PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Predicting gene sequences with AI to study codon usage patterns

Tomer Sidi[a], Shir Bahiri-Elitzur[b], Tamir Tuller[b,c,1], and Rachel Kolodny[a,1]

Affiliations are included on p. 11.

Edited by Michael Levitt, Stanford University, Stanford, CA; received May 23, 2024; accepted November 27, 2024

Selective pressure acts on the codon use, optimizing multiple, overlapping signals that are only partially understood. We trained AI models to predict codons given their amino acid sequence in the eukaryotes *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* and the bacteria *Escherichia coli* and *Bacillus subtilis* to study the extent to which we can learn patterns in naturally occurring codons to improve predictions. We trained our models on a subset of the proteins and evaluated their predictions on large, separate sets of proteins of varying lengths and expression levels. Our models significantly outperformed naïve frequency-based approaches, demonstrating that there are learnable dependencies in evolutionary-selected codon usage. The prediction accuracy advantage of our models is greater for highly expressed genes and is greater in bacteria than eukaryotes, supporting the hypothesis that there is a monotonic relationship between selective pressure for complex codon patterns and effective population size. In *S. cerevisiae* and bacteria, our models were more accurate for longer proteins, suggesting that the learned patterns may be related to cotranslational folding. Gene functionality and conservation were also important determinants that affect the performance of our models. Finally, we showed that using information encoded in homologous proteins has only a minor effect on prediction accuracy, perhaps due to complex codon-usage codes in genes undergoing rapid evolution. Our study employing contemporary AI methods offers a unique perspective and a deep-learning-based prediction tool for evolutionary-selected codons. We hope that these can be useful to optimize codon usage in endogenous and heterologous proteins.

codons prediction | codon AI model | mimicking codons

Significance

Can one predict codon sequences used by an organism to encode a given amino acid sequence? Codon frequencies vary, a phenomenon known as codon bias, yet we improve upon frequency-based predictions using contemporary AI tools that learn complex patterns and capture interactions between codons. Because our predictions are tested fairly, on cases not seen during the training process, accurate predictions suggest that these learned patterns are not random and may be related to the evolutionary process. Our AI model for predicting codons is publicly available. That we can better learn codon patterns in high-expression proteins, and in sequences related to housekeeping cellular processes, suggests that the coding sequences of these genes are populated with complex regulatory codes.

Although in all known organisms, there are 61 codons that each encode one of the 20 amino acids, 18 amino acids are encoded by multiple (two to six) codons. Thus, there are many codon sequences that encode the same amino acid sequence. The selection of codons used impacts protein production, indirectly, by influencing the availability of transfer RNA and free ribosomes in the cell (1–4), and directly by influencing messenger RNA (mRNA) structure and stability (3, 5), transcription (6), splicing (7), and translation kinetics (8–10). Translation kinetics, in turn, influence protein cotranslational folding and regulation (9, 11, 12).

Certain codons are preferred over others, a phenomenon known as codon bias (1, 11), and this bias differs not only across species (9) but also within species in a manner that depends on expression level and protein length (5, 11, 13). Codon usage even differs along individual genes. For example, there are different usage patterns at the beginning of coding regions vs. protein domain boundaries (14, 15). Use of rare codons can slow translation (11, 16) and enable the generation of functionally or structurally stable proteins (3). Because codon usage influences efficiency and accuracy of protein synthesis, it was recognized as a code within the genetic code that is subject to evolutionary selection (3, 13).

Computationally predicting the codon encodings of proteins in different organisms holds practical value, even in studies like ours that do not emulate the evolutionary process. Codon usage can have a significant effect on protein levels in different organisms (17–20). Production of heterologous (nonhost) proteins for use in protein science and biotechnology (21), in bacterial cell factories (22), as vaccines (23), or for agricultural purposes (24) requires optimization of the codon sequence. The disparities in codon biases between the original and host organisms necessitate adjusting the codons to the biases of the new host. Indeed, dozens of measures aim to model codon usage patterns (reviewed in ref. 25). These measures, however, are usually limited to capturing local statistics of codon distribution.

To rigorously approach this prediction challenge, one should establish distinct training and test sets (26). Two meaningful baseline models that were used in previous studies are the naïve Bayes predictor, which predicts the most frequently used codon for each amino

Author contributions: T.S., T.T., and R.K. designed research; T.S. and R.K. performed research; T.S. and S.B.-E. contributed new reagents/analytic tools; T.S., T.T., and R.K. analyzed data; and T.T. and R.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: tamirtul@tauex.tau.ac.il or trachel@cs.haifa.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2410003121/-/DCSupplemental.

Published December 31, 2024.

acid (27) and the bigram frequency model which predicts the most frequently used codon, conditioned on its preceding codon (28, 29). These frequencies can be estimated from the training set. Sen et al. use the codon and codon pairs frequencies as parameters in a mixed integer linear programming formulation and optimize it with a commercial solver (30). There may be patterns of interactions among codons, that are more distant in primary sequence, and learning these may yield even better predictions (3, 31). To learn codon patterns with the limited data available per organism, sophisticated tools that learn data distributions are needed (32).

Indeed, deep networks, and transformers in particular (32), have emerged as the tool of choice to learn the complex distributions characterizing codon usage. Studies suggest that there is a signal within codon sequences that deep networks can harvest: Tunney et al. used a feed-forward network to successfully predict the ribosome density from the sequence neighborhood (33), and several studies have demonstrated the utility of deep networks (RNN, T5, and BiLSTM) to optimize the gene expression levels of two to four proteins (34–36), and predict other aspects of gene expression (37–40). Recently, Outeiral and Deane trained a language model for proteins, in which the input to the model is codons. They show that the embeddings calculated by their language model are more useful on different tasks, compared to the state-of-the-art protein language models whose inputs are amino acids. They argue that the advantage of their model is due to the richer information content of the codon sequence representation (41). Three networks that are close in spirit to our models optimize codon usage directly for specific organisms. CodonBERT is a transformer-based network that uses cross-attention and trained on the human protein atlas (42). Yang et al. used deep networks (BiLSTM) to predict codon sequences of highly expressed proteins in humans and *Escherichia coli*; their predictions failed to improve upon the frequency-based baseline (26). ICOR is a recurrent neural network (RNN) that learned the codon usage biases of *E. coli* (43).

Knowing the codons used in an orthologous protein in another organism may aid in prediction. For example, orthologous proteins may have a codon at a particular position, that differs due to usage bias, but that functions to induce translation pausing that is crucial for proper protein folding. If codons with unique functions can be identified and mimicked, the prediction may be more accurate (8, 22). In support of the utility of this information, it was shown that position-dependent clusters of optimal and nonoptimal codons are conserved among orthologous proteins (44, 45) and that codon usage relates to protein structure with nonoptimal codons aligning domain boundaries (8, 9, 11, 46). Indeed, these observations underlie the design of codon harmonization tools that predict a codon sequence for optimal incorporation of a given amino acid sequence in a non-native organism. Codon harmonization tools are not, however, designed for orthologous proteins but rather for those with the same amino acid sequence (47, 48). In other words, they are not intended to predict the evolutionarily selected codons in one organism based on the codons of a different, albeit orthologous, protein. In practical terms, this means that codon harmonization tools mimic readily described properties like frequency rank (47, 48).

Here, we take a data-driven approach to investigate the naturally occurring codon sequences in four organisms: the eukaryotes *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* and the bacteria *E. coli* and *Bacillus subtilis*. We explored two scenarios: predicting the codon sequence from amino acid sequences and predicting the codon sequence in one organism given the codon sequence of an orthologous protein in another organism. We used mBART, a transformer-based encoder–decoder architecture that extends BART, developed for a single natural language (49, 50). mBART learns a shared model for multiple languages (e.g., English and French) allowing it both generate text in these languages and convert (i.e., translate) text from one language to another. In our setting, the analog of a language is an organism, and our mBART-trained models can both generate codon sequences for multiple organisms and mimic the codons of an orthologous protein. That our best models outperformed the frequency-based baselines suggests that there are patterns of codon interactions between residues (including non-neighboring residues) that can be learned. The extent to which the AI model improves predictions—varies with expression level and protein length and across organisms, and suggests where complex patterns in natural proteins are evolutionarily favored. Furthermore, in *S. cerevisiae* and *E. coli*, we compared the accuracy gain of our model to the frequency-based baseline for functional sets of proteins grouped by Gene Ontology (GO) annotations and found that accuracy gain was higher than average for some molecular functions and biological processes. The AI tool introduced here, with publicly available code and an easy-to-use web interface, will enable future investigations related to evolutionary selection of codon sequences.

Results

Training of mBART Models to Predict Codons from Amino Acid Sequence Using Masking and Mimicking. We consider two tasks: masking, which is prediction of codons from the amino acid sequence, and mimicking, which is prediction of codons based on codons of an ortholog protein in another organism. The rationale for the mimicking mode is that the rate of translation elongation depends on the codon used (51), and the nonuniform rate may be important for cotranslational protein folding (8, 52, 53). Thus, as codons of orthologous proteins in two organisms may encode similar elongation rates, the codon sequence of the orthologous protein may be useful in prediction of codon usage for a protein of interest.

