼1 to 18 min in *Escherichia coli* (39) to a median of ∼10 h in human cells (40). Within an organism, transcripts enriched in optimal codons have longer half-lives, whereas enrichment in nonoptimal codons accelerates decay through mRNA turnover (23, 41). For example, in *Drosophila* and *Neurospora*, ribosomal stalling on nonoptimal codons can induce premature termination (42). Translation-independent decay also contributes, with its relative importance varying across systems (43–45). The computational estimation of mRNA minimum free energy (MFE) (46, 47) was shown to correlate with in vitro mRNA half-life measurements (48).

The ribosome decoding speed depends on the availability of charged tRNA and codon–anticodon wobble (49). The tRNA supply depends on its concentration in the cell (50) and may vary, e.g., depending on cellular state or induced stress (51–54). Codons with abundant cognate tRNAs are decoded rapidly and accurately (20–22). Thus, translation speed, and consequently efficiency, depends on codon use (55, 56). In *E. coli* and in eukaryotic cells, it was shown that positions coding for conserved amino acids are biased to more frequent codons (22, 57–60), suggesting that selection favors optimal codons to minimize translational errors. Tuller and coworkers (61, 62) proposed that the 5′ region of the coding sequence influences the initiation efficiency by using slowly translated codons to reduce ribosome collisions, thereby creating a ramp; this model remains debated (63–65).

A mismatch between codon composition and the tRNA pool places a burden on the organism (66). Codon demand and tRNA abundance can be viewed as an evolved, dynamic market. Disturbing this market affects the efficiency with which genes are translated (67). For example, to offset mismatches imposed on the host tRNA pool by viral codon usage, some viruses enhance translational efficiency by encoding some of their own tRNAs (68, 69). The tRNA adaptation index (tAI) computationally quantifies this relationship (70, 71) and was shown to correlate with protein expression levels (72).

Cotranslational folding, or the folding of the nascent chain as it emerges from the ribosome, is shaped by the pace of translation (73, 74). Excessively slow and excessively fast elongation can promote misfolding and aggregation (75–77). Thus, codon choice regulates nascent protein folding via translation speed. For instance, Zhang et al. (78) showed that replacing rare codons with abundant ones disrupted folding of a bacterial protein both in vitro and in vivo. Buhr et al. (79) reported that differences in synonymous codons between *Bos taurus* and their harmonized heterologously expressed *E. coli* counterparts led to distinct protein conformations. Conformational change can even impact function: Sander et al. (80) showed that using rare codons in a two-state designed fluorescent reporter protein altered its cotranslational folding, leading to marked differences in fluorescence.

As coding sequences encode signals required in the synthesis of well-folded proteins, they are under evolutionary selection (81, 82). In some cases, the selection of a specific codon is essential for proper function, as evidenced by human diseases due to synonymous mutations that alter splicing, mRNA stability, or translation efficiency (83, 84). However, this does not apply to all positions.

Rather, in many positions—possibly most—the choice of the synonymous codon is under nearly neutral selection (81, 85). Furthermore, that the codon sequence encodes multiple interdependent layers of regulation makes the disentangling of individual effects difficult (85).

The effect of a codon depends on its sequence context and epistatic interactions within the sequence. For example, the effects of synonymous mutations in the *CFTR* gene were shown to depend on disease-causing missense mutations in other parts of a sequence, and the combination resulted in a positive epistatic effect on the folding of the nascent peptide chain (86). Such nonadditive, context-dependent interactions represent a significant challenge to the modeling of local and long-range interactions with frequency-based or simple statistical models (87).

Substantial amounts of data are required to achieve statistical significance when identifying those complex patterns. However, as these patterns are species specific, the effective dataset is limited by proteome size, ranging from hundreds to tens of thousands of genes. This restricts the level of confidence with which such signals can be identified. Thus, sophisticated statistical machines like contemporary deep learning models are needed to learn the underlying patterns of codon usage from the available data.

3. CODON-BASED ARTIFICIAL INTELLIGENCE MODELS FOR REPRESENTATION

The design and training choices of an AI model intended for use in downstream tasks are generally guided by those tasks. The task determines which modality encodes the most relevant biological signal, such as the codon sequence, amino acid sequence, or protein structure. One must also decide on the model's architecture. For cLMs, the models are almost universally based on transformers (15), which can efficiently handle input lengths typical of codon sequences (88). One should also determine training data. Large datasets of unlabeled codon sequences are used to self-supervise the training of cLMs. Typically, the datasets are functionally and phylogenetically diverse (89–91), although some models train on specific protein types such as antibodies (88) or specific taxonomic groups (92). When training on taxonomically diverse datasets, an input token specifying the source organism is often added. Larger models, with more parameters, can lead to better results, but they require more training data and computational resources (89, 90, 93).

The loss functions guiding the self-supervised training of codon sequence models vary. Let a codon sequence of length n be marked by $X = (x_1, x_2, \ldots, x_n)$; each of the positions x_i is a codon among the 61 alternatives. The model is trained to assign a probability to each combination of position and codon type (**Figure 1*a***). Most commonly, masked language modeling (MLM) and its variants are used. In MLM, the model is trained to predict one or many masked positions, given all other unmasked codons in the sequence, namely, to maximize $P(x_i|X_{\setminus \text{mask}})$, $\forall i \in$ mask. If the amino acid in each position is known, the model can restrict its predictions to the subset of synonymous codons (88, 91, 94, 95). Causal language modeling (causal-LM; while this is often denoted CLM; we use causal-LM to avoid confusion with cLMs) is an alternative approach where the model, typically a decoder, is trained to predict the kth codon, given all its preceding codons, namely, to maximize $P(x_k|x_1, \ldots, x_{k-1})$. The loss function can also combine more than one task: For example, Li et al. (96) trained on sequence taxonomy prediction (STP), identifying whether two (masked) codon sequences are from the same taxonomic class, in addition to MLM.

As mentioned above, cLMs compute, for each of the n input positions, the probabilities of the 61 codon types. These probabilities are a function of the values of the nodes in the last layer, which in turn are a function of their previous layers. The d-dimensional vectors with the values in this last layer (d is fixed and depends on the model architecture) are referred to as the model's latent space, or embeddings (see **Figure 1*a***). The embeddings contain meaningful information

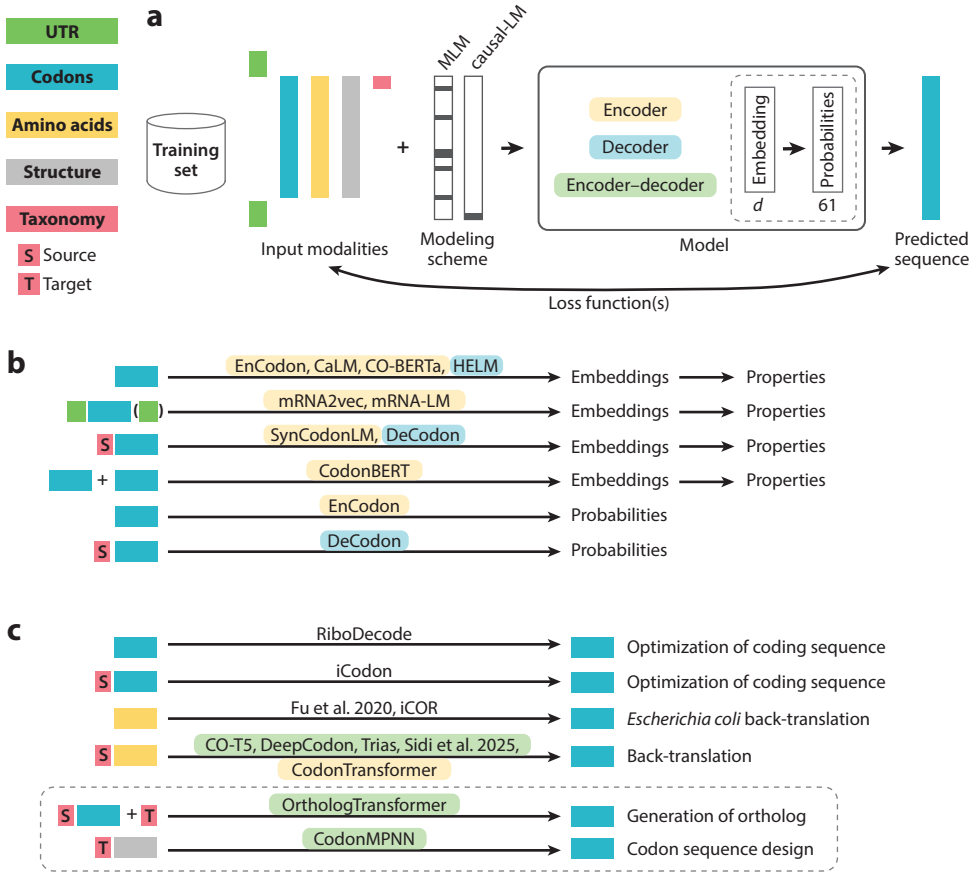

Figure 1

An overview of codon-based artificial intelligence models and their downstream applications. (*a*) cLMs rely on different input modalities (codon sequences, amino acid sequences, protein structure, UTRs, and taxonomic labels) and different modeling schemes (MLM and causal-LM). The models are deep networks with multiple layers. The last embedding layer of *d* dimensions is used to calculate codon probabilities. Weights, or model parameters, are iteratively updated to optimize the loss function(s). (*b*) Discriminative cLMs tokenize coding sequences, in some cases combined with taxonomic information. The learned embeddings are then used as input to fine-tuned networks that predict downstream properties, or used directly in zero-shot prediction. (*c*) Generative models use diverse inputs to sample, or generate, codon sequences for distinct goals: codon optimization, back-translation, and codon sequence design. Circled in a dashed line are the generative models that sample both codon and amino acid sequence space. Abbreviations: causal-LM, causal language modeling; cLM, codon language model; MLM, masked language modeling; UTR, untranslated region.

and can be used as input features to downstream tasks, where specialized networks are fine-tuned on smaller, labeled datasets. Note that any single, or combination, of layers in the trained model can be used as the embeddings for a downstream task. We list examples below and in **Table 1**.

