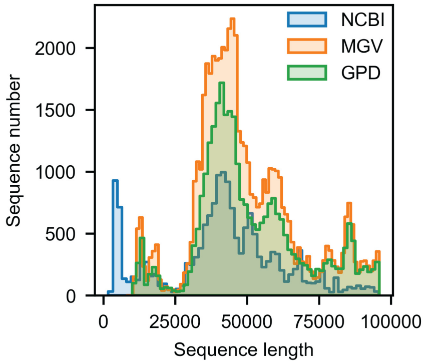
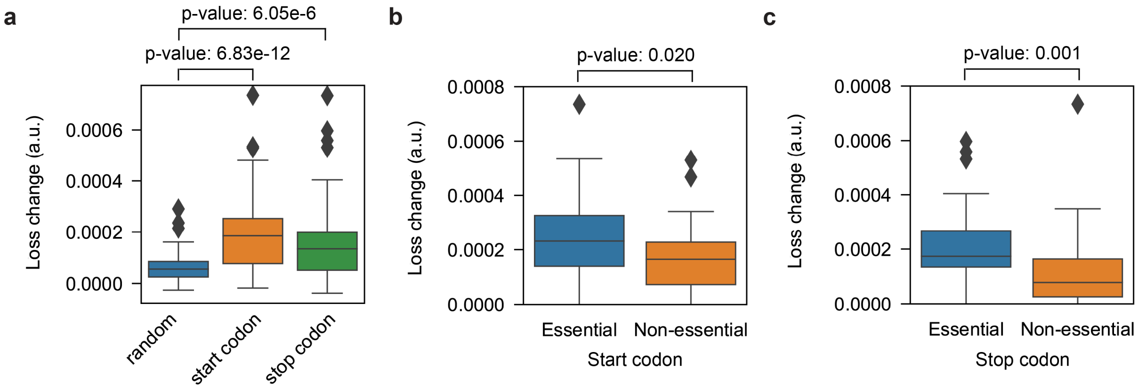
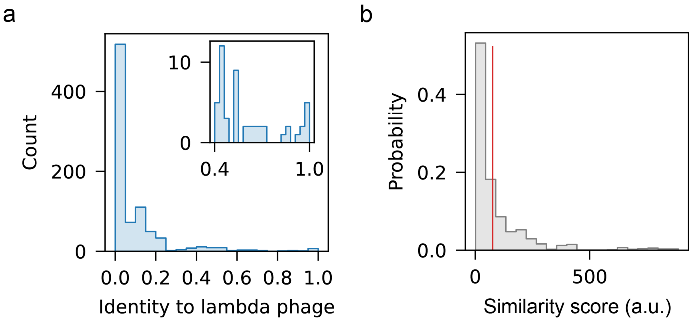
Supplementary figures