We trained several mBART models (49, 50) to support masking and mimicking tasks in four well-studied model organisms: *S. cerevisiae, S. pombe., E. coli,* and *B. subtilis.* Following standard machine learning practices, we divided the protein data from these organisms into three distinct sets: ~70% in the training set, ~10% in the validation set, and ~20% in the test set. All three sets included proteins with a wide range of expression levels. At the amino acid level, none of the proteins in the test set were closely related to those in the training set [based on amino acid sequence clustering using CD-HIT (54) with a threshold of 0.7]. The test set included 1,240 *S. cerevisiae*, 1,024 *S. pombe*, 812 *E. coli*, and 855 *B. subtilis* proteins (of which all but 496, 463, 271, and 247, respectively, have measured expression levels). Thus, evaluation of the models was conducted under stringent conditions: The predictions are evaluated with respect to the evolutionary-selected codons of the test set proteins, and these included a significant number of proteins with a wide range of expression levels and lengths. Also, the test set proteins are different from the training set proteins at the amino acid level (see *Methods* for more detail).

Fig. 1 illustrates the training procedure and the input format of the mBART models. In masking mode, the input data are (only) the amino acid sequence of the target protein; in mimicking mode, the input data are the amino acid sequence *and* the codons of an ortholog protein (Fig. 1*A*). The input format with two concatenated sequences supports both tasks and includes tokens indicating the organisms of the proteins (Fig. 1*B*). We trained multiple models, each with a specific window size (Fig. 1*C*). Preprocessing

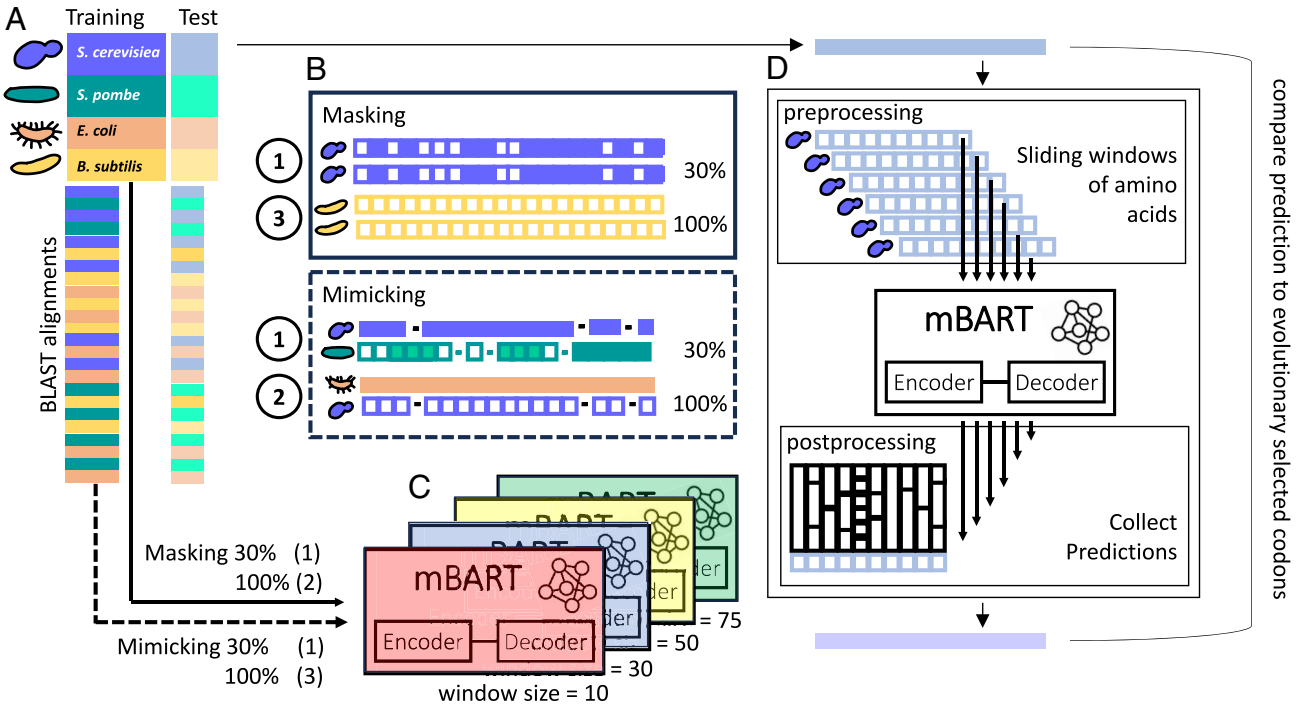

Fig. 1. Strategy to learn codon usage patterns in *S. cerevisiae*, *S. pombe*, *E. coli*, and *B. subtilis*. (*A*) Our dataset includes protein amino sequence data from the four organisms and BLAST-identified protein alignments. Following standard practices in AI, we split our data into training and test sets (after clustering the proteins with CD-HIT and ensuring that closely related proteins were in either the training or the test set; see *Methods* for details). (*B*) To support both the masking mode and the mimicking mode, the input format has two sequences, with each sequence preceded by its source organism. In the cartoon representation, codons are shown by solid-colored boxes, and their corresponding amino acids by hollow boxes. In masking mode, both sequences are the same, and the input is either a sequence of codons where 30% or 100% of the positions are masked. In mimicking mode, the first sequence are the codons in an orthologous protein, and the second sequence is a masked sequence of codons, where 30% or 100% of the positions are masked. (*C*) The training set was used to train mBART models with varying window sizes (10, 30, 50, and 75 codons). First (1), we pre-trained the models with 30% masking and mimicking. Next (2), we fine-tuned the model with the 100% masked sequences to generate the fine-tuned masking model. Third (3), we further fine-tuned the model with 100% masking and mimicking data to generate the fine-tuned mimicking model. (*D*) During inference, there are pre- and post-processing steps: For each protein in the test set, all sliding windows corresponding to the model window size were considered, and each sequence of codons was fully masked. The cartoon example shows predictions in masking mode for a sliding window of 10 codons in an *S. cerevisiae* protein. The network softmax output predicts for each amino acid a distribution over its possible codons. These predictions are combined to yield the final codon prediction for the sequence, and we measured the accuracy of the prediction with respect to the evolutionarily selected codon sequence.

and postprocessing steps are illustrated in Fig. 1*D*. The performance of the trained models was evaluated based on accuracy of prediction of the codons of all proteins in the test set. We observed that the frequency-based model trained on highly expressed proteins was more accurate when predicting the codons of highly expressed proteins (Fig. 2); therefore, we added a 6-class classification token of the expression level of the protein within its organism (omitted from the illustration in Fig. 1*A* for brevity). In masking mode, the input is two copies of the same amino acid sequence and the organism token. In mimicking mode, the input is a gap-infused alignment of two orthologous proteins, where the first sequence is the codons of the source protein and its organism, and the second sequence is the amino acids of the target protein and its organism. Some of the codons in the target sequences may be passed as input for context (i.e., they are not predicted). For example, during pretraining, only 30% of the positions are masked by amino acids and predicted by the model.

Masking-Mode mBART Predictions Have Better Accuracy than the Frequency-Based Baseline Suggesting That Long-Range Codon Interaction Patterns Can Be Learned. Fig. 2 shows the accuracies of codon predictions for the test-set proteins in masking mode for different models, as a function of expression level in the four organisms. The mBART models were pretrained on the masking and mimicking tasks and then fine-tuned (FT) on the masking task. The models varied by window size, with windows of 10, 30, 50, and 75 codons evaluated. We fixed the mBART window size in each model so that we could reason about the scale of the learnable long-range interactions. The accuracies of the baseline frequency-based models, which are based on the most frequently used codon for each amino acid in each organism for all training-set proteins (cyan), in the 10% most highly expressed training-set proteins (magenta), and the frequency-based bigram model (black) were calculated for comparison. The mBART models are more accurate than the frequency-based models, suggesting that there are patterns that can be learned from the long-range relationships among codons.

We sort the proteins by expression level and show their ranked position along the *x*-axis; *SI Appendix*, Fig. S1 shows the distribution of the log(expressions) of the test-set proteins. *SI Appendix*, Figs. S2 and S3 show the same data as in Fig. 2, with the *y*-axis range optimized per organism and with expression values along the *x*-axis. *SI Appendix*, Fig. S2 also includes the frequency-based prediction (ochre) using Sen et al.'s model (30), and for *E. coli* ICOR predictions (light blue) (43) (see *Methods* for details). *SI Appendix*, Fig. S4 shows a collated version of the prediction accuracies using boxplots. In Fig. 2 and *SI Appendix*, Figs. S2 and S3, the data are smoothed with a Gaussian kernel (50 proteins window), and the horizontal solid lines indicate the average accuracy for the proteins with no measured expression. *SI Appendix*, Figs. S5 and S6 show the accuracy differences between our models and the naïve frequency-based baselines. *SI Appendix*, Table S1