Embeddings from cLMs can also be visualized after dimensionality reduction [with t-SNE (97) or UMAP (98)] to better understand their learned interrelationships. Embeddings of synonymous codons differ from each other (89, 90, 95, 96), yet they are closer to each other than to those of nonsynonymous ones (89, 90, 96). Embeddings of codons in the same wobble pair are also closer to each other (90, 95). The cLM latent spaces are also taxonomically coherent: Averaged embeddings

Table 1 Datasets and expression systems used for evaluating codon language models across biological processes

Biological process	Dataset	Type of expression host	Evaluated codon language models	Dataset reference(s)
mRNA abundance	RNA sequencing datasets	Endogenous	CaLM, CodonBERT	109–112
	TDH3 in *Saccharomyces cerevisiae*	Endogenous	SynCodonLM	108
	GFP in *S. cerevisiae*	Heterologous	SynCodonLM	108
	Transcript abundance in human tissues	Endogenous	mRNA-LM	132, 133
mRNA stability	mRNA decay rate in human and mouse dataset	Endogenous	EnCodon, DeCodon, mRNA-LM	113
	mRNA decay rates of randomized 100–amino acid transcripts in zebrafish embryos	Heterologous	CodonBERT, mRNA2vec, HELM	114
	RNA decay rate of synonymous variants of enhanced GFP, multi-epitope SARS-CoV-2 vaccine, and NanoLuc luciferase	In vitro	CodonBERT, mRNA-LM	48
	Degradation of participant-designed RNA vaccine sequences	In vitro	HELM, CodonBERT	115
RNA toxicity	Effect of expression of synonymous GFP variants on *Escherichia coli* growth	Heterologous	SynCodonLM	116
Ribosome occupancy	Ribosome profiling data	Endogenous	EnCodon, DeCodon	118
Protein function	Gene Ontology terms	NA	CaLM	192
	Subcellular localization	NA	CaLM	192
	Pathogenecity effect in human	Endogenous	EnCodon, DeCodon	138
	Fitness derived from deep mutational screening dataset	Endogenous	EnCodon, DeCodon	137
Protein abundance	FolA in *E. coli*	Endogenous	CO-BERTa	92
	Viral protein in human cells	Target host	CodonBERT, HELM	96
	Curated datasets	Endogenous	CaLM	135
	Viral proteins in human cells	Target host	mRNA-LM	130
	Proteins in human tissues	Endogenous	mRNA-LM	132
	GFP and VHH in *E. coli*	Heterologous	CO-BERTa	92
	Monomeric red fluorescent protein in *E. coli*	Heterologous	EnCodon, DeCodon, CodonBERT, SynCodonLM, mRNA2vec, HELM	129
	GFP in *E. coli*	Heterologous	SynCodonLM	128
	Proteins from all superkingdoms in *E. coli*	Heterologous	CodonBERT	134
Protein properties	Solubility in human cell lysate	Endogenous	CaLM	139
	Stability in 13 species	Endogenous	CaLM	141
Species	Species identity	NA	CaLM	192

Abbreviations: GFP, green fluorescent protein; mRNA, messenger RNA; NA, not available.

of proteins cluster according to their source species; Outeiral & Deane (90) showed that ones from different superkingdoms are distinct, whereas those from the same order overlap, and Li et al. (99) showed clustering by biological taxa. An exception to this are highly conserved sequences such as ribosomal proteins or cell cycle enzymes from different species that are embedded near each other (90). Dimensionality reductions aim to preserve local neighborhoods in the original latent space, but their visual representation may be distorted, and thus should be interpreted with caution. Outeiral & Deane (90) argue that as cLMs learn the nonequivalence of synonymous codons—a signal inaccessible to pLMs, which treat them as a single token—their embeddings are richer than those output by pLMs.

3.1. Highlighting Trained cLMs

EnCodon is a suite of cLM models of varying scales (from 80M to 1B trainable parameters), trained with MLM on the National Center for Biotechnology Information (NCBI)'s Genomes database (60M sequences from over 5,000, mostly bacterial, species) (89). The DeCodon models were trained with causal-LM and conditioned on a species token (89). EnCodon and DeCodon also have variants that were further trained on data from 227 eukaryotic species. Naghipourfar et al. (89) compared the performance of EnCodon/DeCodon models to other gLMs and cLMs on prediction tasks including protein expression, mRNA stability, fitness, and pathogenicity effects. They show that larger models perform better and that the fined-tuned causal-LM model outperforms its MLM counterpart.

CaLM, the codon adaptation language model, is an ESM-like model (100) with 86M parameters, trained with MLM on over 9M nonredundant, diverse-taxa codon sequences from the European Nucleotide Archive (90). Outeiral & Deane (90) compared the performance of CaLM to other pLMs in predicting protein function, subcellular localization, melting point, and solubility. They showed that their cLM matches or outperforms pLMs of comparable size, and in some cases, even ones 50 times larger.

CodonBERT is a 110M-parameter model trained on 10M codon sequences spanning a diverse set of species (mammals, *E. coli*, human viruses, and yeast) (96). Its loss function is unique in that it combines MLM and STP, with the latter predicting whether two input sequences are from the same taxonomical category. Li et al. (96) compared the performance of CodonBERT to more naïve methods (in that all but one are smaller, nontransformer models) on a diverse set of downstream tasks related to mRNA translation, stability, regulation, and protein expression and showed that the codon-based models outperform the nucleotide-based ones on most tasks. mRNA-LM extends CodonBERT by coembedding the codons with encodings of the 5′ and 3′ untranslated regions (UTRs) (99). Li et al. (99) show that mRNA-LM performs better on tasks related to mRNA stability, translation rate, transcript abundance, and protein expression. Thus, they demonstrate that the UTR of mRNAs, known to play an integral role in protein synthesis, can enrich the learned signal in codon-based AI models.

CodonTransformer is a 90M-parameter model based on the BigBird transformer, a BERT variant developed for long sequences (91). It is trained with MLM on two datasets: Pretraining is on 1M codon sequences from 164 species. Then, it is trained on the top 10% of the genes from 13 representative species, with highest codon similarity index (CSI). The computational CSI is used to select predicted high-expression genes (101). CodonTransformer implements an amino acid–specific masking to inform the model of the amino acid sequence. Although CodonTransformer is an encoder-only LM, Fallahpour et al. (91) use it to generate codon sequences (see Section 4.1).

SynCodonLM is trained on 66M codon sequences of nearly 35,000 unique and diverse species from NCBI RefSeq (95). In SynCodonLM, approximate information about the species' codon bias

is added to the input codon tokens. The species-specific information is approximated to avoid assigning a unique token per species. Rather, these are clustered into 501 groups based on their relative synonymous codon usage profiles. The amino acid sequence of each protein is passed as input and SynCodonLM uses a variant of MLM, which restricts predictions to its synonymous codons. Heuschkel et al. (95) compared the performance of SynCodonLM to other cLMs (including CodonTransformer) on fivefold cross-validation prediction of mRNA toxicity, transcript abundance, and protein expression. They show that on the seven datasets they consider, SynCodonLM performs comparably or better than other cLMs.

mRNA2vec (102) is trained on over 500,000 mRNAs from human, rat, mouse, chicken, and zebrafish. mRNA2vec includes up to 102 nucleotides of 5′ UTR, tokenized as codons together with the coding sequence. Also, it is not a transformer-based model; rather, it follows the data2vec self-distillation framework (103). The model has a student network that is presented with masked input (where the masking probability is higher around the start codon) and a teacher network that is presented with unmasked input. Both networks output embeddings for two downstream prediction tasks of MFE and RNA secondary structure. The training loss has terms comparing the student and teacher outputs and the accuracy of the predictions. Zhang et al. (102) compared their model with CodonBERT and demonstrated that it performs equivalently in predicting protein expression and outperforms CodonBERT in predicting mRNA stability.

3.2. Codon-Related Downstream Tasks

Downstream tasks for cLMs differ from those of pLMs for two main reasons: The benchmark datasets that were originally developed to evaluate pLMs (104–106) include protein properties, yet their codon sequences are often unavailable. In fact, it is often impossible to retrieve the coding sequence used in experiments, as proteins are frequently codon-optimized with proprietary algorithms and the coding sequence is not documented. Also, cLMs encode codons, and as such they can predict downstream tasks related to mRNA. These include predicting the abundance of transcripts, the stability of mRNAs (either in vivo or in vitro), and the overall translation and folding success. Note, however, that most cLMs process only the protein-coding sequence even though mature mRNAs contain 5′ and 3′ UTRs essential for mRNA stability and for its translation by the ribosome. These downstream tasks are used not only as tasks in their own right but also to compare the performance of cLMs. **Figure 1*b*** outlines the inputs and trained cLMs, and **Table 1** lists the downstream prediction tasks and models whose embeddings were used to solve them.

3.3.1. mRNA abundance. mRNA abundance can be quantified at the transcriptome level using RNA sequencing (RNA-seq) measurements from deep sequencing of converted complementary DNA (cDNA) fragments (107). Chen et al. (108) quantified mRNA abundance of degenerate libraries of green fluorescent protein (GFP) and TDH3 proteins in *Saccharomyces cerevisiae*, representing both heterologous and endogenous expression. More broadly, curated databases, such as the Gene Expression Omnibus (109), Expression Atlas (110), Sequence Read Archive (111), and the MycoCosm database (112), include experimental datasets for diverse tissues and species. Predicting transcript abundance integrates multiple transcript-level factors, including transcription rate and mRNA stability.