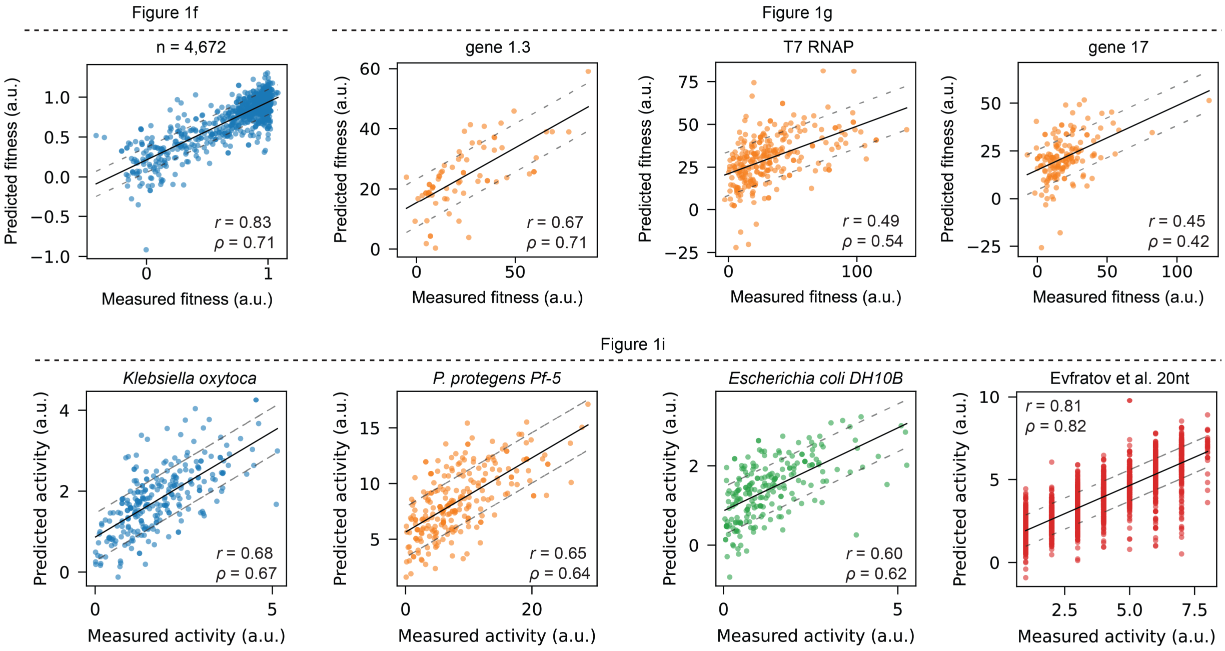
Supplementary Figure 1: Genome size distributions of three data sources. Distributions of genome sizes within the training dataset: NCBI (sample size: n = 16,609), MGV (sample size: n = 53,032) and GPD (sample size: n = 30,032).



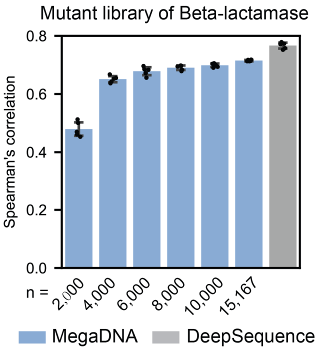
Supplementary Figure 2: Model loss changes due to codon mutations. **a**) Changes in model loss for random 3-nt mutations and mutations in start codon and stop codons of genes in the lambda phage genome (sample size: n = 73). **b**) Changes in model loss for mutations in start codons of essential genes (sample size: n = 29) and non-essential genes (sample size: n = 44). **c**) Changes in model loss for mutations in stop codons of essential genes (sample size: n = 29) and non-essential genes (sample size: n = 44). For a, b and c, p-values from the Mann-Whitney U test are shown. The central line inside the box represents the median value. The top and bottom borders of the box represent the third (upper) and first (lower) quartiles, respectively.



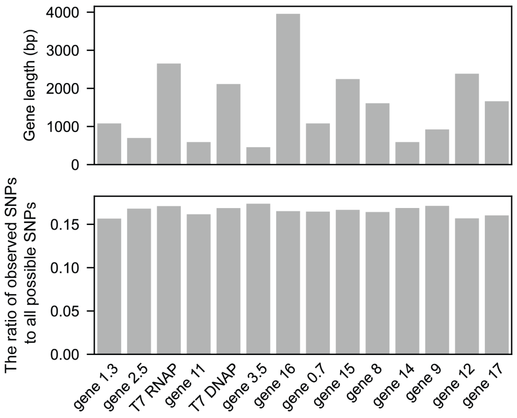
Supplementary Figure 3: Sequence similarity of training dataset and the lambda phage genome. a) Histogram showing the distribution of identity of training sequences that aligned with the lambda phage genome through BLAST analysis (n = 847). The identity for a specific sequence is calculated by summing the lengths of all aligned segments with the lambda phage genome and then dividing this total by the full sequence length. Inset, the counts of sequences with an identity above 0.4. **b**) Histogram showing the distribution of the similarity score of randomly sampled phage genomes from the training dataset (n = 2,000). For a specific genome, the similarity score is defined as the sum of identity for all the training sequences with significant hits. The similarity score for lambda phage is denoted in red.