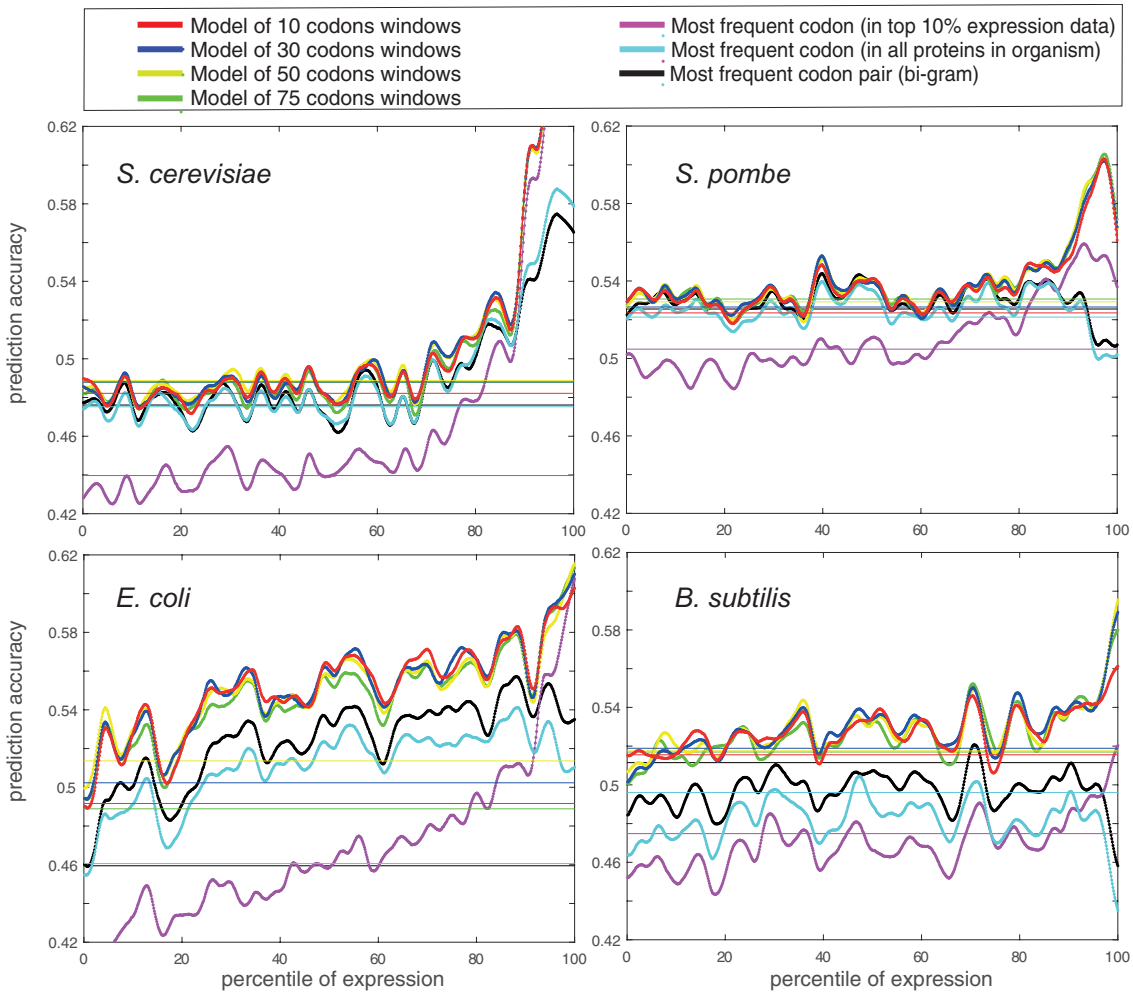

Fig. 2. The codon prediction accuracies for the test set proteins with inference in masking mode show that mBART-trained models predict codons better than the frequency-based models; the model with a 30-codon window is generally the top performer. Prediction accuracies for proteins in the test sets of *S. cerevisiae, S. pombe, E. coli*, and *B. subtilis* plotted vs. percentile ranking of expression. The average accuracies for proteins for which expression has not been measured are shown as solid horizontal lines. Data were smoothed with a Gaussian filter and a window size of 50 proteins. The mBART masking models with 10, 30, 50, and 75 codon windows are shown in red, blue, yellow, and green, respectively. The frequency-based model accuracies calculated on all data are shown in cyan, frequency-based model accuracies calculated on the top 10% of proteins based on expression levels are shown in magenta, and the bigram frequency-based model in black. In all four organisms, predictions improve when considering proteins that are more highly expressed. The improvement in accuracy is most pronounced when considering the bacterial proteins. That the accuracy of our models is better than the frequency-based baselines demonstrates that there is an evolutionary pattern of long-range dependencies among codons, and it is sufficiently pronounced that the AI models considered here can learn it.

and Table 1 list the accuracy differences, the *P*-values, and the effect size (normalized Cohen d-values) comparing pairs of models. *SI Appendix*, Table S1 compares prediction accuracies of different mBART models, showing that the 30-codon window performs best; Table 1 compares prediction accuracies of the 30-codon window to the frequency-based models. The statistical test considers the null hypothesis that the data in the differences (e.g., accuracy of prediction of the mBART model minus the accuracy of the frequency-based model) comes from a normal distribution with zero mean and unknown variance, using the paired-sample *t* test. All *P*-values for comparisons with the baseline models are (far) smaller than 0.05 (Table 1), thus rejecting the null hypothesis, and suggesting the accuracies of the mBART predictions are indeed better than that of the frequency-based predictions. In all four organisms, the 30-codon window mBART model predicts codons more accurately than the frequency-based models (including bigram model and the model by Sen et al.). For *E. coli*, the mBART model's predictions are also more accurate than those of the RNN ICOR, which is consistent with the report by Yang et al. that the accuracy of their RNN predictions for *E. coli* proteins was comparable to the frequency-based model (26).

Fig. 3 compares the perplexity of different models on the test-set proteins, as a function of the expression level rank of the proteins (with a Gaussian kernel smoothing of 50 proteins). Perplexity is a commonly used measure in AI to assess model predictions and is the computed exponentiated average of the cross-entropy loss, implying that better models have lower perplexity (e.g., see ref. 26). *SI Appendix*, Fig. S7 shows a collated version of the perplexities using boxplots. The intuition behind perplexity is modeling the "surprise" of a model with respect to the naturally occurring codon sequences in the test sets, in terms of the probability the models assign these sequences. Namely, it complements the accuracy measure, which only rewards cases where the model assigns the highest probability to the correct codon as perplexity considers the probability assigned to the true codon regardless if it is the one with the highest probability or not. *SI Appendix*, Fig. S8 shows the prediction accuracies vs. perplexity for the test-set proteins and the different models. The two are negatively correlated, with the degree of correlation varying, depending on the model and the organism, and thus emphasizing that the information in these two measures is not redundant.

Table 1. Significance of differences in accuracies between models

	S. cerevisiae			*S. pombe*		
		Effect size			Effect size	
Compared predictions mBART (ws 30) vs.	*P*-value	Absolute (<% diff>)	Normalized (Cohen d)	*P*-value	Absolute (<% diff>)	Normalized (Cohen d)
Frequency-based (all)	$<10^{-5}$	1.5395	0.2047	$<10^{-5}$	1.2272	0.2830
Frequency-based (bigram)	$<10^{-5}$	1.5332	0.1791	$<10^{-5}$	0.7951	0.1602
Frequency-based (top 10%)	$<10^{-5}$	4.2307	0.9644	$<10^{-5}$	2.7519	0.7327
Sen et al.'s model	$<10^{-5}$	1.4390	0.1636	$<10^{-5}$	1.9056	0.4914
	E. coli			*B. subtilis*		
		Effect size			Effect size	
mBART (ws 30) vs.	*P*-value	Absolute (<% diff>)	Normalized (Cohen d)	*P*-value	Absolute (<% diff>)	Normalized (Cohen d)
Frequency-based (all)	$<10^{-5}$	4.0541	0.7363	$<10^{-5}$	4.5887	1.1914
Frequency-based (bigram)	$<10^{-5}$	2.8260	0.4966	$<10^{-5}$	3.0508	0.8132
Frequency-based (top 10%)	$<10^{-5}$	8.4185	1.4942	$<10^{-5}$	5.8368	1.4337
Sen et al.'s model	$<10^{-5}$	3.6533	0.6456	$<10^{-5}$	2.7311	0.6952
ICOR model	$<10^{-5}$	2.1384	0.3897			

P-values and normalized Cohen d values were determined using a paired-sample *t* test on the test set proteins (grouped by organism) for the test decision for the null hypothesis that differences between models come from a normal distribution with mean equal to zero and unknown variance. We see that in all cases, there is a difference, as evidenced by the absolute mean percent, the *P*-values that indicate that the difference is significant and the normalized Cohen d value indicating that the effect size is meaningful.

We also evaluated the accuracies of the mBART and the frequency-based models as a function of the protein length ranking (Fig. 4 and *SI Appendix*, Figs. S9–S11). *SI Appendix*, Fig. S12 shows the protein length distributions in the test sets. *SI Appendix*, Table S3 lists the Pearson Correlation coefficients (and associated *P*-values) between the prediction accuracies and protein lengths for the different models. These vary between −0.18 and 0.3, i.e., at best these are low correlations. In *S. cerevisiae, E. coli*, and *B. subtilis*, longer proteins are more accurately predicted; in *S. pombe* this is not the case. The accuracy advantage of the mBART models is more pronounced in the shorter proteins of the eukaryotic organisms and the longer proteins of the bacterial organisms (*SI Appendix*, Figs. S10 and S11).