3.3.2. mRNA stability. Datasets with mRNA stability information obtained in vitro or in vivo are used for evaluation. Agarwal & Kelley (113) compiled mRNA half-life data of diverse coding sequences across diverse conditions from 39 human and 27 mouse experiments. Diez et al. (114) injected in vitro–transcribed libraries of randomly generated 100–amino acid proteins, i.e., proteins that were not under evolutionary selection, into zebrafish embryos and measured mRNA

abundance over time. Leppek et al. (48) measured in vitro RNA decay of synonymous variants of the enhanced GFP (eGFP), multi-epitope SARS-CoV-2 vaccine, and NanoLuc luciferase, by tracking over time the integrity of mRNA in a magnesium-containing buffer. Wayment-Steele et al. (115) also have in vitro RNA degradation profiles.

3.3.3. RNA toxicity. Heterologous expression can decrease a cell's fitness by creating a burden due to increased demand on its resources. Mittal et al. (116) characterized the effect of expressing 347 synonymous variants of GFP in *E. coli* on bacterial fitness. They demonstrated that the reported impact of expression was independent of translation and that growth inhibition was due to mRNA toxicity.

3.3.4. Ribosome occupancy. Ribosome profiling (Ribo-seq) experiments measure nucleotide and transcript resolution of ribosome occupancy by sequencing mRNA fragments that are ribosome-protected during translation from nuclease digestion (117, 118). These data reveal slowly or rapidly translated regions and identify codons that cause ribosomal stalling. Although cLMs can model such patterns, the ones described above did not, and only a simpler codon-aware machine learning model did (119). Note that Ribo-seq data are noisy and susceptible to experimental biases (120–122).

3.3.5. Species identity. CaLM demonstrated that the codons carry a richer species-specific signal than amino acids (123–125). Thus, cLMs can predict taxonomic information (89, 92, 95). Solving this task is valuable in contexts where the species is unknown, for example, in the taxonomic classification of metagenomic contigs (126, 127).

3.3.6. Protein expression. The amount of produced protein can be measured directly or with a readily measured function as proxy (e.g., fluorescence). Expression datasets vary widely in scope and design. Some datasets contain only synonymous variants of one protein, denoted degenerate libraries, and aim to capture the relationship between codon choice and translation efficiency. Others include measurements for codon sequences of diverse amino acid sequences. Constant et al. (92) designed degenerate libraries for GFP and an anti-HER2 VHH domain, measuring expression in *E. coli* via fluorescence and activity-specific cell enrichment. They also constructed a degenerate library for *E. coli* FolA, where highly expressed variants were selected with two folate biosynthesis inhibitors. Schmitz & Zhang (128) measured the expression of a GFP degenerate library in *E. coli*. Nieuwkoop et al. (129) designed three degenerate libraries for the monomeric red fluorescent protein, in which codons were uniformly and randomly picked, more frequent synonymous codons were favored, and less frequent synonymous codons were favored. They quantified abundance by a combination of fluorescence-activated cell sorting, flow cytometry, and a microplate reader. Li et al. (96) transcribed a library encoding a viral hemagglutinin protein, transfected it into HeLa cells, and quantified protein levels in lysates by ELISA (enzyme-linked immunosorbent assay). Zhang et al. (130) transfected in HEK293 cells synonymous codon sequences for viral proteins designed with their LinearDesign algorithm. Eichhorn et al. (131) measured protein expression per transcript using Ribo-seq and RNA-seq from two human cell types, while Wang et al. (132) curated transcript and protein abundance data from 29 healthy human tissues from the Human Protein Atlas (133). Boël et al. (134) generated a dataset of 6,348 naturally evolved proteins from all three superkingdoms (<60% amino acid identity) expressed in *E. coli*. They visually inspected the expression in whole-cell lysates on SDS-PAGE and classified it into one of five categories. These data capture translation efficiency in a heterologous expression context, where codon–tRNA relationships differ from those in the native host. Protein expression data for a diversity of species is curated in the Protein Abundance Database (135). Recent advances in mass spectrometry–based proteomics approaches have made it possible to

measure protein abundance rapidly, quantitatively, sensitively, and specifically across diverse sample types, including bulk tissues, single cells, and heterogeneous populations (136). These methods represent a potential rich source of data for fine-tuning cLMs.

3.3.7. Protein function. Predicted functions from cLM embeddings are subcellular localization annotations, as well as Gene Ontology (GO) terms (molecular function, cellular component, and biological process) of SwissProt proteins. Another way of framing this task is to predict the impact of codon sequence variants on fitness or pathogenicity. Data from deep mutational screening datasets, such as MaveDB (137), and repository datasets, such as ClinVar, which catalogs human variants (138), were used to this end (89).

3.3.8. Protein stability and solubility. Sridharan et al. (139) includes a dataset of protein-wide solubility, using as proxy measurements of the protein abundance in an SDS-treated fraction of human cell lysates (140). The Meltome Atlas has melting temperature data determined by mass spectrometry for proteins from 13 species and 19 human sample types (141).

4. GENERATIVE MODELS

Generative models sample codon sequences for a protein sequence to be expressed in a target organism, and potentially a specific cell type or tissue. AI generative models offer a path to train models from naturally occurring sequences, i.e., those selected by evolution, and then sample statistically similar patterns. The data available to train an organism-specific model are limited to the genes of that organism. Larger datasets can be constructed by pooling sequences from many, possibly taxonomically related, species.

Once trained, the sequences generated by a model should be evaluated. The simplest case is when the protein is naturally expressed in the organism, meaning there is a textbook solution identified by evolution. This evolutionarily selected codon sequence can be used to evaluate the sequences generated by the AI model (Section 4.1) or to study the evolutionary process itself (Section 5). There are different ways to sample, or generate, sequences from a trained model, and we highlight those that are currently used (Section 4.2).

AI and earlier computational models were mostly developed in the context of heterologous expression, where the goal is to generate codon sequences for which there are no existing solutions provided by evolution (Section 4.3). In heterologous expression, recombinant proteins are produced in a host organism different from their native one. This process has applications ranging from biochemical research to pharmaceutical manufacturing and industrial biotechnology. For example, *E. coli* BL21(DE3) cells are among the most widely used systems owing to their rapid growth and ease of manipulation (142), and mammalian cells are better suited for expressing glycoproteins such as antibodies because they perform necessary posttranslational modifications (143, 144). The constraints on the generation of the codon sequence for heterologous expression can be relaxed, allowing for a species-specific adaptation of the protein sequence (Section 4.4). We also describe the use of generative models to recode the codon sequences of naturally occurring proteins within the same species for pharmaceutical application (Section 4.5). Finally, we describe how generative models are used to score codon sequences (Section 4.6).

4.1. Generative Models and Their Evaluation

Early studies trained models for codon sequence generation in *E. coli* and used recurrent neural network architectures because these can be effectively trained on the relatively scarce data. Fu et al. (145) used ~5,000 *E. coli* sequences to train a bidirectional long short-term memory (Bi-LSTM) network on amino acid sequences that autoregressively outputs codon sequences (or a

codon box, which is a group of codons). iCOR is another Bi-LSTM model that was trained on ∼7,000 amino acid–redundant *E. coli* genes (146). More recent models use attention layers, which incorporate information from distant positions. CO-T5 (92) is an amino acid–to-codon encoder–decoder model with transformer layers that was pretrained on a larger set of protein sequences from high-quality genomes from the *Enterobacterales* order and conditioned on the host genome. Thus, CO-T5 learned from a richer dataset while maintaining specific *E. coli* knowledge. Recently, the transformer model DeepCodon was trained to autoregressively generate a codon sequence for *E. coli* from an amino acid sequence (147). DeepCodon was trained on a larger dataset of *Enterobacteriaceae* sequences: first pretraining on a nonredundant set of ∼1.5 million sequences and then fine-tuning on ∼68,000 sequences.

CodonTransformer is a multispecies transformer model (91). Typically, the architecture of transformer-based generative models is an encoder–decoder (e.g., as in 94). CodonTransformer, however, is an encoder-only model that is used for generation (one can view its last layer as a trivial decoder). During training, CodonTransformer is fed with partially masked codon sequences and full amino acid information from different organisms, with the goal of predicting the masked codons. During inference, the input is the target organism and an amino acid sequence and the model predicts all codons, albeit not autoregressively.

Evaluating generated codon sequences is difficult. When generating codon sequences of nonnaturally occurring proteins (see Section 4.3), the most reliable way is to experimentally characterize these generated sequences, measuring protein expression or function. However, as this is extremely resource intensive, it is often limited to a small number of proteins, with an overrepresentation of fluorescent proteins selected for their ease of functional screening. Constant et al. (92) generated codon sequences for two *E. coli* proteins and experimentally validated that their CO-T5 sequences were more highly expressed than sequences optimized by the commercial GenScript, Genewiz, IDT, and Twist algorithms. Fu et al. (145) experimentally showed that their tool generated codon sequences for two *E. coli* proteins that were more highly expressed than those of the commercial optimizers Genewiz and Thermo. Han et al. (147) compared the expression levels of the codon encodings for 20 *E. coli* proteins generated by their DeepCodon tool to those optimized by GenScript. Ravi et al. (148) measured the expression level of their AI-optimized codon sequences for eGFP and firefly luciferase (FLuc) in corresponding cell lines, which were optimized for mouse brain, liver, and muscle tissues. They showed that all their AI-optimized codon sequences increased expression compared to the wild-type and GenScript-optimized sequences, although it was not exclusive to the targeted cell line.