<u>Supplementary Figure 4:</u> Correlation between model predictions and experimental measurements of protein fitness and regulatory element activities. Each dot represents an observed vs predicted value of either measured protein fitness (Figure 1f and 1g) or translational activity of 5'UTR sequences (Figure 1i). *r* and *p* are Pearson and Spearman correlation coefficients between the true and predicted values, respectively. The black line represents the linear regression fit, and the dashed lines show the fitted values plus or minus standard deviation of the prediction residuals.

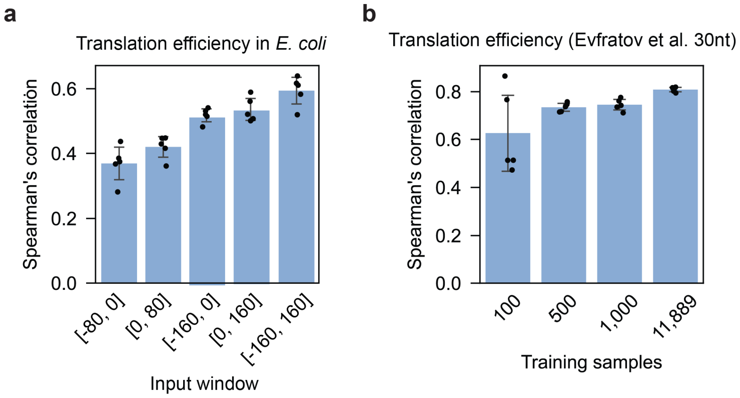


Supplementary Figure 5: **Prediction of mutational effects of beta-lactamase.** We used a codon mutant library of the *Escherichia coli* TEM-1 β-lactamase gene to evaluate model's performance[36]. Spearman correlation of the predicted and reported fitness for mutants from 5-fold cross validation tests are shown. n is the number of training samples for megaDNA. Blue and gray colors represent results from megaDNA and DeepSequence[15]. Error bars denote standard deviation.

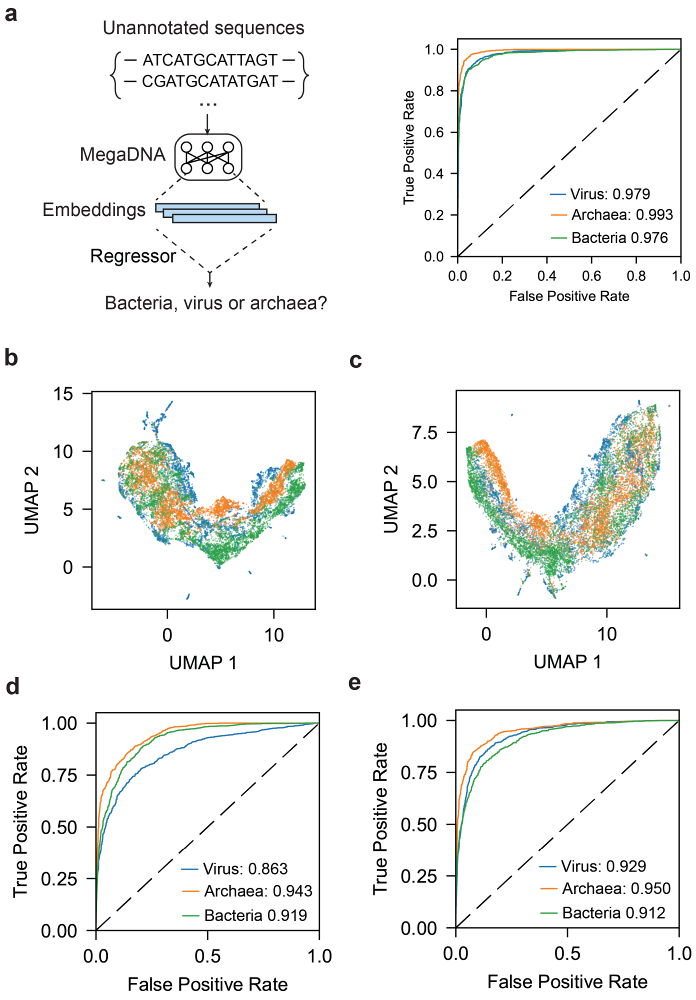


Supplementary Figure 6: Length and mutation coverage for the genes in T7 bacteriophage genome.