SI Appendix, Figs. S13 and S14 show distributions of other measures for the codon predictions for our models and the frequency-based baselines. To extend the binary metrics accuracy, precision, recall, and F1 to the multilabeled codon prediction task, we average with equal contributions these metrics for each sample-class pair ("micro" averaging) (55). *SI Appendix*, Fig. S13 shows the accuracy and F1 distributions of the mBART models are with higher values than for frequency-based models. *SI Appendix*, Fig. S14 shows that the precision and recall distributions of the mBART models are with higher values than for frequency-based models. *SI Appendix*, Fig. S15 shows the boxplots of CAI (56) values of different predictions. The most frequent codon prediction is shown only for reference as by definition, its CAI values are 1. We also show the CAI values of the true, evolutionary-selected, codons. We see that even for this measure, on which the mBART models were not optimized, their codon predictions are more similar to the true ones, compared to predictions by the frequency-based models.

To determine whether similarities between the training and test sets artificially boost prediction accuracies, we identified similarities at the amino acid level that remained after we used CD-HIT clustering for the training set/test set split. To do this, we BLAST aligned (with an E-value threshold of 10^{-2}) each test protein to the training set. Then, we calculate its average percent identity to the training set by identifying for each residue the aligned segment with the highest percent identity and averaging these values over the protein residues. This analysis showed that 31% of the test set proteins have no sequence identity to proteins in the training set (28.1%, 28.1%, 30.3%, and 37.2% for *S. cerevisiae*, *S. pombe*, *E. coli*, and *B. subtilis* respectively; *SI Appendix*, Fig. S16). The accuracies of the predictions of the mBART and frequency-based models as a function of the average identity were then calculated (*SI Appendix*, Fig. S17). The accuracy gap between the mBART and frequency-based models was similar regardless of whether or not the test set protein had close neighbors in the training set, suggesting that the performance of the mBART models is not artificially boosted by these similarities.

To better understand the contributions to the models' accuracies, we trained models with less inputs and analyze the differences in their prediction accuracies. To evaluate the contribution of the mimicking mode training phase, we trained models with only masking mode. As there is no need to accommodate for the mimicking mode, we also simplified the input format of the comparison models so that it includes only a single copy of codons/masked amino acids. *SI Appendix*, Fig. S18 compares the accuracies of our reported models and the ones trained without the mimicking mode phase and shows that for all models and in all organisms, there is an accuracy boost, albeit a modest one. We also evaluate the contribution of the expression level class input on the accuracies of the predictions. *SI Appendix*, Fig. S19 shows the accuracy and perplexity differences between our reported models and ones trained and tested without the expression level tokens (using the only masking models). Most noticeably, the accuracy is greater and the perplexity is lower in the top 10% most highly expressed proteins and the differences are more pronounced in the eukaryotic organisms.

Figs. 2 and 3 and *SI Appendix*, Figs. S2–S7 and S9–S14 and Tables S1 and S2 show that the mBART models predict more accurately than the frequency-based models (including the models

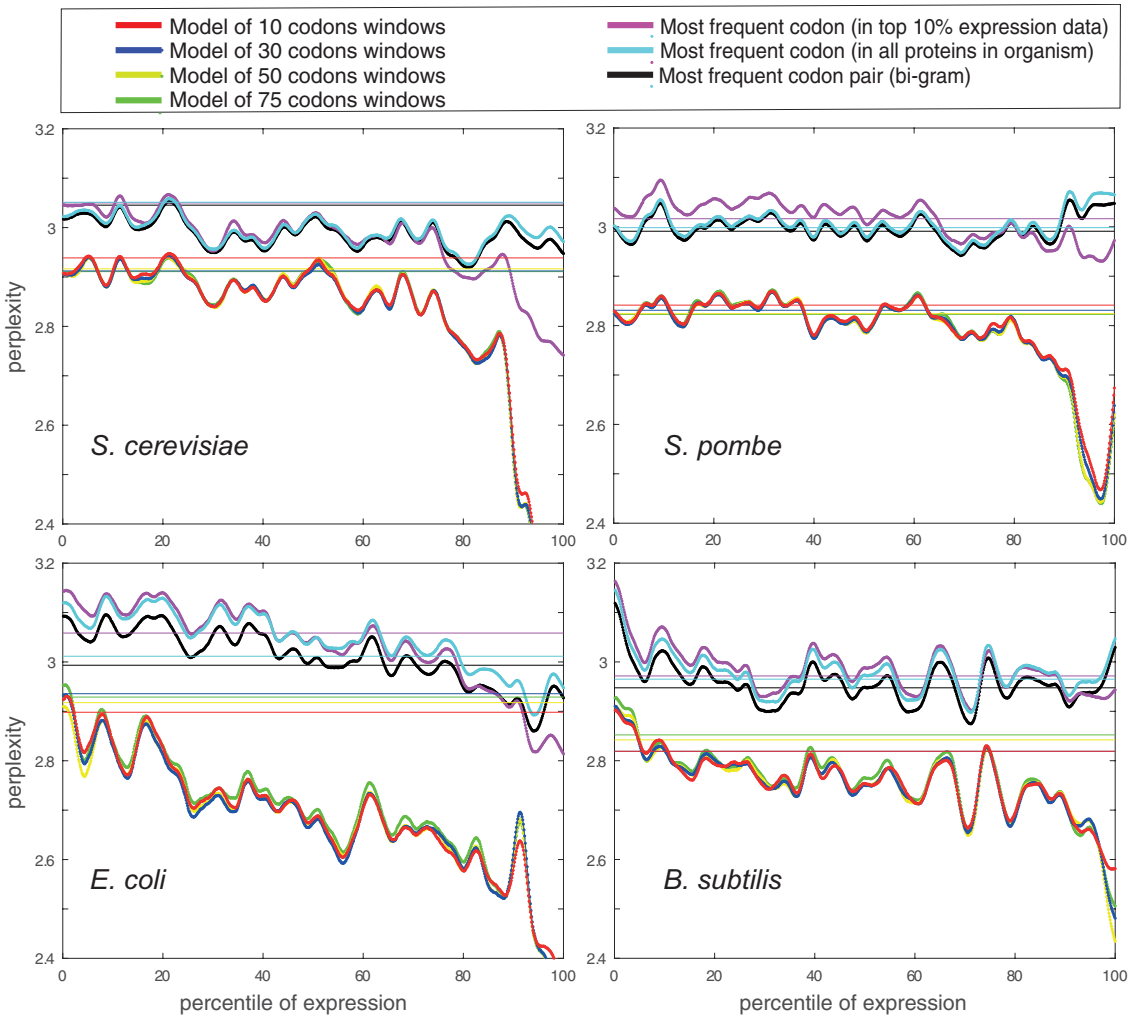

Fig. 3. Calculated perplexities for masking-mode mBART predictions are lower than those of the frequency-based models. Perplexity, which is the computed exponentiated average of the cross-entropy loss, plotted vs. percentile ranking of expression. The average perplexities for proteins with no measured expression levels are shown by solid horizontal lines. The data were smoothed with a Gaussian filter and a window of 50 proteins. These graphs complement Fig. 2 showing the accuracies: The mBART models perform better (lower perplexity) than the frequency-based models, the model with a window size of 30 codons is a top performer, and the perplexities are lower when considering proteins that are more highly expressed. This provides further support to that there is an evolutionary pattern of long-range dependencies among codons, and it is sufficiently pronounced that the AI models considered here can learn it.

by Sen et al., and ICOR for *E. coli)*, demonstrating there are patterns that can be learned from the long-range relationships among codons. Both for the frequency-based models and the mBART models it is easier to predict the codons of more highly expressed proteins. In *S. cerevisiae*, the advantage of mBART is the smallest and is in the highly expressed proteins. In *S. pombe*, the best mBART models are better than the frequency-based models, and for the bacteria, the mBART models offer the largest improvement in prediction accuracy and perplexity with respect to the frequency-based models, across all levels of protein expression. Furthermore, considering both the accuracy and the perplexity shows that the AI-based models can learn patterns of codon usage even better when relying on a window that is even longer than 10 codons.

Mimicking-Mode mBART Predictions Are Only Marginally Better than Masking-Mode Predictions. To identify orthologous segments that can be used in mimicking-mode predictions, we used BLAST (with an E-value cutoff of 10^{-2}). *SI Appendix*, Fig. S20 shows histograms quantifying the similarities among the orthologous segments in the test set using percent identity and $\log_{10}$(E-values). The average percent identities among the orthologous segments in the test set are 32%, 33%, 30%, and 29% for *S. cerevisiae*, *S. pombe*, *E. coli*, and *B. subtilis*, respectively. During training, the data are of aligned proteins that are both from the training set, and in testing, orthologous proteins that were both in the test set were evaluated.

We evaluated different mBART models in masking-mode inference and mimicking-mode inference, on the test set of orthologous segments. First, we consider the same models (FT on the masking task with windows of 30 and 50 residues) and the same masking-mode inference as described above, only on this different test set. We further fine-tuned these two models on the masking and mimicking task, and evaluated these refined models in masking-mode inference and in mimicking-mode inference. Finally, we also calculated the frequency-based baseline (similarly to the what is described above, only on this test set), and a naïve frequency-mimicking model where we mimic the codon with the same frequency rank as in the orthologous protein (see *Methods* for details). Fig. 5 and *SI Appendix*, Figs. S21 and S22 show the accuracies of the predictions as a function of the sequence identity to the ortholog; *SI Appendix*, Fig. S21 adapts the *y*-axis per organism, and *SI Appendix*, Fig. S22 shows these accuracies with respect to the frequency-based baseline.