Many studies complement, and even replace altogether, the experimental evaluation with a computational one. When generating a coding sequence for the heterologous expression of proteins with native or evolutionarily selected counterparts, these target proteins can be excluded from the training set (91, 94, 149). The sequence identity between the generated and native coding sequences can be computed to evaluate generation. This, however, is a misleading framing of the challenge, as evolutionary constraints do not define a unique solution. Fallahpour et al. (91) compared 50 random held-out genes from highly expressed sequences in five organisms to their and others' generated sequences (the commercial optimizers Genewiz, IDT, and Twist, and iCOR for *E. coli*). Rather than measure the percentage identity between native and generated sequences, they compared global statistical properties (such as GC content and MFE). They also demonstrate that the corresponding sequences have similar codon patterns using %MinMax (150), which measures the agreement in rare and common codon usage patterns, in a sliding window along the sequence. Notice that to compare distributions of properties in a generated and naturally occurring set, there is no need for a one-to-one correspondence between the two. Indeed, evaluation of global statistical measures can also be done for heterologous proteins with no naturally occurring counterparts

in the target organism, and multiple studies compare the distribution of different such statistics: codon adaptation index (CAI), content, tAI, codon frequency distribution, pairwise frequencies, negative repeat elements, negative *cis*-regulatory elements, and the retention of conserved rare codon clusters (91, 145–148). Only %MinMax has a short-window resolution; other metrics capture correlations at the complete coding-sequence level (151–154). Indeed, these measures cannot capture instances in which a single codon is critical and a synonymous substitution has dramatic impact on expression (155). Optimizing sequences to reproduce global statistical features, instead of validating experimentally, risks missing the true determinants of expression.

4.2. How to Generate, or Sample, Codon Sequences

The term "generative" emphasizes that during inference, trained models can be used to sample, or generate, codon sequences from the learned distribution. The most straightforward approach is to use the generative models and autoregressively sample one or many sequences (156). The conditioning used during training can guide the model during inference. For example, Ortholog-Transformer (Section 4.4) is conditioned on the codon sequence and species in the source protein, as well as the target species (149). The selection of codons during generation can be further tuned with additional constraints. For example, Akiyama et al. (149) use a Monte Carlo tree search strategy to look ahead during generation and consider the three most favorable extensions, as evaluated based on their GC content and mRNA stability, while Patel et al. (157) use the Monte Carlo tree guidance with multiple objectives. Constant et al. (92) selected for in vivo testing the 10 best and 10 worst sequences from a much larger pool of 10^8 generated sequences, based on their predicted expression by their companion discriminative AI model. The sampling procedure in RiboDecode (Section 4.5) is based on gradient ascent on the scores of trained networks (158). The iCodon model generates codon sequences that are optimized or deoptimized for mRNA stability by using an algorithm that simulates an evolution-like process where the sequence iteratively accumulates synonymous iterations under the supervision of mRNA stability predictions (114).

4.3. Generative Models Can Be Used to Generate Sequences for Heterologous Expression

The most common goal of recombinant codon optimization is to sample, or generate, a sequence among the exponentially many codon sequence alternatives coding for a given amino acid sequence that optimizes yield (159). This has been a long-standing task, for which there are non-AI computational tools (160); for example, CHARMING (161) uses harmonization, which replicates the codon frequency usage pattern in the source species with the codon bias of a target species. Other proprietary codon optimization algorithms (e.g., GenScript, Genewiz, IDT, Twist, Thermo) optimize predefined metrics and integrate negative design elements, typically with the result of increasing the frequency of the target species' frequent codons.

Curating data for training AI models to generate highly expressed sequences can be challenging because evolution is not necessarily optimizing expression, suggesting that not all data are suited for training a model with this goal. Selecting only sequences from highly expressed proteins limits the dataset, which may already be limited by the number of proteins in a single species. Also, expression-level information may not be available for the target species.

Many of the abovementioned generative models (see Section 4.1) were trained for generating highly expressed recombinant proteins. To this end, the developers of iCOR restricted their training data to *E. coli* genes with high CAI values, i.e., genes that use *E. coli*'s more frequent codons and are assumed to be of highly expressed proteins (146). Constant et al. (92) trained alongside CO-T5 the CO-BERTa network, for predicting *E. coli* expression from the codon sequence. CO-BERTa,

pretrained on the same dataset as CO-T5, was then separately fine-tuned on *E. coli* expression data of ∼150,000 codon variants of three proteins: the *E. coli* FolA and two heterologous proteins (GFP from *Aequorea victoria* and a camelid VHH antibody domain). DeepCodon (147) was fine-tuned on ∼68,000 sequences that were filtered for high CAI and tAI values, used as proxies for high expression (24).

When optimizing codon sequences in humans, one may need to also consider the expressing tissues. Hernandez-Alias et al. (162) showed that transcriptomics and proteomics data in the GTEx and Human Protein Atlas include enriched/depleted codon patterns in specific tissues. They further offered an experimental proof of concept whereby the codon sequences for heterologous expression of eGFP and mCherry were optimized for human kidney and lung tissues. Ravi et al. (148) developed Bi-LSTM models tailored for human brain, liver, and muscle tissues for heterologous expression of eGFP and the FLuc, trained on genes with the most abundant transcripts in these tissues.

Maximizing translational efficiency by the near-exclusive use of frequent codons may lead to excessively rapid or abundant translation, which can be toxic or impose a burden on the limited resources of the host, thereby reducing cellular fitness (66, 163–167). Love & Nair (168) propose the alternative strategy of designing coding sequences that minimize competition with host translational demands rather than maximize the use of common codons. Their implementation empirically identifies codons linked to higher or lower heterologous protein expression to guide coding sequence design. It is hard to imagine how to reproduce this orthogonal approach with AI models, which generate codon sequences from learned patterns in naturally evolved proteins.

4.4. Relaxing the Constraints on Heterologous Proteins Codon Generation

The constraint imposed in heterologous gene optimization, whereby only synonymous codons are considered, may be too strict. In fact, as Akiyama et al. (149) point out, naturally evolving orthologous genes adapt to their host organism using nonsynonymous mutations, insertions, and deletions. There are settings where it is not necessary to maintain the exact original amino acid sequence, for example, when the goal is to efficiently produce a protein with a desired function, and even more so if the protein will be further engineered in a tractable expression system. Two studies propose codon generation of related, yet possibly different amino acid sequences. Akiyama et al. (149) use a biocatalytical perspective to reframe the problem and train the deep learning model OrthologTransformer to generate a codon sequence in a target organism orthologous to a given input codon sequence in a source organism. CodonMPNN generates a codon sequence that will fold in a target organism to an input structure (169).

The generative OrthologTransformer (149) is an encoder–decoder model, trained on ∼5M pairs of orthologous genes from 2,138 bacterial species. During training, the goal is to generate, given a source species codon sequence, the orthologous codon sequence in the target species. Pretraining is on diverse bacterial species, and this diversity is advantageous: When testing on cases with known orthologs in the target species, models trained on more organisms generate sequences more similar to the evolutionarily selected ones. Finally, they fine-tune on data from specific pairs of source and target species.

OrthologTransformer was used to generate codon sequences for a PETase protein from *Ideonella sakaiensis* to be expressed in *Bacillus subtilis*, where there is no natural ortholog. Akiyama et al. (149) validated that the proteins were expressed by quantifying the mRNA level with quantitative real-time PCR (polymerase chain reaction), the protein production and export out of *B. subtilis* cells with western blot, and the activity by quantifying the PET (polyethylene terephthalate) degradation product with HPLC (high-performance liquid chromatography) and scanning

electron microscopy assay measuring the MHET [mono(2-hydroxyethyl) terephthalate] levels that are the primary hydrolysis product (149). Their experiments included 12 codon sequences generated with OrthologTransformer (see below on sampling from generative models) and, as controls, the sequence of the wild-type codons in *I. sakaiensis* and the sequence of most frequent codons in *B. subtilis*. Their approach is successful: One of their generated orthologs dramatically improved PETase expression and activity in *B. subtilis*.

Stark et al. (169) further relax the constraints with CodonMPNN and specify only the target structure and organism. Their formulation is a redefined version of the classic protein design problem (2). Rather than the established two-step approach—the first for generating the amino acid sequence that will fold to the desired structure and the second for a codon sequence that will lead to efficient expression of the protein in the target species—they propose a single step to generate the codon sequence directly. Removing the intermediate constraint possibly renders the challenge easier. They trained a model with an encoder–decoder architecture and graph neural network layers modeling the protein structure. Also, rather than listing one token per organism, they cluster the organisms based on the taxonomy map to pool the data into larger, equally sized sets. The formulations and models by Akiyama et al. (149) and Stark et al. (169) are a step toward unifying protein engineering with gene optimization, conceptually bringing protein design closer to how evolution operates—introducing variation at the DNA level and selecting on function.

4.5. Recoding Sequences for Naturally Occurring Proteins (in the Same Species)

Another area for optimization is the recoding of proteins expressed in their native host. This differs from the case of heterologous expression, where no evolutionarily selected codon sequence exists for the protein sequence in the target species. Here, such an encoding does exist, yet the goal is to recode it, intentionally replacing the evolutionarily selected codons to enhance a desired property, namely mRNA stability or protein yield. Indeed, natural codon sequences have been optimized by evolution for their function, which does not necessarily maximize those properties (156). Computational evaluation of recoded sequences presents a challenge because, by design, they differ from their naturally evolved counterparts. Metrics that measure rare codon conservation in functionally relevant locations (156) or discriminative networks trained on features like mRNA stability and protein expression (96) are used to evaluate designed sequences.