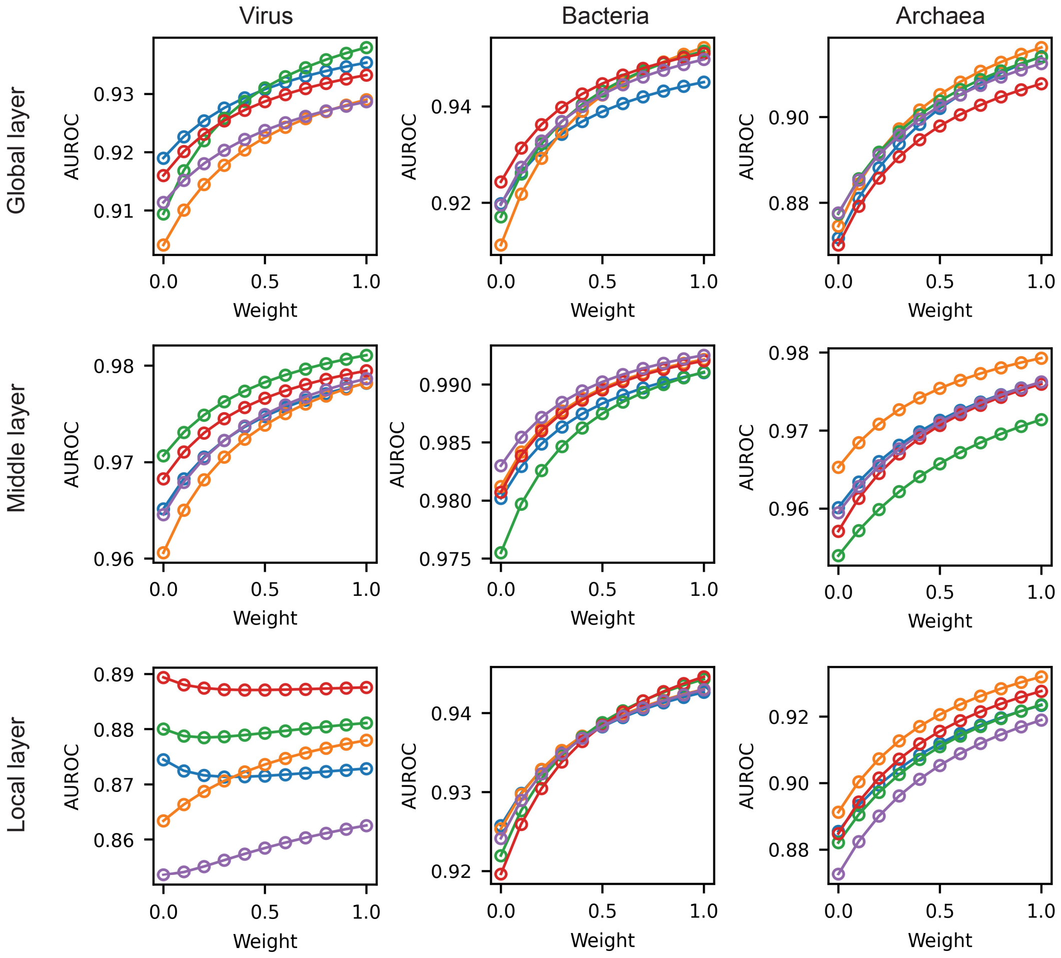
Upper**:** gene lengths. Lower: the ratio of observed SNPs to all possible SNPs per gene.