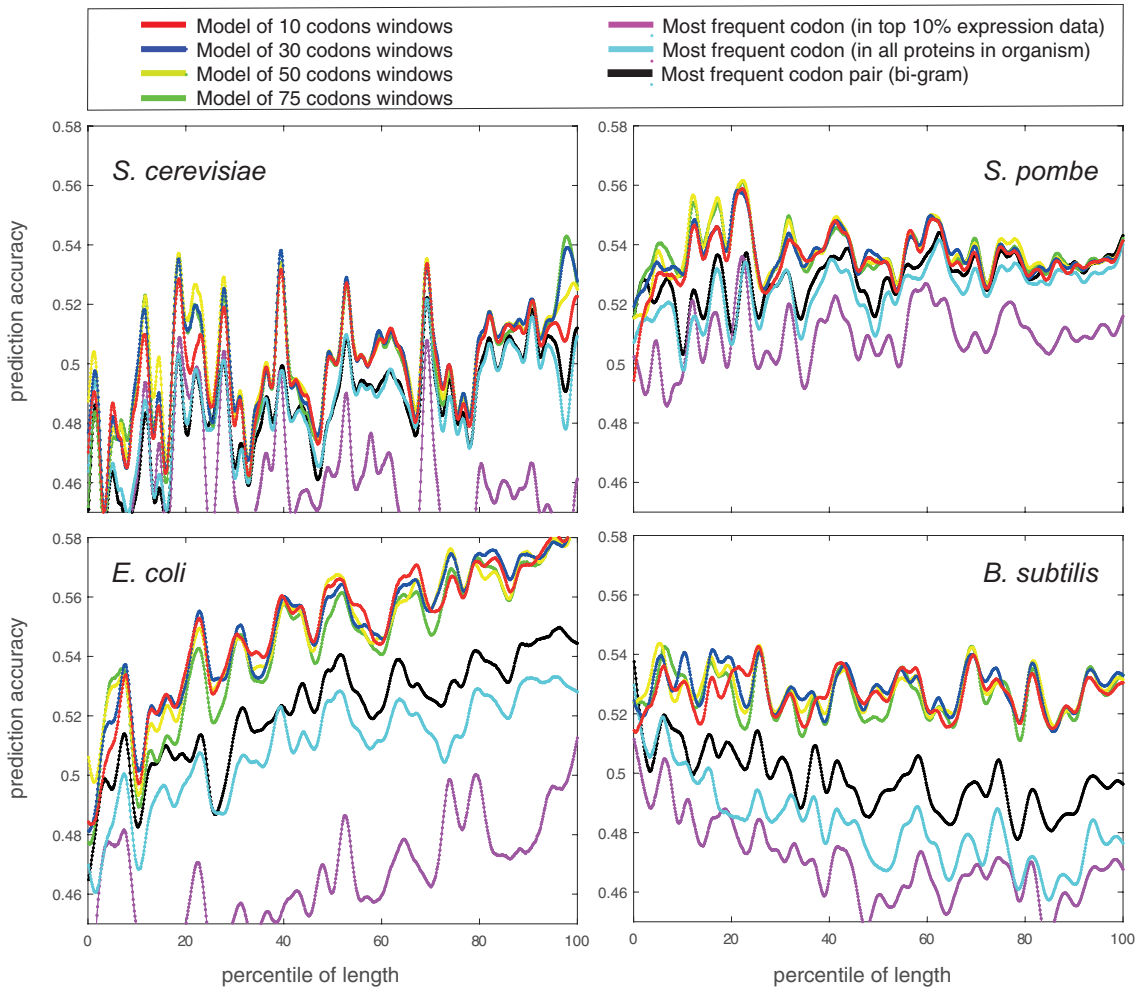

Fig. 4. Codon prediction accuracies for the test set proteins with inference in masking mode do not have a clear dependency on length. Prediction accuracies vs. length percentile ranking, with the data smoothed with a Gaussian filter (window size of 50 proteins). There is no simple connection between the accuracies of the models (any of them) and the lengths of the proteins: In *S. cerevisiae* and *E. coli*, the mBART models are more accurate for longer proteins, in *B. subtilis*, the models are more accurate for the shorter proteins, and in *S. pombe*, there is no clear dependency between the two.

The mBART model's mimicking-mode predictions have accuracies that are very similar to the masking-mode predictions (see also *SI Appendix*, Fig. S18). The mimicking mode resulted in a modest accuracy boost in the two eukaryotic species and for orthologs that have segments with high amino acid sequence identity. Prediction accuracy may be higher due to the codons used in a particular segment are easier to predict, namely the most frequently used ones. In the masking-mode frequency-based predictions, ortholog codons are not used, and hence percent identity is only meaningful in terms of characterizing the protein as a conserved one. In the eukaryotic species, the accuracies of both the naïve frequency-based mimicking and the mBART models became higher than the frequency-based baseline as the percent identity of the orthologs increased. This implies that there is a signal in the codons of the orthologs, especially when the percent identity is high, that improves prediction accuracy. However, the masking-based predictions have almost identical accuracies, so this extra information does not further improve the mBART predictors. It should be noted that even when there is a signal, the information from the orthologs could be detrimental. For example, although the naïve frequency-based mimicking improved due to this signal, it was still less accurate than the frequency-based masking model. Our mimicking-based predictions have similar, and in some cases, improved accuracy compared to the naïve frequency-based mimicking model.

SI Appendix, Table S2 shows the accuracy differences, *P*-values, and effect sizes (normalized Cohen d values) comparing pairs of models. The comparison of the mBART model in mimicking mode and the frequency-based mimicking model for all four organisms shows that the difference is significant (e.g., all *P*-values are less than 10^{-5}). In contrast, the comparison of the mBART model in masking mode vs. in mimicking mode shows their differences are not significant.

mBART Outperforms the Frequency-Based Baseline More Significantly for Certain Types of Proteins. To determine whether prediction accuracies depend on molecular functions or type of biological process, we use the GO-XL slim classifications of *S. cerevisiae* and *E. coli* proteins and considered cases for which there were at least 10 proteins with that annotation in the test sets of these organisms. Each such classification is placed along the x-axis and the proteins with that annotations are shown as box plots and scatter data. We used the Mann–Whitney rank-sum test to determine whether the accuracy difference values for the proteins with a specific GO annotation are likely from the same population of values as the rest of the organism's test-set proteins (*SI Appendix*, Table S4).

For most GO-annotation categories, the accuracy boost of the mBART model is not different from the background of all test-set proteins in either organism, but there are some that stand out