In mRNA-based vaccines and protein-replacement therapy, mRNA is delivered directly to the patient cell, where the protein synthesis machinery produces the encoded protein (24, 158, 159). The protein expression level affects the vaccines' potency, immunogenicity, and efficacy (96). To achieve strong immune responses with minimal dose, dose-efficient vaccines are optimized so that the delivered mRNA yields robust antigen expression (157, 158, 170–173). Protein-replacement therapy also benefits from stable mRNA that expresses proteins efficiently (24, 158, 174). In mRNA vaccines and protein-replacement therapy, the sequence of the proteins cannot be altered, and optimization is constrained to synonymous mutations. There are many optimizations: the UTR, the 5′-cap, the poly(A) tail, the presence of modified nucleotides, and the delivery system of the mRNA construct (99, 174, 175). Here, we focus on the optimization of codon recoding.

RiboDecode generates mRNA codon sequences for enhanced protein production (158). The model consists of two trained networks: the translation model to predict expression level and an mRNA stability predictor. The translation model, conditioned on tissue type, is trained on 320 paired Ribo-seq and RNA-seq datasets from 24 human tissues and cell lines, with over 10,000 mRNAs per dataset. The mRNA stability predictions are in terms of MFE, but unlike dynamic programming tools like ViennaRNA/RNAfold (176) and LinearFold (177), it is a differentiable network. The optimized codon sequence is generated iteratively, starting from the

naturally occurring sequence and taking gradient ascent steps to optimize the scores predicted by either one, or both, networks. Only synonymous codons are considered, leaving the amino acid sequence intact.

Li et al. (158) use RiboDecode for recoding in both abovementioned tasks. For vaccine development, they generate an optimized influenza hemagglutinin sequence and compare its expression to the wild-type sequence to show that it has enhanced expression in vitro, with improved initial and boosted immune responses. For protein-replacement therapy, they generate a nerve growth factor and show in a mouse model that the optimized mRNA sequence has an equivalent neuroprotection at one-fifth the dose of the unoptimized sequence.

One could also consider recoding the sequence to attenuate expression levels. Theoretically, synonymous recoding of viral proteins can reduce pathogenic fitness to create live-attenuated viruses (159). This possibility was elegantly demonstrated with the recoding of major capsid gene 10 of the bacteriophage T7: Synonymous recoding of the gene with nonpreferred codons relative to *E. coli* was inversely correlated with fitness (178). CocoVax is a web server for codon-based deoptimization of genes for live attenuated vaccines (179). The approach of CocoVax is not machine learning per se. Rather, it deoptimizes coding sequences by selecting single codons or neighboring codons that are statistically unfavored and characterizes the generated sequences by the number of mutations, dinucleotide ratios, and RNA stability predicted with ViennaRNA.

4.6. Generative Models Can Be Used to Score Codon Sequences

Beyond their capacity to generate sequences, generative models are also used to score or rank codon sequences provided to them through the likelihood score that they produce. The underlying assumption is that the trained models have learned patterns from evolutionary optimized sequences and that these patterns are likely to correlate with experimental features such as expression or fitness. This is a usage of generative models that is reminiscent of discriminative models. The latent representation of the generative model DeCodon was even fine-tuned toward the specific task of scoring mRNA stability and protein expression and indeed achieved top results. Stark et al. (169) consider a dataset of experimentally characterized fitness effects of 250 synonymous mutations in 21 different yeast genes (a subset with the most significant effect from a larger set). Conditioned on the wild-type structure of the protein, they compare CodonMPNN likelihoods of the wild-type coding sequence to the synonymous variant and show that in 72% of the cases, CodonMPNN assigns a higher likelihood to the more highly expressed sequence in the pair. Faizi et al. (156) compare the likelihoods assigned by their generative model Trias to experimental measurements of 30 synonymous GFP variants expressed in human HEK293 cells. They show that the scores from their model, trained on codon sequences from 640 vertebrate species, correlate well with GFP protein expression (measured as total fluorescence), mRNA half-life, and mean ribosome load. Naghipourfar et al. (89) show that the difference between likelihoods of wild-type and mutated sequences of the generative model DeCodon successfully ranks a set of ~48,000 pathogenic human variants from ClinVar.

5. USING ARTIFICIAL INTELLIGENCE CODON MODELS TO BETTER UNDERSTAND EVOLUTION

Studies that trained AI models on codon data provide insights into evolutionarily selected codon usage patterns. The fact that a nonoverfitted AI model can predict a property means that it could learn a useful signal from the data. Namely, the AI model identifies the presence of a pattern, though it does not reveal the pattern itself. In many studies, the focus is on identifying where such patterns occur. Because AI models can easily memorize training data, studies that reason

about the evolutionary signal must carefully separate training/validation from test data to ensure there are no proteins, or even homologous instances, in both sets [e.g., using CD-HIT (180), as in References 94 and 181]. One can even exclude the entire proteome of organisms from the training set (e.g., as in 90). We highlight different perspectives: codon patterns that can be derived from the amino acid sequence, functional properties that are more accurately derived from the codon sequences, and codon patterns that are shared among homologs.

5.1. Codon Patterns That Were Derived from the Amino Acid Sequence

Successful prediction of naturally occurring codon sequences from the amino acid sequence (back-translation in **Figure 1**) is an example of identifying such patterns. The accuracy baselines used for comparison are frequency-based predictions: the most frequent codons, or predictions based on codon–pair frequencies (182, 183). Yang et al. (181) used an LSTM model to predict the codons of four *E. coli* and a larger set for human highly expressed proteins, with modest success. Sidi et al. (94) trained transformer models presented with long amino acid sequence context (ranging from 10 to 150 residues) on proteins from two bacterial species (*E. coli*, *B. subtilis*) and two eukaryotes (*Schizosaccharomyces pombe*, *S. cerevisiae*). They predict the naturally occurring codon sequences in carefully separated datasets of several hundred proteins, with greater accuracy (and lower perplexity) than the baselines. Analyzing the properties of codon-predictable proteins reveals where more constrained, or learnable, patterns are found. Sidi et al. (94) show that the prediction accuracy in bacteria is greater, suggesting its sequences include more easily identifiable patterns. Highly expressed proteins are predicted more accurately, in agreement with earlier suggestions (25, 184, 185). In *S. cerevisiae*, there are GO functional annotations associated with proteins of more accurately predicted codons, e.g., genes related to the translation process.

The encoder–decoder model Trias was trained on the evolutionary selected eukaryotic codon sequences to study principles governing its generated sequences (156). By manually forcing the use of rare codons in the prefix of generated sequences, Faizi et al. (156) found that in this setting, Trias favors generating rarer codons. This suggests that the model learned that rare codons are not randomly distributed but rather occur in clusters, as previously reported (186).

To focus on the process of translation in yeast, Sakharova & Lareau (85) trained an AI model on a related task: predict the codon translation elongation speed (as opposed to the codon itself) from the amino acid sequence, or more specifically, from its pLM embedding. The predicted fast or slow labels were derived from the average transcriptome-wide elongation rates in ribosome profiling experiments in *S. cerevisiae*, and the authors focus on the subset of 26 codons from eight amino acids with robust elongation rates in different yeast species. They also balance the data by sampling equal numbers of slow and fast codons for each gene and amino acid type, setting the model's expected prediction at 50% for each label. Their fully connected deep network predicts labels from the ESM2 embeddings (100, 187) with an overall accuracy of 54.5%. They include an interesting control: A random forest can predict codon choice from the amino acid type, protein structure, and position, an overall lower accuracy of 51.8%. Sakharova & Lareau (85) also use their model to select cases of codon choices that are critical enough to have measurable phenotype differences in experimental growth assays compared to counterparts with synonymous mutations. Among 6,874 synonymous slow-to-fast or fast-to-slow mutations, most had no measurable effect in laboratory growth conditions, yet 3% had significant effects on fitness (137 deleterious and 67 advantageous). Interestingly, their experiment detects cases that may be due to translation speed changes (slow-to-fast or fast-to-slow mutations), as well as cases that are not related to the translation speed (slow to slow or fast to fast), as expected from the impact of the codon sequence on diverse processes.

5.2. Functional Properties Derived from Codon Sequences

Outeiral & Deane (90) argue that modeling with the 61-codon alphabet, rather than the 20–amino acid alphabet, can lead to better performance on downstream tasks. Indeed, CaLM outperforms pLMs of similar size in predicting protein solubility, subcellular localization, and function, and it outperforms all pLMs in predicting species and protein melting temperatures. CaLM is also competitive with larger pLMs in predicting transcript and protein abundance. All experiments are fivefold cross-validation after clustering the data with strict sequence identity cutoffs to ensure the removal of homologous sequences. This work includes interesting controls to demonstrate that codons carry a meaningful signal: First, the prediction networks are not merely learning the source species. They show this by training, models that are also given the taxonomic identity, and their cLM maintains its lead. Most interestingly, they train alternative models that learn from codon sequences in which the evolutionary signal is muddled by replacing, in training and test sets, an increasing proportion with synonymous mutations, i.e., ones that maintain the amino acid sequence signal. The performance of the models trained on the corrupted signal indicates that the cLMs fundamentally rely on codon-level signals to build their representations rather than simply enriching the amino acid signal. When the codon signal is corrupted, the models' predictive capacity falls below that of pLMs. The superiority of CaLM compared to all tested pLMs in predicting melting temperature is consistent with Boshar et al. (16), who showed that gLMs also outperform pLMs on this task. The authors hypothesize that gLMs outperform pLMs on this task by extracting species information, which is relevant as different proteomes have different melting point profiles. However, they demonstrate that pLMs enhanced with GC content and species information continue to underperform compared to gLMs in terms of melting point prediction, suggesting the success of gLMs relies on signals beyond species-related features. Outeiral & Deane (90) report a similar finding with their cLM. Overall, this shows that the most informative biological representation for a given task does not necessarily coincide with the biological level at which that property arises (16).