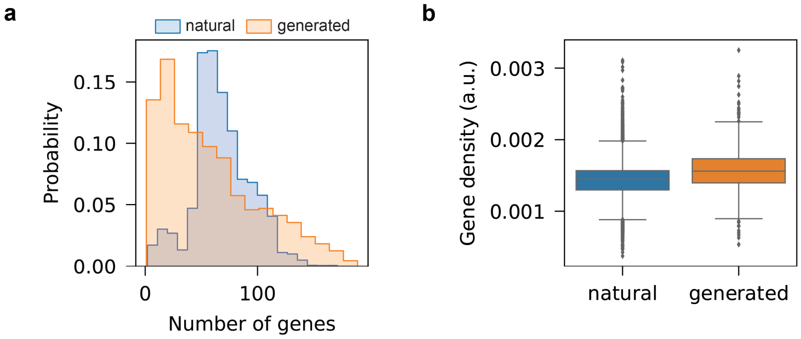
<u>**Supplementary Figure 7:**</u> **Impact of window position and training sample size on translation efficiency prediction. a**) Impact of the input window position for the prediction of translation efficiency of endogenous genes in *E. coli*. Positions are reported relative to the start codons. **b**) Effect of training sample number on the prediction performance of 5'UTR activity for the Evfratov et al. dataset. For a) and b), results from 5-fold cross validation tests are shown (n = 5 folds). Error bars denote standard deviation.



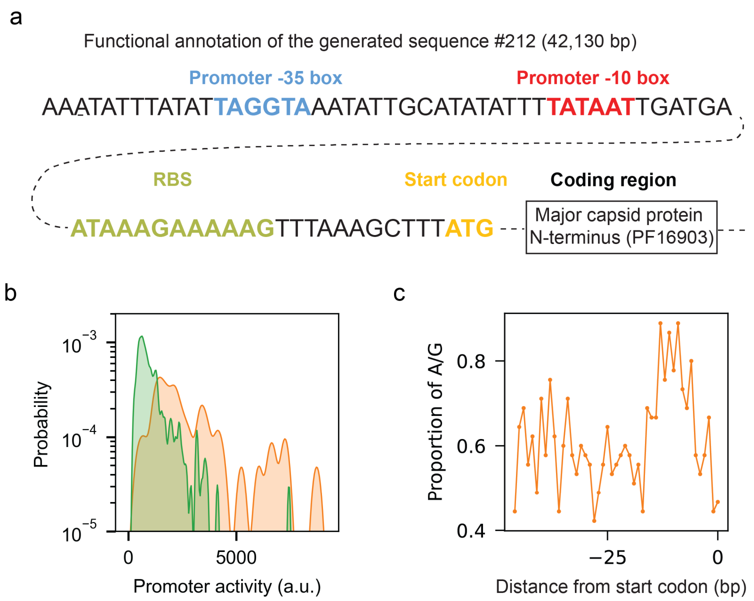
Supplementary Figure 8: Taxonomy prediction of unannotated sequences. **a)** Embeddings from the middle layer was used to classify sequences into virus, bacteria, and archaea. Visualization of sequence embeddings from the local layer (**b**) and global layer (**c**), and their taxonomy prediction performances (**d**) and (**e**) are shown. For a), d) and e), the model's performance was assessed using 5-fold stratified cross-validation tests. The receiver operating characteristic (ROC) curves are shown (sample size: n = 5,000 for each category). The mean AUROC scores from 5-fold cross-validation tests are reported.