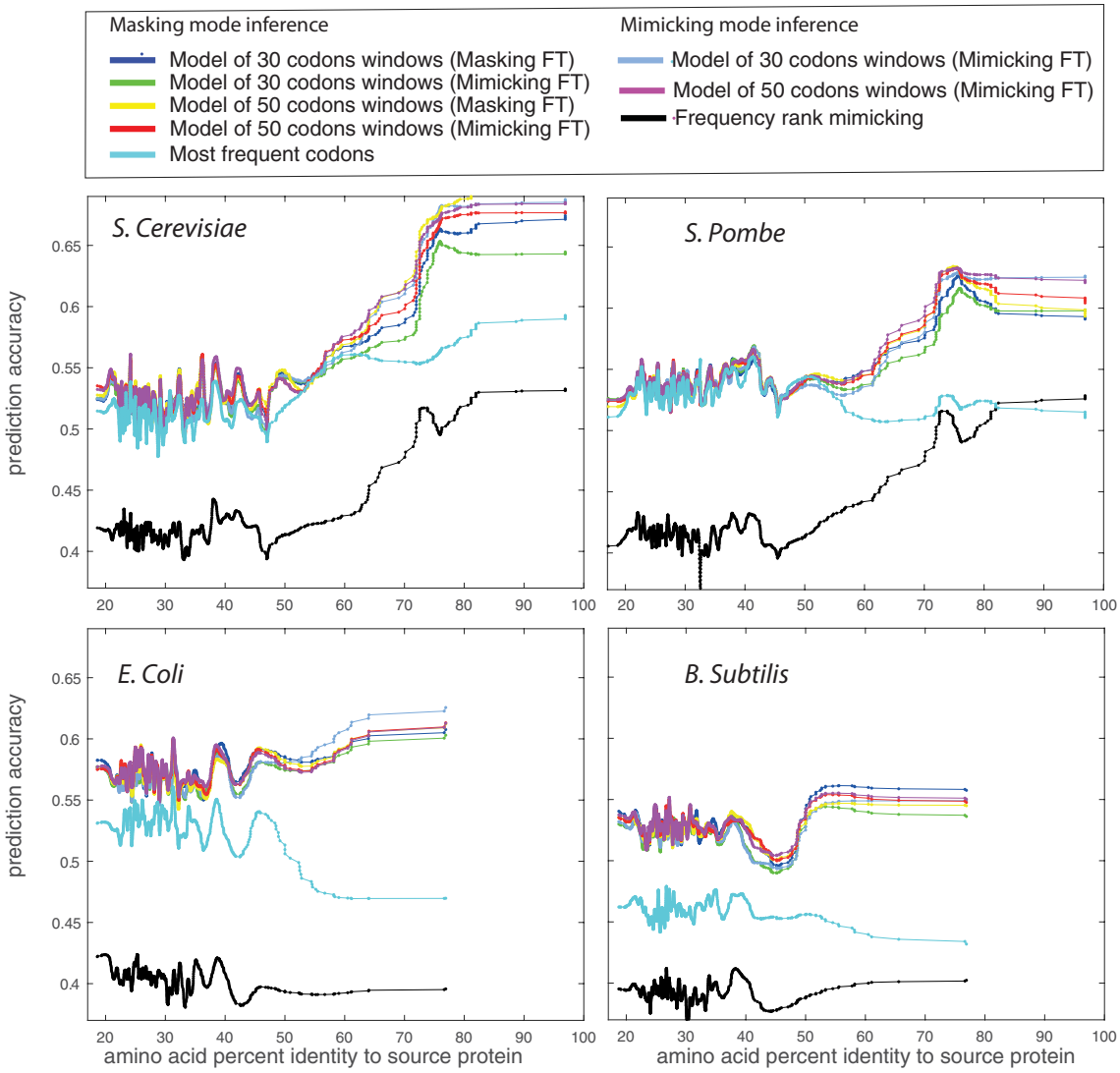

Fig. 5. mBART mimicking-mode inference accuracy is on par with the masking mode inference. Prediction accuracies vs. amino acid percent identify to the orthologous protein. For the dataset of homologous pairs, our goal is to predict the codon encoding of one of homologues. However, there are two modes of using the models. In the first mode, denoted "masking," the input is (two identical copies of the) masked amino acid sequence of the target protein. In the second mode, denoted "mimicking," the input is the masked amino acids of the target protein and the aligned codons in its homologue. We show masking-mode predictions by two models that were fine-tuned on the masking task (window size of 30 and 50 in blue and yellow, respectively). These are the same models and inference mode shown in Fig. 2, only here they are evaluated on the protein segments found in the alignment dataset. We also further fine-tuned these models on the masking and mimicking task, and the accuracies of the predictions of these FT models in masking-mode inference are shown in green and red, respectively. Alternatively, we used these same models in mimicking-mode inference and the accuracies of these predictions are shown in maroon and magenta. For comparison, we show the frequency-based model on this dataset in cyan and the prediction accuracy of a frequency-based mimicking model in black. Data were smoothed with a Gaussian filter (window size of 50 proteins). The signal of codons in the orthologous proteins is not sufficiently strong that our AI models can exploit it to improve their predictions, and the performance of the mimicking-mode inference (in maroon/magenta) is on par with that of the masking mode inference (in green and red), with a slight advantage for proteins with very close orthologs in bacterial organisms.

(Fig. 6 and *SI Appendix*, Figs. S23–S26). In *S. cerevisiae*, the molecular function categories in which the mBART model predicts even better than the frequency-based baseline are 1) "structural constituent of ribosome," 2) nucleic acid binding function "rRNA binding," "RNA binding," and "DNA binding," and 3) molecular function that falls under the broad category of catalytic activity and specifically "transferase activity," "nucleotidyltransferase activity," "nuclease activity," and "peptidase activity." For *E. coli*, these are 1) "structural molecular activity," 2) "RNA binding," and 3) both the broader "catalytic activity" and more specifically "catalytic activity acting on DNA." Considering biological processes, in *S. cerevisiae* these are 1) "ribosomal large subunit biogenesis," 2) the broad category of translation and specifically "cytoplasmic translation" and "regulation of translation," 3) the broad category of nucleic acid metabolic process and specifically "DNA recombination" and "rRNA processing," and 4) "transposition." In *E. coli:* 1) "ribosome biogenesis," a parent category of "ribosomal large subunit biogenesis" identified in *S. cerevisiae*, 2) "cytoplasmic translation," a process also identified in *S. cerevisiae*, 3) and "protein containing complex assembly."

To understand which gene ontologies (GOs) annotate genes that tend (or do not tend) to use predictable codons, we employed GO enrichment tools (57, 58) to find GOs that appear in genes with best/worst performances in three cases: 1) our mBART model prediction accuracy, 2) frequency-based model prediction accuracy, and 3) improvement of the prediction accuracies by mBART over the frequency-based model. We focused on *S. cerevisiae* as it is the organism with most comprehensive and accurate functional annotations. Our detailed enrichment analysis and data are in *SI Appendix*, *Supplementary Text* and Datasets S1–S6.

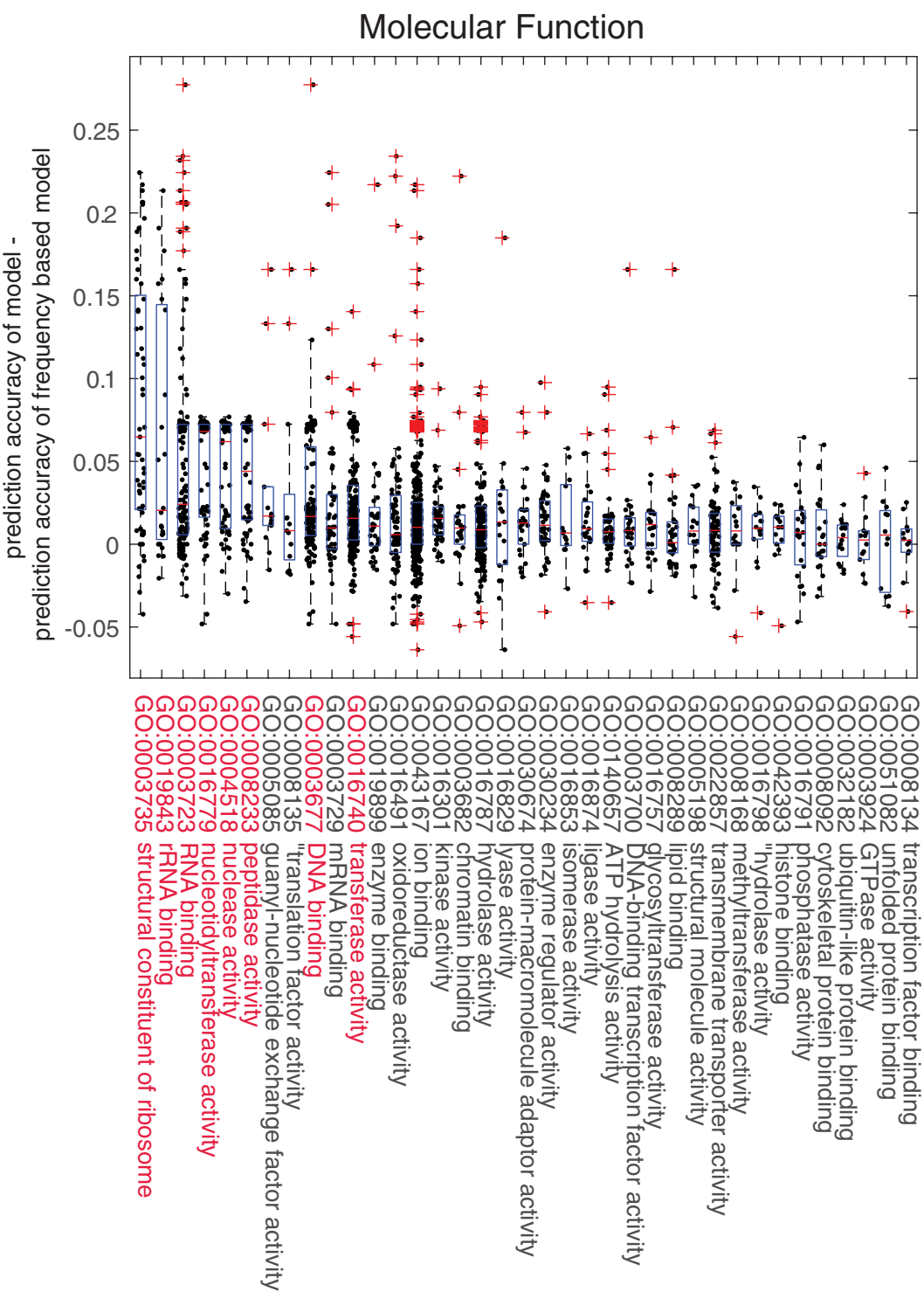

Fig. 6. Prediction accuracy is high for proteins in certain GO functional groups. Differences between masking-task-FT mBART model (30-codon window) predictions and the frequency-based baseline for *S. cerevisiae* test-set proteins grouped by GO molecular function terms. Terms were sorted in descending order by the mean average difference. The terms for which the *P*-value is <0.05 in a Mann-Whitney rank sum test for the functionally grouped set vs. the rest of the test-set proteins are highlighted in red. We see that the mBART model outshines the naïve approach more significantly for proteins with the function "structural constituent of ribosome," nucleic acid binding function, and sub-functions of catalytic activity.

Discussion

We have employed contemporary tools from AI to predict codons in genes from the eukaryotes *S. cerevisiae* and *S. pombe* and the bacteria *E. coli* and *B. subtilis*, given the protein amino acid sequences. Our approach is data-driven: mBART models were trained on data from each organism and are evaluated on a separate and diverse test set. Because the difficulty of codon prediction varies, evaluation on a large and diverse test set rather than on only a few proteins as has been done previously (26) results in a more comprehensive test of predictive power and allows us to study the accuracy of the models as a function of protein expression, length, conservation, and functionality. The lower bound for the accuracy of the model are the frequency-based models, and we expect the effective upper to be significantly lower than 100% because not all positions are under a strong evolutionary pressure (53). Given this, the improvement reported here relative to

the frequency-based model is significant. The datasets we used for training/testing and our mBART models are publicly available (59), both as source code (60) and trained models (59) and through a web-based user-friendly interface (61).