5.3. Codon Patterns That Are Shared by Homologs

AI models also have the potential of identifying patterns shared by the codon sequences of orthologs. This extends the observation that rare codon clusters are conserved in coding sequences of highly expressed orthologous proteins (161, 186, 188–190). The mBART network developed by Sidi et al. (94) was trained to predict codons of the target species from amino acids and possibly mimic the codons of an ortholog protein. Conditioning on orthologous codon sequences did not, surprisingly, improve prediction accuracy. This may indicate that coding sequences of orthologous proteins do not follow equivalent or directly transferable patterns, or at least not ones that can be learned to improve beyond only learning the patterns in the target organism. OrthologTransformer (149) tested the accuracy of predictions from orthologs in a held-out set. This evaluation considers three codon sequences: the one generated by the model and the two in the ortholog pair, the source used as the model's input and target. In all cases, the generated sequence is closer to the target sequence than it is to the source sequence, suggesting the AI model successfully learned a relevant signal. Further supporting this, the similarity between the OrthologTransformer-generated and target codon sequences (38–74% identity) is higher than that of CodonTransformer-generated sequences (29–34% identity).

6. FUTURE WORK

The full potential of AI models trained on the language of codons has not been exhausted. We list three open directions in which these can contribute: annotating genes, improving pipelines for

designing sequences with desired functions, and advancing our understanding of protein synthesis and evolution.

Current gene-prediction pipelines do not include discriminative cLM-based models. They do include indirect measures of codon bias, e.g., GC content at the third codon position (191). Awareness of meaningful relationships can be learned by cLMs directly from sequence data and incorporated in the annotation pipelines. Furthermore, as suggested by Yang et al. (181), AI models could learn to identify codon patterns of endogenous genes from recent horizontal transfer (whose codons have not yet adapted) to improve genome annotation and evolutionary inference.

A promising direction is integrating codon sequence design into AI-assisted protein design. Current de novo design strategies focus on amino acid sequence design. The coding sequence is then back-translated for the target host using generic tools as a final step. This disconnects the significant codon constraints (123–125) from the design process, often producing codon sequences that are poorly adapted to the host. CodonMPNN (169) demonstrates that codon-aware design is feasible by generating a coding sequence conditioned on the target protein backbone and the host. Guiding codon sequence generation with multiple factors, such as structure, function, expression host, or genomic context (as implemented in the gLM Evo 2) (12), will possibly lead to the design of valid coding sequences that are biologically and functionally optimized for the target host expression system.

Codon choice can influence cotranslational folding by modulating translation kinetics. Codon-level AI models may therefore offer new ways to explore the coupling between sequence and folding. In principle, cLMs could relate coding sequences to their corresponding folded structures, revealing structural features embedded in the translation process. Moreover, given folding-relevant and large datasets, AI models could be leveraged to train models directly on data. To date, small datasets linking nonsynonymous mutations to folding outcomes exist, but no large-scale resource relates codon usage to cotranslational folding dynamics. Using AI to model the language of codons holds the promise of revealing how evolution has encoded multiple layers of regulatory, structural, and functional information within the very sequences that define life.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

R.K. was supported by Israel Science Foundation (ISF) grant 1764/21. C.L.-S.-D. was supported by the Zuckerman STEM Leadership Program. R.K. and C.L.-S.-D. were supported by the Data Science Research Center at the University of Haifa.

LITERATURE CITED

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–89
2. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, et al. 2022. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 378:49–56
3. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, et al. 2025. Simulating 500 million years of evolution with a language model. *Science* 387:850–58
4. Yanai G, Axel G, Longo LM, Ben-Tal N, Kolodny R. 2025. Contrastive learning unites sequence and structure in a global representation of protein space. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2025.09.05.674454v4**

5. Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Mirdita M, et al. 2024. Bilingual language model for protein sequence and structure. *NAR Genom. Bioinform.* 6:lqae150
6. Lee JS, Abdin O, Kim PM. 2025. Language models for protein design. *Curr. Opin. Struct. Biol.* 92:103027
7. Heinzinger M, Rost B. 2025. Teaching AI to speak protein. *Curr. Opin. Struct. Biol.* 91:102986
8. Ünlü A, Ulusoy E, Yiğit MG, Darcan M, Doğan T. 2025. Protein language models for predicting drug–target interactions: novel approaches, emerging methods, and future directions. *Curr. Opin. Struct. Biol.* 91:103017
9. Bjerregaard A, Groth PM, Hauberg S, Krogh A, Boomsma W. 2025. Foundation models of protein sequences: a brief overview. *Curr. Opin. Struct. Biol.* 91:103004
10. Harding-Larsen D, Funk J, Madsen NG, Gharabli H, Acevedo-Rocha CG, et al. 2024. Protein representations: encoding biological information for machine learning in biocatalysis. *Biotechnol. Adv.* 77:108459
11. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, et al. 2024. Sequence modeling and design from molecular to genome scale with Evo. *Science* 386:eado9336
12. Brixi G, Durrant MG, Ku J, Naghipourfar M, Poli M, et al. 2026. Genome modelling and design across all domains of life with Evo 2. *Nature* 652:1349–61. **https://doi.org/10.1038/s41586-026-10176-5**
13. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, et al. 2025. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* 22:287–97
14. Benegas G, Ye C, Albors C, Li JC, Song YS. 2025. Genomic language models: opportunities and challenges. *Trends Genet.* 41:286–302
15. Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, et al. 2025. Transformers and genome language models. *Nat. Mach. Intell.* 7:346–62
16. Boshar S, Trop E, de Almeida BP, Copoiu L, Pierrot T. 2024. Are genomic language models all you need? Exploring genomic language models on protein downstream tasks. *Bioinformatics* 40:btae529
17. Grantham R, Gautier C, Gouy M. 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.* 8:1893–912
18. Carbone A, Képès F, Zinovyev A. 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Mol. Biol. Evol.* 22:547–61
19. Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* 12:32–42
20. Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. *J. Biol. Chem.* 262:11351–55
21. Hanson G, Coller J. 2018. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* 19:20–30
22. Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136:927–35
23. Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, et al. 2015. Codon optimality is a major determinant of mRNA stability. *Cell* 160:1111–24
24. Paremskaia AI, Kogan AA, Murashkina A, Naumova DA, Satish A, et al. 2024. Codon-optimization in gene therapy: promises, prospects and challenges. *Front. Bioeng. Biotechnol.* 12:1371596
25. Bahiri-Elitzur S, Tuller T. 2021. Codon-based indices for modeling gene expression and transcript evolution. *Comput. Struct. Biotechnol. J.* 19:2646–63
26. Hia F, Takeuchi O. 2021. The effects of codon bias and optimality on mRNA and protein regulation. *Cell. Mol. Life Sci.* 78:1909–28
27. Zhou Z, Dang Y, Zhou M, Li L, Yu C-h, et al. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *PNAS* 113:E6117–25
28. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLOS Biol.* 4:e180
29. Fu J, Dang Y, Counter C, Liu Y. 2018. Codon usage regulates human KRAS expression at both transcriptional and translational levels. *J. Biol. Chem.* 293:17929–40
30. Harigaya Y, Parker R. 2016. Analysis of the association between codon optimality and mRNA stability in *Schizosaccharomyces pombe*. *BMC Genom.* 17:895

31. Radrizzani S, Kudla G, Izsvák Z, Hurst LD. 2024. Selection on synonymous sites: the unwanted transcript hypothesis. *Nat. Rev. Genet.* 25:431–48
32. Bergman S, Tuller T. 2020. Widespread non-modular overlapping codes in the coding regions. *Phys. Biol.* 17:031002
33. Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* 3:285–98
34. Geisberg Joseph V, Moqtaderi Z, Fan X, Ozsolak F, Struhl K. 2014. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* 156:812–24
35. Jia L, Mao Y, Ji Q, Dersh D, Yewdell JW, Qian S-B. 2020. Decoding mRNA translatability and stability from the 5′ UTR. *Nat. Struct. Mol. Biol.* 27:814–21
36. Sachs AB. 1993. Messenger RNA degradation in eukaryotes. *Cell* 74:413–21
37. Mayr C. 2017. Regulation by 3′-untranslated regions. *Annu. Rev. Genet.* 51:171–94
38. Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, et al. 2019. mRNA structure regulates protein expression through changes in functional half-life. *PNAS* 116:24075–83
39. Bernstein JA, Khodursky AB, Lin P-H, Lin-Chao S, Cohen SN. 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *PNAS* 99:9697–702
40. Yang E, Van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, et al. 2003. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* 13:1863–72
41. Bae H, Coller J. 2022. Codon optimality-mediated mRNA degradation: linking translational elongation to mRNA stability. *Mol. Cell* 82:1467–76
42. Yang Q, Yu C-H, Zhao F, Dang Y, Wu C, et al. 2019. eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. *Nucleic Acids Res.* 47:9243–58
43. Mercier BC, Labaronne E, Cluet D, Guiguettaz L, Fontrodona N, et al. 2024. Translation-dependent and -independent mRNA decay occur through mutually exclusive pathways defined by ribosome density during T cell activation. *Genome Res.* 34:394–409
44. Bicknell AA, Reid DW, Licata MC, Jones AK, Cheng YM, et al. 2024. Attenuating ribosome load improves protein output from mRNA by limiting translation-dependent mRNA decay. *Cell Rep.* 43:114098
45. Roy B, Jacobson A. 2013. The intimate relationships of mRNA decay and translation. *Trends Genet.* 29:691–99
46. Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9:133–48
47. Zuker M. 1994. Prediction of RNA secondary structure by energy minimization. *Methods Mol. Biol.* 25:267–94
48. Leppek K, Byeon GW, Kladwang W, Wayment-Steele HK, Kerr CH, et al. 2022. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat. Commun.* 13:1536
49. Liu Y, Yang Q, Zhao F. 2021. Synonymous but not silent: the codon usage code for gene expression and protein folding. *Annu. Rev. Biochem.* 90:375–401
50. Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. 2016. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep.* 14:1787–99
51. Torrent M, Chalancon G, De Groot NS, Wuster A, Madan Babu M. 2018. Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Sci. Signal.* 11:eaat6409
52. Gingold H, Tehler D, Christoffersen Nanna R, Nielsen Morten M, Asmar F, et al. 2014. A dual program for translation regulation in cellular proliferation and differentiation. *Cell* 158:1281–92
53. Sagi D, Rak R, Gingold H, Adir I, Maayan G, et al. 2016. Tissue- and time-specific expression of otherwise identical tRNA genes. *PLOS Genet.* 12:e1006264
54. Goodarzi H, Nguyen HCB, Zhang S, Dill BD, Molina H, Tavazoie SF. 2016. Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell* 165:1416–27
55. Sørensen MA, Kurland C, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.* 207:365–77