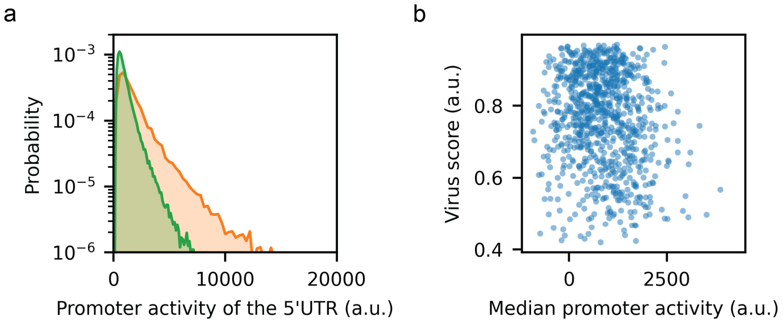
<u>Supplementary Figure 9:</u> Effect of sequence similarity on the accuracy of taxonomy prediction. Model predictions on the test dataset are weighted by their pairwise genome similarity to the training dataset: a weight of 0 excludes sequences in the test dataset that have at least one similar sequence in the training dataset, while a weight of 1 includes all the sequences (methods). Each color represents AUROC from one of the five folds in the cross-validation tests.



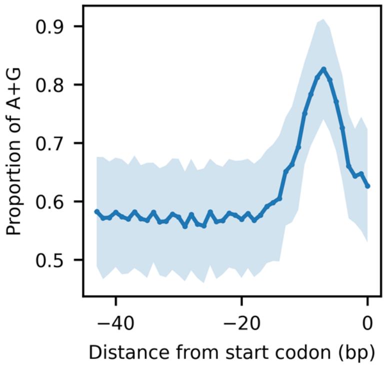
Supplementary Figure 10: Predicted gene numbers and densities for the generated sequences and the training dataset. a) Comparison of the number of predicted genes in generated sequences (sample size: n = 607) versus those in the training dataset (sample size: n = 99,673). **b**) Gene density distributions for the generated and training sequences. Gene density is defined as the ratio of gene numbers and the sequence length. The central line inside the box represents the median value. The top and bottom borders of the box represent the third (upper) and first (lower) quartiles, respectively.



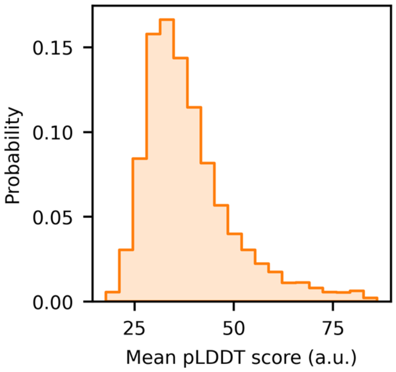
Supplementary Figure 11. Example of a generated sequence. a) Functional annotation of a selected sequence fragment (generated sequence #212). **b**) Predicted promoter activity for all the 5'UTRs in the generated sequence (orange, sample size: n = 45), along with the promoter activity of the random sequences with the same length (green). Promoter activities were calculated using the Promoter Calculator[22]. Two-sided Kolmogorov-Smirnov test: p value = $6.8x10^{-15}$. **c**) Proportions of adenine (A) and guanine (G) nucleotides preceding the start codon of the predicted genes in the generated sequence #212.