Learning the statistical patterns of codon usage in an organism is challenging because the amount of available data is limited by the number of proteins in the organism. Previous studies highlighted associations between the frequencies of neighboring codons and attributed these to their effect on ribosomal pausing, frameshifting, and other gene expression steps (31, 62, 63). Here, we used mBART, a transformer-based architecture (32), to learn correlations in codon use in residues that are 10 to 75 codons apart to improve predictions. Codon usage patterns inferred by our models are therefore related not only to translation but also to gene expression steps including transcription, splicing, methylation, RNA processing, mRNA stability, and genomic stability (7). Each of our models relies on a fixed window size, allowing us to reason about the distance among codons in which our models can still improve, presumably by learning patterns among codons that are separated this far apart. That the model with a 30-codon window was a better performer than models with longer window sizes may be due to statistical aspects such as the amount of data given the number of parameters in the model. However, this window size may be optimal because it is closest to the length of the ribosomal exit tunnel (64).

That our trained models can better learn the codon patterns of a subset of proteins suggests that their codon usage is more constrained, and this may be due to the encodings of these proteins being under more pronounced evolutionary selection. Here, we study how the accuracy and perplexity of our trained models vary as a function of the expression levels, lengths of the proteins, conservation, and GO annotations. It was suggested that selection for codon usage increases with the expression levels of a gene (25, 65). One reason for this is that we expect a silent mutation to have a greater effect on organism fitness in highly expressed genes (25). Indeed, the codons of high-expression proteins in all four organisms, are more accurately predicted both by the naïve frequency-based models and our models, and with lower perplexity.

It was shown that coding sequences may include overlapping codes related to all gene expression steps (7, 52, 66–68). Many of these codes are longer than a single codon (for example motifs related to transcription are longer than 3 codon); thus, they cannot be detected by measures that are based on a single, or pairwise, codon distributions. On the other hand, our mBART model is expected to capture such patterns. The fact that we see an increase in prediction accuracy of our model in coding sequences of highly expressed genes and an improvement compared to the frequency-based models, may suggest that they are also under stronger selection for longer codes.

We also observed that in most organisms (excluding *S. pombe*) the gap between the performances of our models and the frequency-based approach is greater for long proteins than for the short ones. It is possible that complex signals related to cotranslational folding regulation, needed for the tighter control of in multidomain proteins (69, 70), are "encoded" by codon patterns at distant positions and thus better detected by our models. It is also possible that our models detect gene expression codes beyond those important for cotranslational folding that are encoded in longer genes; these could be codes related to binding sites of transcription factors or RNA binding proteins (7).

Our models performed better for genes with conserved orthologs than for those without. The latter are either recent or have undergone very rapid evolution. This suggests that evolution first shapes the amino acid sequence and later the complex codon usage patterns. It may also suggest that older genes tend to include more complex gene expression codes than newer ones, possibly as a result of selection for tighter regulation, and it is these complex codes that are detected and exploited by our models.

The gap between the performances of our models and the frequency-based baselines varied among different GO functional groups. This result may be related to factors mentioned above and that seem to introduce complex codes into the coding sequence: expression levels, gene age and conservation, and gene length. For example, genes related to the translation process (e.g., those encoding ribosomal proteins), that were better predicted by our models, are known to be highly expressed and old. DNA and RNA binding proteins are also known to include ancient domains, and this may explain the better performance of our models in predicting their codons (71). Our GO-based analysis suggests that our models can be useful for studying and annotating novel genes given that predictive performance is associated with functionality.

Finally, our models predict better than the frequency-based approach more significantly in bacteria than in the yeast species evaluated. This may be due to the larger effective population size in bacteria (72). A larger population should induce stronger selection pressure on codon usage, which in turn will result in complex-long-range signals that can be learned by our models but not by the frequency-based models. Also, the mean number of ribosomes per mRNA in bacteria is greater than in eukaryotes making traffic jams more common, and possibly triggering stronger selection for complex codon usage signals (73–76). Moreover, horizontal gene transfer occurs in bacteria (77), which may accelerate the evolutionary rate.

We trained our mBART models to mimic, i.e., to predict the codons of a protein given the codon encodings of an ortholog. An accurate mimicking tool is useful for predicting codons that are optimal in a non-native host. To design a mimicking tool for protein codons, the first step is to characterize patterns of codon usage in orthologous proteins. Then, the learned insights (or AI models) can be used, given the codon encoding of a protein in a source organism, to predict a codon encoding for a similar protein in a target organism. The final, sometimes overlooked step, is to evaluate the predicted codon encodings in the target organism.

Previous studies used "hand-crafted" measures to characterize patterns of codon usage in orthologous proteins: Pechman and Frydman devised a translational efficiency scale and applied it to yeast species to show evolutionary conservation of codon optimality in eukaryotes (44). Jacobs and Shakhnovich measured local rare-codon enrichment and studied its conservation across multiple-sequence alignments (45). Chaney et al. used the MinMax measure to identify clusters of rare codons and showed that they are conserved among homologous proteins across eukaryotic and bacteria species (78). These measures are only a few of the many available codon bias indices (25). In contrast to these hand-crafted features, our mBART models learn patterns from the data, both indirectly by training a single model on the masking task for multiple organisms and directly by training on the mimicking task from alignment data. As in all contemporary AI-based models, the patterns learned help improve accuracy during inference, even though we do not have an explicit description of what the models learned.

Evaluating mimicking tools is challenging because there is a discrepancy between their desired use and the data on which we can evaluate them. Their use is to predict the optimal codons for expression of a protein in a non-native host. Evaluation can be carried out on evolutionary data: the codon encodings in two organisms of merely similar, namely nonidentical, proteins. This

meaningful distinction has two important consequences. First, because evolutionary data do not include pairs of identical protein sequences in two organisms as in our desired use, current tools, like CHARMING (48) and CodonWizard (47), do not evaluate their measures, or even predict codons, based on the codon encodings of a similar, yet nonidentical, protein ortholog. Second, it is unclear what is a correct threshold for approximating this desired use, in terms of the amino acid sequence identity of the ortholog. We observed that mimicking predictions are not consistently better than the masking predictions. This may be because we evaluated our mimicking models on orthologs that are too remote. Alternatively, it may be that it is not easier to learn to mimic than to predict the codon encoding from the amino acid sequence. Evaluating codon encodings for heterologous proteins, and our mimicking tool in particular, can only be done in a laboratory experiment. Nonetheless, because accuracies of our mimicking-mode predictions improved as the amino acid sequence identity increased, which more closely approximates the desired use case, we believe that our mimicking-based predictions have the potential of being useful for optimizing expression of proteins in a non-native host. To ease their inclusion in future experimental evaluations, we provide the code and a web interface for this task.

In summary, we used AI to study codon usage bias. Our approach can be used to design the coding sequence of heterologous genes and also to study where there are learnable patterns of codon encodings. We believe that future research on codon usage will involve extensive application of AI-based approaches, like the one presented here.

Methods

We trained multiple AI models to predict the codon sequences of proteins in *S. cerevisiae*, *S. pombe*, *E. coli*, and *B. subtilis*. Our models are fixed window-size implementations of mBart–a transformer-based Encoder–Decoder architecture with an autoregressive decoder. Relying on the Huggingface infrastructure, we implemented and trained multiple models that vary by their window size and training protocol. We compare predictions by these different models to each other, and to baseline models, including most frequent codon, bigram frequencies, ICOR (43), and the predictions by Sen et al. (30). More details can be found in *SI Appendix*.

Data, Materials, and Software Availability. All study training/test data, code, and models are available and have been deposited in github (60) and Hugging Face (59). A web-interface for easy access to this tool has been set on ref. 61.

ACKNOWLEDGMENTS. We thank Prof. Rita Osadchy of University of Haifa and Michael Peeri, Lorna Bakhit, and Prof. Nir Ben-Tal of Tel-Aviv University for insightful discussions. S.B.-E. and T.T. were supported by the Safra Center for Bioinformatics at Tel-Aviv University, and T.S. and R.K. were supported by the Data Science Research Center at the University of Haifa.