56. Pedersen S. 1984. *Escherichia coli* ribosomes translate in vivo with variable rate. *EMBO J.* 3:2895–98
57. Stoletzki N, Eyre-Walker A. 2006. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* 24:374–81
58. Gilchrist MA, Shah P, Zaretzki R. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics* 183:1493–505
59. Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–52
60. Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *J. Mol. Evol.* 52:275–80
61. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* 12:R110
62. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141:344–54
63. Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, et al. 2019. A short translational ramp determines the efficiency of protein synthesis. *Nat. Commun.* 10:5774
64. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin Joshua B. 2013. Rate-limiting steps in yeast protein translation. *Cell* 153:1589–601
65. Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, et al. 2014. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.* 10:770
66. Ceroni F, Boo A, Furini S, Gorochowski TE, Borkowski O, et al. 2018. Burden-driven feedback control of gene expression. *Nat. Methods* 15:387–93
67. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *PNAS* 115:E4940–49
68. Dreher TW. 2010. Viral tRNAs and tRNA-like structures. *WIREs RNA* 1:402–14
69. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W. 2003. Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* 67:86–156
70. Dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32:5036–44
71. Dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.* 31:6976–85
72. Man O, Sussman JL, Pilpel Y. 2006. Examination of the tRNA adaptation index as a predictor of protein expression levels. In *Systems Biology and Regulatory Genomics*, ed. E Eskin, T Ideker, B Raphael, C Workman. Springer
73. Stein KC, Frydman J. 2019. The stop-and-go traffic regulating protein biogenesis: how translation kinetics controls proteostasis. *J. Biol. Chem.* 294:2076–84
74. Holtkamp W, Kokic G, Jäger M, Mittelstaet J, Komar AA, Rodnina MV. 2015. Cotranslational protein folding on the ribosome monitored in real time. *Science* 350:1104–7
75. Cortazzo P, Cerveñansky C, Marín M, Reiss C, Ehrlich R, Deana A. 2002. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 293:537–41
76. Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM. 2010. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J. Mol. Biol.* 396:1310–18
77. Nedialkova Danny D, Leidel Sebastian A. 2015. Optimization of codon translation rates via tRNA modifications maintains proteome integrity. *Cell* 161:1606–18
78. Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* 16:274
79. Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, et al. 2016. Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell* 61:341–51
80. Sander IM, Chaney JL, Clark PL. 2014. Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. *J. Am. Chem. Soc.* 136:858–61
81. Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu. Rev. Genet.* 42:287–99
82. Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. B* 365:1203–12

83. Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* 12:683–91
84. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156:1324–35
85. Sakharova H, Lareau LF. 2025. Protein language models reveal evolutionary constraints on synonymous codon choice. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2025.08.05.668603v1**
86. Rauscher R, Bampi GB, Guevara-Ferrer M, Santos LA, Joshi D, et al. 2021. Positive epistasis between disease-causing missense mutations and silent polymorphism with effect on mRNA translation velocity. *PNAS* 118:e2010612118
87. Moss MJ, Chamness LM, Clark PL. 2024. The effects of codon usage on protein structure and folding. *Annu. Rev. Biophys.* 53:87–108
88. Yazdani-Jahromi M, Prakash M, Mansi T, Moskalev A, Liao R. 2025. HELM: hierarchical encoding for mRNA language modeling. Preprint, arXiv:2410.12459v2 [cs.LG]
89. Naghipourfar M, Chen S, Howard M, Macdonald C, Saberi A, et al. 2024. A suite of foundation models captures the contextual interplay between codons. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2024.10.10.617568v1**
90. Outeiral C, Deane CM. 2024. Codon language embeddings provide strong signals for use in protein engineering. *Nat. Mach. Intell.* 6:170–79
91. Fallahpour A, Gureghian V, Filion GJ, Lindner AB, Pandi A. 2025. CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nat. Commun.* 16:3205
92. Constant DA, Gutierrez JM, Sastry AV, Viazzo R, Smith NR, et al. 2023. Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2023.02.11.528149v1**
93. Greener JG, Kandathil SM, Moffat L, Jones DT. 2022. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23:40–55
94. Sidi T, Bahiri-Elitzur S, Tuller T, Kolodny R. 2025. Predicting gene sequences with AI to study codon usage patterns. *PNAS* 122:e2410003121
95. Heuschkel J, Kingsley L, Pefaur N, Nixon A, Cramer S. 2025. Advancing codon language modeling with synonymous codon constrained masking. *Nucleic Acids Res.* 54:gkag166
96. Li S, Moayedpour S, Li R, Bailey M, Riahi S, et al. 2024. CodonBERT large language model for mRNA vaccines. *Genome Res.* 34:1027–35
97. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9:2579–605
98. McInnes L, Healy J, Melville J. 2020. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint, arXiv:1802.03426v3 [stat.ML]
99. Li S, Noroozizadeh S, Moayedpour S, Kogler-Anele L, Xue Z, et al. 2025. mRNA-LM: full-length integrated SLM for mRNA analysis. *Nucleic Acids Res.* 53:gkaf044
100. Rives A, Meier J, Sercu T, Goyal S, Lin Z, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS* 118:e2016239118
101. Sabath N, Wagner A, Karlin D. 2012. Evolution of viral proteins originated de novo by overprinting. *Mol. Biol. Evol.* 29:3767–80
102. Zhang H, Gao X, Zhang J, Lai L. 2024. mRNA2vec: mRNA embedding with language model in the 5'UTR-CDS for mRNA design. Preprint, arXiv:2408.09048v2 [q-bio.QM]
103. Baevski A, Hsu W-N, Xu Q, Babu A, Gu J, Auli M. 2022. data2vec: a general framework for self-supervised learning in speech, vision and language. Preprint, arXiv:2202.03555v3 [cs.LG]
104. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, et al. 2019. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, Vol. 32, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett. Curran
105. Shen Y, Chen Z, Mamalakis M, He L, Xia H, et al. 2024. A fine-tuning dataset and benchmark for large language models for protein understanding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE
106. Dallago C, Mou J, Johnston KE, Wittmann BJ, Bhattacharya N, et al. 2022. FLIP: benchmark tasks in fitness landscape inference for proteins. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2021.11.09.467890v2**

107. Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10:57–63
108. Chen S, Li K, Cao W, Wang J, Zhao T, et al. 2017. Codon-resolution analysis reveals a direct and context-dependent impact of individual synonymous mutations on mRNA level. *Mol. Biol. Evol.* 34:2944–58
109. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. 2012. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41:D991–95
110. Moreno P, Fexova S, George N, Manning JR, Miao Z, et al. 2021. Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.* 50:D129–40
111. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. 2021. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* 50:D387–90
112. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, et al. 2013. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42:D699–704
113. Agarwal V, Kelley DR. 2022. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol.* 23:245
114. Diez M, Medina-Muñoz SG, Castellano LA, da Silva Pescador G, Wu Q, Bazzini AA. 2022. iCodon customizes gene expression based on the codon composition. *Sci. Rep.* 12:12126
115. Wayment-Steele HK, Kladwang W, Watkins AM, Kim DS, Tunguz B, et al. 2022. Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nat. Mach. Intell.* 4:1174–84
116. Mittal P, Brindle J, Stephen J, Plotkin JB, Kudla G. 2018. Codon usage influences fitness through RNA toxicity. *PNAS* 115:8639–44
117. Ingolia NT, Hussmann JA, Weissman JS. 2019. Ribosome profiling: global views of translation. *Cold Spring Harb. Perspect. Biol.* 11:a032698
118. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, et al. 2022. Standardized annotation of translated open reading frames. *Nat. Biotechnol.* 40:994–99
119. Tunney R, McGlincy NJ, Graham ME, Naddaf N, Pachter L, Lareau LF. 2018. Accurate design of translational output by a neural network model of ribosome distribution. *Nat. Struct. Mol. Biol.* 25:577–82
120. Hussmann JA, Patchett S, Johnson A, Sawyer S, Press WH. 2015. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLOS Genet.* 11:e1005732
121. Diament A, Tuller T. 2016. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol. Direct.* 11:24
122. Bartholomäus A, Del Campo C, Ignatova Z. 2016. Mapping the non-standardized biases of ribosome profiling. *Biol. Chem.* 397:23–35
123. Singer GAC, Hickey DA. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol. Biol. Evol.* 17:1581–88
124. Lightfield J, Fram NR, Ely B. 2011. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLOS ONE* 6:e17677
125. Li J, Zhou J, Wu Y, Yang S, Tian D. 2015. GC-content of synonymous codons profoundly influences amino acid usage. *G3* 5:2027–36
126. Mathieu A, Leclercq M, Sanabria M, Perin O, Droit A. 2022. Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. *Front. Microbiol.* 13:811495
127. Kim J, Steinegger M. 2024. Metabuli: sensitive and specific metagenomic classification via joint analysis of amino acid and DNA. *Nat. Methods* 21:971–73
128. Schmitz A, Zhang F. 2021. Massively parallel gene expression variation measurement of a synonymous codon library. *BMC Genom.* 22:149
129. Nieuwkoop T, Terlouw BR, Stevens KG, Scheltema Richard A, de Ridder D, et al. 2023. Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* 51:2363–76
130. Zhang H, Zhang L, Lin A, Xu C, Li Z, et al. 2023. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* 621:396–403
131. Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA, Shin C, et al. 2014. mRNA destabilization is the dominant effect of mammalian MicroRNAs by the time substantial repression ensues. *Mol. Cell* 56:104–15