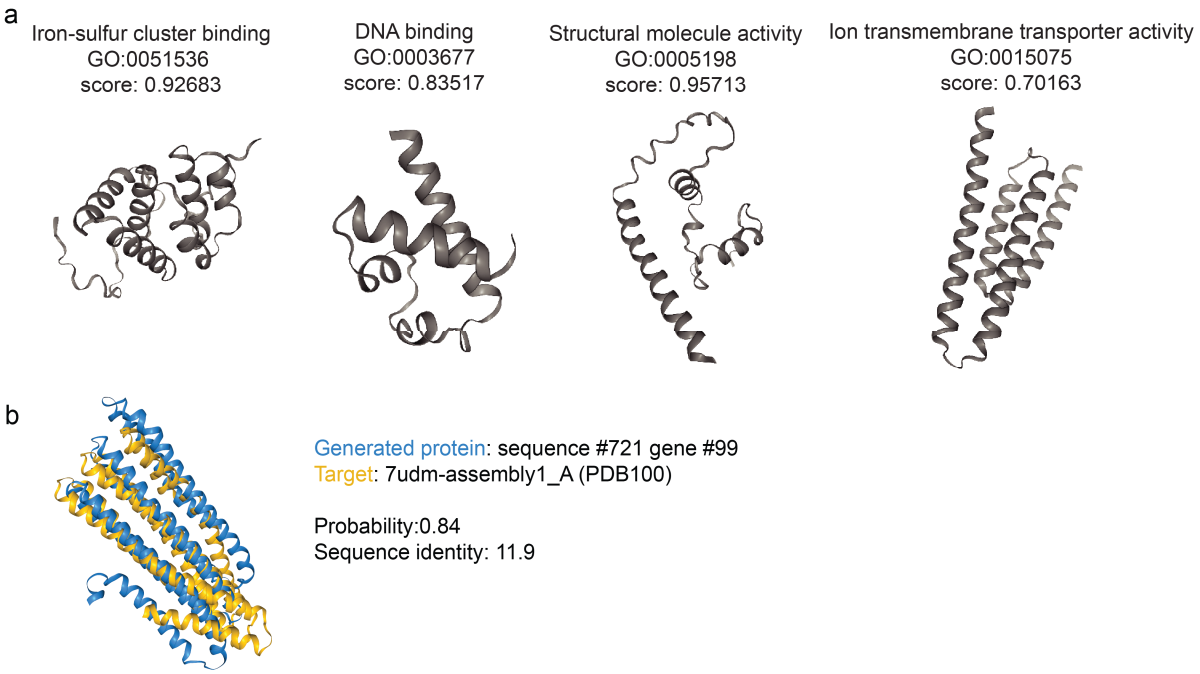
<u>Supplementary Figure 12:</u> **Predicted promoter activity for 5’UTRs in generated sequence and its relationship to the virus score**. **a**) Histograms showing the distribution of predicted promoter activities for the 5’UTRs in all generated sequences (sample size: n = 49,931, orange) and for an equal number of random sequences with the same length (green). Promoter activities were calculated using the Promoter Calculator[22] . **b**) Correlation of the promoter activity and the virus score for the generated sequences. The promoter activity is reported as the difference in medians for random sequences and generated promoters. Each dot represents one generated sequence, and the Spearman correlation coefficient is -0.15.



<u>Supplementary Figure 13:</u> Proportions of adenine (A) and guanine (G) nucleotides preceding the start codon for all the generated sequences. Blue line denotes the mean A+G nucleotides proportion profile for all the generated sequences with a virus score larger than zero (sample size: n = 607). The shaded region represents the standard derivation of all profiles.

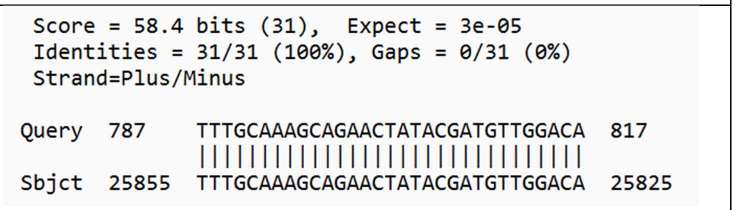
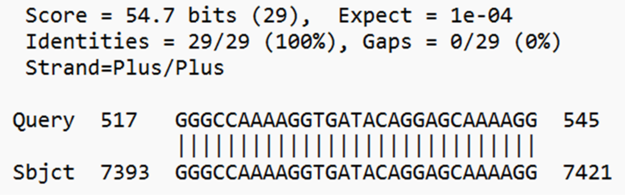
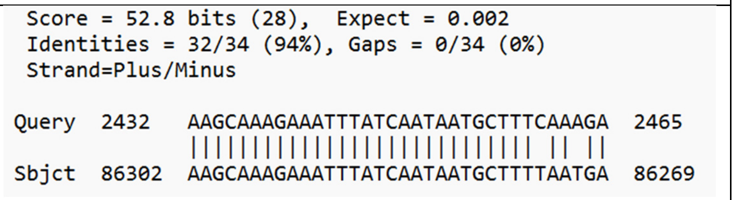
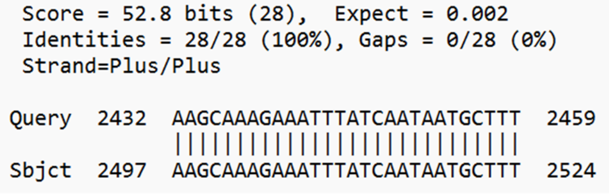