Author affiliations: [a]Department of Computer Science, University of Haifa, Haifa 3303221, Israel; [b]Department of Biomedical Engineering, Tel-Aviv University, Tel Aviv 6139001, Israel; and [c]The Sagol School of Neuroscience, Tel-Aviv University, Tel Aviv 6139001, Israel

1. T. Ikemura, Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.* **151**, 389–409 (1981).
2. I. Frumkin *et al.*, Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4940–E4949 (2018).
3. E. Samatova *et al.*, Translational control by ribosome pausing in bacteria: How a non-uniform pace of translation affects protein production and folding. *Front. Microbiol.* **11**, 619430 (2020), 10.3389/fmicb.2020.619430.
4. M. V. Rodnina, The ribosome in action: Tuning of translational efficiency and protein folding. *Protein Sci.* **25**, 1390–1406 (2016).
5. G. Hanson, J. Coller, Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* **19**, 20–30 (2018).
6. E. Cohen, Z. Zafrir, T. Tuller, A code for transcription elongation speed. *RNA Biol.* **15**, 81–94 (2018).
7. S. Bergman, T. Tuller, Widespread non-modular overlapping codes in the coding regions. *Phys. Biol.* **17**, 031002 (2020).
8. K. C. Stein, J. Frydman, The stop-and-go traffic regulating protein biogenesis: How translation kinetics controls proteostasis. *J. Biol. Chem.* **294**, 2076–2084 (2019).
9. Y. Liu, Q. Yang, F. Zhao, Synonymous but not silent: The codon usage code for gene expression and protein folding. *Annu. Rev. Biochem.* **90**, 375–401 (2021).
10. F. Buhr *et al.*, Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell* **61**, 341–351 (2016).
11. J. L. Chaney, P. L. Clark, Roles for synonymous codon usage in protein biogenesis. *Annu. Rev. Biophys.* **44**, 143–166 (2015).
12. M. Thommen, W. Holtkamp, M. V. Rodnina, Co-translational protein folding: Progress and methods. *Curr. Opin. Struct. Biol.* **42**, 83–89 (2017).
13. R. Hershberg, D. A. Petrov, Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299 (2008).
14. T. Tuller, H. Zur, Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Res.* **43**, 13–28 (2014).
15. R. Saunders, C. M. Deane, Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.* **38**, 6719–6728 (2010).
16. T. F. Iv Clarke, P. L. Clark, Rare codons cluster. *PLoS One* **3**, e3412 (2008).
17. T. Ben-Yehezkel *et al.*, Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. *RNA Biol.* **12**, 972–984 (2015).
18. G. Kudla *et al.*, Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255–258 (2009).
19. D. B. Goodman, G. M. Church, S. Kosuri, Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475–479 (2013).
20. I. Weiner *et al.*, Enhancing heterologous expression in Chlamydomonas reinhardtii by transcript sequence optimization. *Plant J.* **94**, 22–31 (2018).
21. C. Gustafsson, S. Govindarajan, J. Minshull, Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**, 346–353 (2004).
22. Y. Rong *et al.*, Folding of heterologous proteins in bacterial cell factories: Cellular mechanisms and engineering strategies. *Biotechnol. Adv.* **63**, 108079 (2023).
23. Y.-A. Kim *et al.*, Computational design of mRNA vaccines. *Vaccine* **42**, 1831–1840 (2023).
24. S. T. Parvathy, V. Udayasuriyan, V. Bhadana, Codon usage bias. *Mol. Biol. Rep.* **49**, 539–565 (2022).
25. S. Bahiri-Elitzur, T. Tuller, Codon-based indices for modeling gene expression and transcript evolution. *Comput. Struct. Biotechnol. J.* **19**, 2646–2663 (2021).
26. D. K. Yang, S. L. Goldman, E. Weinstein, D. Marks, Generative Models for Codon Prediction and Optimization. Machine Learning in Computational Biology. https://mlcb.github.io/mlcb2019_proceedings/papers/paper_29.pdf. Accessed 16 December 2024.
27. R. Duda, P. Hart, D. Stork, *Pattern Classification* (Wiley-Interscience, ed. 2, 2001).
28. G. Moura *et al.*, Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS One* **2**, e847 (2007).
29. Y. Harigaya, R. Parker, The link between adjacent codon pairs and mRNA stability. *BMC Genomics* **18**, 364 (2017).
30. A. Şen *et al.*, Codon optimization: A mathematical programming approach. *Bioinformatics* **36**, 4012–4020 (2020).
31. H. Zur, T. Tuller, Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. *Bioinformatics* **31**, 1161–1168 (2015).
32. A. Chandra *et al.*, Transformer-based deep learning for predicting protein properties in the life sciences. *Elife* **12**, e82819 (2023).
33. R. Tunney *et al.*, Accurate design of translational output by a neural network model of ribosome distribution. *Nat. Struct. Mol. Biol.* **25**, 577–582 (2018).
34. D. R. Goulet *et al.*, Codon optimization using a recurrent neural network. *J. Comput. Biol.* **30**, 70–81 (2022).
35. D. A. Constant *et al.*, Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.02.11.528149 (Accessed 12 February 2023).
36. H. Fu *et al.*, Codon optimization with deep learning to enhance protein expression. *Sci. Rep.* **10**, 17617 (2020).
37. K. Jaganathan *et al.*, Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
38. S. Zhang *et al.*, TITER: Predicting translation initiation sites by deep learning. *Bioinformatics* **33**, i234–i242 (2017).
39. V. Agarwal, J. Shendure, Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
40. S. E. McGeary *et al.*, The biochemical basis of microRNA targeting efficacy. *Science* **366**, 6472 (2019).
41. C. Outeiral, C. M. Deane, Codon language embeddings provide strong signals for use in protein engineering. *Nat. Mach. Intell.* **6**, 170–179 (2024).
42. Z. Ren *et al.*, CodonBERT: A BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics* **40**, btae330 (2024).
43. R. Jain *et al.*, ICOR: Improving codon optimization with recurrent neural networks. *BMC Bioinformatics* **24**, 132 (2023).
44. S. Pechmann, J. Frydman, Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237 (2013).
45. W. M. Jacobs, E. I. Shakhnovich, Evidence of evolutionary selection for cotranslational folding. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 11434 (2017).
46. M. Zhou *et al.*, Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.* **97**, 974–987 (2015).
47. P. Rehbein *et al.*, "CodonWizard"–An intuitive software tool with graphical user interface for customizable codon optimization in protein expression efforts. *Protein Expr. Purif.* **160**, 84–93 (2019).
48. G. Wright *et al.*, CHARMING: Harmonizing synonymous codon usage to replicate a desired codon usage pattern. *Protein Sci.* **31**, 221–231 (2022).

49. M. Lewis *et al.*, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv [Preprint] (2019). https://doi.org/10.48550/arXiv.1910.13461 (Accessed 29 October 2019).
50. Y. Liu *et al.*, Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguist.* **8**, 726–742 (2020).
51. A. Dana, T. Tuller, The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.* **42**, 9171–9181 (2014).
52. T. E. F. Quax *et al.*, Codon bias as a means to fine-tune gene expression. *Mol. Cell* **59**, 149–161 (2015).
53. C.-H. Yu *et al.*, Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol. Cell* **59**, 744–754 (2015).
54. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
55. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. P. M. Sharp, W.-H. Li, The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
57. L. Kolberg *et al.*, g: Profiler–Interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* **51**, W207–W212 (2023).
58. F. Supek *et al.*, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
59. T. Sidi, mBART models for codon generation. Hugging Face. https://huggingface.co/siditom. Deposited 5 November 2023.
60. T. Sidi, ReverTra – Codon Optimization Tool code. Github. https://github.com/siditom-cs/ReverTra. Deposited 6 February 2024.
61. T. Sidi, ReverTra – Codon Optimization Tool web-access. aa2codons. https://www.aa2codons.info/. Accessed 6 February 2024.
62. S. Boycheva, G. Chkodrov, I. Ivanov, Codon pairs in the genome of Escherichia coli. *Bioinformatics* **19**, 987–998 (2003).
63. A. M. Alonso, L. Diambra, Dicodon-based measures for modeling gene expression. *Bioinformatics* **39**, btad380 (2023).
64. J. Frank *et al.*, A model of the translational apparatus based on a three-dimensional reconstruction of the Escherichia coli ribosome. *Biochem. Cell Biol.* **73**, 757–765 (1995).
65. M. dos Reis, L. Wernisch, Estimating translational selection in eukaryotic genomes. *Mol. Biol. Evol.* **26**, 451–462 (2009).
66. Y. Liu, A code within the genetic code: Codon usage regulates co-translational protein folding. *Cell Commun. Signal.* **18**, 145 (2020).
67. I. Menuhin-Gruman *et al.*, Evolutionary stability optimizer (ESO): A novel approach to identify and avoid mutational hotspots in DNA sequences while maintaining high expression levels. *ACS Synth. Biol.* **11**, 1142–1151 (2022).
68. F. Hia, O. Takeuchi, The effects of codon bias and optimality on mRNA and protein regulation. *Cell. Mol. Life Sci.* **78**, 1909–1928 (2021).
69. C. M. Kaiser, K. Liu, Folding up and moving on–Nascent protein folding on the ribosome. *J. Mol. Biol.* **430**, 4580–4591 (2018).
70. E. P. O'Brien *et al.*, Understanding the influence of codon translation rates on cotranslational protein folding. *Acc. Chem. Res.* **47**, 1536–1544 (2014).
71. W. H. Hudson, E. A. Ortlund, The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.* **15**, 749–760 (2014).
72. P. K. Ingvarsson, Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. *Mol. Biol. Evol.* **24**, 836–844 (2007).
73. D. Zenklusen, D. R. Larson, R. H. Singer, Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* **15**, 1263–1271 (2008).
74. J. R. Warner, The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24**, 437–440 (1999).
75. H. Bremer, P. P. Dennis, Modulation of chemical composition and other parameters of the cell by growth rate. *E. coli Salmonella Cell Mol. Biol.* **2**, 1553–1569 (1996).
76. A. Bartholomäus *et al.*, Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**, 20150069 (2016).
77. T. Tuller *et al.*, Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res.* **39**, 4743–4754 (2011).
78. J. L. Chaney *et al.*, Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput. Biol.* **13**, e1005531 (2017).