132. Wang D, Eraslan B, Wieland T, Hallström B, Hopf T, et al. 2019. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* 15:e8503
133. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, et al. 2015. Tissue-based map of the human proteome. *Science* 347:1260419
134. Boël G, Letso R, Neely H, Price WN, Wong K-H, et al. 2016. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* 529:358–63
135. Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, et al. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteom.* 11:492–500
136. Guo T, Steen JA, Mann M. 2025. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature* 638:901–11
137. Rubin AF, Stone J, Bianchi AH, Capodanno BJ, Da EY, et al. 2025. MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays. *Genome Biol.* 26:13
138. Landrum MJ, Kattman BL. 2018. ClinVar at five years: delivering on the promise. *Hum. Mutat.* 39:1623–30
139. Sridharan S, Kurzawa N, Werner T, Günthner I, Helm D, et al. 2019. Proteome-wide solubility and thermal stability profiling reveals distinct regulatory roles for ATP. *Nat. Commun.* 10:1155
140. Vihinen M. 2020. Solubility of proteins. *ADMET DMPK* 8:391–99
141. Jarzab A, Kurzawa N, Hopf T, Moerch M, Zecha J, et al. 2020. Meltome atlas—thermal proteome stability across the tree of life. *Nat. Methods* 17:495–503
142. Terpe K. 2006. Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* 72:211–22
143. Tripathi NK, Shrivastava A. 2019. Recent developments in bioprocessing of recombinant proteins: expression hosts and process development. *Front. Bioeng. Biotechnol.* 7:420
144. Jefferis R. 2016. Posttranslational modifications and the immunogenicity of biotherapeutics. *J. Immunol. Res.* 2016:5358272
145. Fu H, Liang Y, Zhong X, Pan Z, Huang L, et al. 2020. Codon optimization with deep learning to enhance protein expression. *Sci. Rep.* 10:17617
146. Jain R, Jain A, Mauro E, LeShane K, Densmore D. 2023. ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinform.* 24:132
147. Han X, Shao X, Shuo L, Shi Z, Huang R, et al. 2025. DeepCodon: a deep learning codon-optimization model to enhance protein expression. *BioDesign Res.* 7:100042
148. Ravi S, Sharma T, Yip M, Yang H, Xie J, et al. 2025. A deep learning model trained on expressed transcripts across different tissue types reveals cell-type codon-optimization preferences. *Nucleic Acids Res.* 53:gkaf233
149. Akiyama M, Tashiro M, Huang Y, Uehara M, Kanzaki T, et al. 2025. Generative AI-driven artificial DNA design for enhancing inter-species gene activation and enzymatic degradation of PET. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2025.05.08.652991v1**
150. Rodriguez A, Wright G, Emrich S, Clark PL. 2018. %MinMax: a versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. *Protein Sci.* 27:356–62
151. Man O, Pilpel Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.* 39:415–21
152. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–58
153. Zaytsev K, Bogatyreva N, Fedorov A. 2024. Exploring individual codon influence on protein expression: a predictive approach. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2024.06.30.601029v1**
154. Jansen R, Bussemaker HJ, Gerstein M. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res.* 31:2242–51
155. Simhadri VL, Hamasaki-Katagiri N, Lin BC, Hunt R, Jha S, et al. 2017. Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. *J. Med. Genet.* 54:338–45

156. Faizi M, Sakharova H, Lareau LF. 2025. A generative language model decodes contextual constraints on codon choice for mRNA design. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2025.05.13.653614v2**
157. Patel S, Tang S, Zhang Y, Chatterjee P, Yao S. 2025. Multi-objective-guided generative design of mRNA with therapeutic properties. *Paper presented at ICML 2025 Generative AI and Biology (GenBio) Workshop*
158. Li Y, Wang F, Yang J, Han Z, Chen L, et al. 2024. Deep generative optimization of mRNA codon sequences for enhanced protein production and therapeutic efficacy. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2024.09.06.611590v1**
159. Lin BC, Kaissarian NM, Kimchi-Sarfaty C. 2023. Implementing computational methods in tandem with synonymous gene recoding for therapeutic development. *Trends Pharmacol. Sci.* 44:73–84
160. Demissie EA, Park S-Y, Moon JH, Lee D-Y. 2025. Comparative analysis of codon optimization tools: advancing toward a multi-criteria framework for synthetic gene design. *J. Microbiol. Biotechnol.* 35:e2411066
161. Wright G, Rodriguez A, Li J, Milenkovic T, Emrich SJ, Clark PL. 2022. CHARMING: harmonizing synonymous codon usage to replicate a desired codon usage pattern. *Protein Sci.* 31:221–31
162. Hernandez-Alias X, Benisty H, Radusky LG, Serrano L, Schaefer MH. 2023. Using protein-per-mRNA differences among human tissues in codon optimization. *Genome Biol.* 24:34
163. Eguchi Y, Makanae K, Hasunuma T, Ishibashi Y, Kito K, Moriya H. 2018. Estimating the protein burden limit of yeast cells by measuring the expression limits of glycolytic proteins. *eLife* 7:e34595
164. Moriya H. 2015. Quantitative nature of overexpression experiments. *Mol. Biol. Cell* 26:3932–39
165. Stoebel DM, Dean AM, Dykhuizen DE. 2008. The cost of expression of *Escherichia coli lac* operon proteins is in the process, not in the products. *Genetics* 178:1653–60
166. Bolognesi B, Lehner B. 2018. Reaching the limit. *eLife* 7:e39804
167. Mauro VP. 2018. Codon optimization in the production of recombinant biotherapeutics: potential risks and considerations. *BioDrugs* 32:69–81
168. Love AM, Nair NU. 2024. Specific codons control cellular resources and fitness. *Sci. Adv.* 10:eadk3485
169. Stark H, Padia U, Balla J, Diao C, Church G. 2024. CodonMPNN for organism specific and codon optimal inverse folding. Preprint, arXiv:2409.17265v1 [cs.LG]
170. Castillo-Hair SM, Seelig G. 2021. Machine learning for designing next-generation mRNA therapeutics. *Acc. Chem. Res.* 55:24–34
171. Verbeke R, Lentacker I, De Smedt SC, Dewitte H. 2021. The dawn of mRNA vaccines: the COVID-19 case. *J. Control. Release* 333:511–20
172. Kim SC, Sekhon SS, Shin W-R, Ahn G, Cho B-K, et al. 2022. Modifications of mRNA vaccine structural elements for improving mRNA stability and translation efficiency. *Mol. Cell. Toxicol.* 18:1–8
173. Andreas NK, Tim B, Petra S, Britta V, Janina B, et al. 2012. mRNA as a versatile tool for exogenous protein expression. *Curr. Gene Ther.* 12:347–61
174. Morrow AK, Thornal A, Flynn ED, Hoelzli E, Shan M, et al. 2025. ML-driven design of 3' UTRs for mRNA stability. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2024.10.07.616676v2**
175. Gote V, Bolla PK, Kommineni N, Butreddy A, Nukala PK, et al. 2023. A comprehensive review of mRNA vaccines. *Int. J. Mol. Sci.* 24:2700
176. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA Websuite. *Nucleic Acids Res.* 36:W70–74
177. Huang L, Zhang H, Deng D, Zhao K, Liu K, et al. 2019. LinearFold: linear-time approximate RNA folding by 5′-to-3′ dynamic programming and beam search. *Bioinformatics* 35:i295–304
178. Bull J, Molineux I, Wilke C. 2012. Slow fitness recovery in a codon-modified viral genome. *Mol. Biol. Evol.* 29:2997–3004
179. Li J, Shuai S. 2025. CocoVax: a web server for codon-based deoptimization of viral genes in live attenuated vaccine design. *Nucleic Acids Res.* 53:W178–86
180. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–59
181. Yang DK, Goldman SL, Weinstein E, Marks D. 2019. Generative models for codon prediction and optimization. In *MLCB 2019 Proceedings*

182. Moura G, Pinheiro M, Arrais J, Gomes AC, Carreto L, et al. 2007. Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLOS ONE* 2:e847
183. Harigaya Y, Parker R. 2017. The link between adjacent codon pairs and mRNA stability. *BMC Genom.* 18:364
184. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* 8:197
185. Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10:7055–74
186. Clarke TFIV, Clark PL. 2008. Rare codons cluster. *PLOS ONE* 3:e3412
187. Lin Z, Akin H, Rao R, Hie B, Zhu Z, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379:1123–30
188. Chaney JL, Steele A, Carmichael R, Rodriguez A, Specht AT, et al. 2017. Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLOS Comput. Biol.* 13:e1005531
189. Jacobs WM, Shakhnovich EI. 2017. Evidence of evolutionary selection for cotranslational folding. *PNAS* 114:11434
190. Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20:237
191. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* 11:119
192. The UniProt Consortium. 2025. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* 53:D609–17