<u>Supplementary Figure 14:</u> **Mean pLDDT scores for proteins derived from the generated sequences.** The distribution of mean pLDDT score for a randomly sampled subset of all the generated proteins is shown (sample size: n = 10,000; median value: 36).


a
Iron-sulfur cluster binding
GO:0051536
score: 0.92683
DNA binding
GO:0003677
score: 0.83517
Structural molecule activity
GO:0005198
score: 0.95713
Ion transmembrane transporter activity
GO:0015075
score: 0.70163
b
Generated protein: sequence #721 gene #99
Target: 7udm-assembly1_A (PDB100)
Probability:0.84
Sequence identity: 11.9

Supplementary Figure 15: Representative proteins from the generated sequence with predicted functions and structures. a) The protein structures were predicted using ESMfold[34] and the functions were annotated using deepFRI[28]. Predicted scores and GO terms from deepFRI are shown. **b)** Alignment of a generated protein to the PDB database using Foldseek. The structure of sequence #721 gene #99 was predicted using ESMfold[34] with default parameters. The predicted structure was then searched against the PDB100 database via the Foldseek online server in 3Di/AA mode (https://search.foldseek.com/search)[26].

<u>Supplementary Table 1:</u> BLAST analysis of the generated genes. We conducted BLAST analysis to compare the generated genes with geNomad markers (n = 343) against the training dataset. Using default settings, we found hits for 3 out of the 343 genes. Only one hit is shown if there are multiple identical matches across reference genomes.

Query	Reference	BLAST results
Generated sequence #202, gene #97 (length = 2,346)	MGV-GENOME-0232742 (length = 33,859)	 Score = 58.4 bits (31), Expect = 3e-05 Identities = 31/31 (100%), Gaps = 0/31 (0%) Strand=Plus/Minus Query 787 TTTGCAAAGCAGAACTATACGATGTTGGACA 817 ||||||||||||||||||||||||||||||| Sbjct 25855 TTTGCAAAGCAGAACTATACGATGTTGGACA 25825
Generated sequence #209, gene #102 (length = 888)	MN176228.1 Bacillus phage 049ML003 (length = 44,817)	 Score = 54.7 bits (29), Expect = 1e-04 Identities = 29/29 (100%), Gaps = 0/29 (0%) Strand=Plus/Plus Query 517 GGGCCAAAAGGTGATACAGGAGCAAAAGG 545 ||||||||||||||||||||||||||||| Sbjct 7393 GGGCCAAAAGGTGATACAGGAGCAAAAGG 7421
Generated sequence #561 gene #14 (length = 2,721)	MGV-GENOME-0359033 1-89291/89291 (length = 89,291)	 Score = 52.8 bits (28), Expect = 0.002 Identities = 32/34 (94%), Gaps = 0/34 (0%) Strand=Plus/Minus Query 2432 AAGCAAAGAAATTTATCAATAATGCTTTCAAAGA 2465 |||||||||||||||||||||||||||| || || Sbjct 86302 AAGCAAAGAAATTTATCAATAATGCTTTTAATGA 86269
	MGV-GENOME-4432828 (length = 88,796)	 Score = 52.8 bits (28), Expect = 0.002 Identities = 28/28 (100%), Gaps = 0/28 (0%) Strand=Plus/Plus Query 2432 AAGCAAAGAAATTTATCAATAATGCTTT 2459 |||||||||||||||||||||||||||| Sbjct 2497 AAGCAAAGAAATTTATCAATAATGCTTT 2524