Predicting microbial transcriptome using genome sequence

Guohao Fu[1, 4], Yujing Yan[2, 4], Bin Shao[3, 4 *]

[1] Xuteli School, Beijing Institute of Technology, Beijing 100081, P. R. China

[2]Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Harvard University, Cambridge MA 02138, USA

[3] Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, P. R. China

[4] These authors contribute equally to this work.

*Correspondence should be addressed to Bin Shao (shaobinlx@gmail.com)

Abstract: We present TXpredict, a transformer-based framework for predicting microbial transcriptomes using annotated genome sequences. By leveraging information learned from a large protein language model, TXpredict achieves an average Spearman correlation of 0.53 in predicting gene expression for new bacterial genomes. We further extend this framework to predict transcriptomes for 900 additional microbial genomes spanning 276 genera, 69% of which remain uncharacterized at the transcriptional level. Additionally, TXpredict enables the prediction of condition-specific gene expression, providing a powerful tool for understanding microbial adaptation and facilitating the rational design of gene regulatory sequences.

For a newly sequenced microbe, transcriptome profiling provides critical insights into its adaptation, survival, and pathogenicity, and bridges the gap between raw genomic data and functional biology[1–6]. Despite its importance, only a small fraction of sequenced microbes have been characterized at the transcriptional level. A major challenge is that most microbes remain uncultivated due to the limited knowledge of their growth requirements or their inhabitation of extreme environments[7]. Recent advances in predicting physicochemical requirements have begun to address these challenges, representing an important first step[8–10]. Moreover, even when cultivation is successful, RNA-seq protocols often require organism-specific optimizations, such as cell lysis and rRNA depletion, which can involve extensive trial and error for novel organisms. These technical issues hinder comprehensive studies of microbial gene expression and the engineering of functional components in non-model organisms.

Current machine learning approaches for predicting gene expression have primarily focused on analyzing untranslated regions (UTRs)[11–16] or large DNA windows[17,18]. While these methods show promise, their application to non-model organisms remains limited, as trained models often fail to generalize across species or even within the same species[19,20]. Additionally, previous studies have demonstrated evolutionary constraints on protein synthesis, such as conserved enzyme stoichiometry in metabolic pathways[21–23]. Yet these constraints have not been translated into predictive models that are broadly applicable to diverse microbial species.

Here we introduce TXpredict, a novel framework for predicting microbial transcriptomes using only annotated genome sequences. Our approach assumes that gene expression is shaped by evolutionary constraints on protein sequences and leverages a protein language model, ESM2[24], to extract predictive features. Using transcriptome data from 22 bacterial and 10 archaeal species obtained from the NCBI GEO database, we trained a transformer-based model that achieves a Spearman correlation up to 0.64 in leave-one-genome-out evaluations. This framework was further extended to predict transcriptomes for 900 additional genomes and characterize expression patterns across 3.11 million genes. Additionally, we trained the model to account for condition-dependent gene expression using data from 13 bacterial species across 4.6k unique conditions. Our trained model identifies sequence determinants of condition-dependent gene expression and facilitates promoter design. The models developed in this study are publicly available: https://github.com/lingxusb/TXpredict

We first collected microbial RNA-seq datasets from the NCBI GEO datasets and selected the dataset with the highest number of samples for each microbial species. Read counts were normalized with the same

pipeline (Methods). Taken together, our dataset includes 22 bacterial and 10 archaeal species and a total of 11.5M gene expression measurements (Supplementary Table 1). For each microbial species, gene expression was averaged across all samples to construct a standard expression profile. Protein embeddings for each gene were computed using the ESM2-650M model and combined with basic protein statistics to serve as the input features for TXpredict (Fig. 1a, Methods). For model evaluation, we employed a leave-one-genome-out cross-validation strategy. For bacterial species, the model was trained on genes from 21 species and tested on the held-out species. We followed the same approach for archaeal species. Model performance was evaluated using Spearman correlation coefficients between predicted and measured expression values for the held-out genome.

We found our model achieved a mean Spearman correlation coefficient of 0.53 for the bacterial species (Fig. 1b). It also performs well for non-model organisms, with correlation coefficients of 0.64, 0.62 and 0.62 for *B. hinzii, B. thetaiotaomicron*, and *C. beijerinckii*. It is worth noting that transcriptome predictions for these bacteria were based solely on the computational analysis of annotated genome sequences, which provides an efficient alternative to labor-intensive RNA-seq experiments. The model's performance in predicting gene expression variation decreased modestly as the variation calculations depend on the diversity of experimental conditions (Supplementary Fig. 1). For archaea, TXpredict achieved a mean Spearman correlation of 0.42, partly due to the limited availability of training data (Fig. 1b). However, the model demonstrated broad applicability to extremophiles from distinct natural niches (ranging from Dead Sea to volcanic springs), further highlighting the robustness of our method.

We then applied the trained TXpredict model to predict transcriptomes for new genomes including 600 bacterial species and 300 archaeal genomes (Fig. 1c and d). Collectively, these new genomes span 191 new bacterial genera and 85 archaeal genera, representing a 10-fold expansion compared to the training datasets (18 for bacteria and 9 for archaea, Fig. 1e, and Supplementary Fig. 2). Our model predicted the expression of up to 3.11M genes across these genomes, and the predicted gene expressions in bacteria and archaea followed similar distributions (Fig. 1f). To explore the relationship between predicted gene expression and protein function, we used DeepFRI[25] to predict Gene Ontology (GO) terms for the new genes (Supplementary Fig. 3). Among the GO terms with the highest predicted expression levels were structural constituents of ribosome (GO:0003735) and tRNA ligase (GO:0004812), in accord with their essential role in protein synthesis (Fig. 1g). In contrast, GO terms with the lowest predicted expression included proteins with transporter activity (GO:0005215), which are typically expressed in a condition-dependent manner. We further applied tSNE analysis to visualize the protein embeddings of the 3.11M

genes (Fig. 1i, Supplemenarty Fig. 4). We observed a clear separation of bacterial and archaeal genes. Small clusters of ribosomal genes located at the periphery of the plot and showed high predicted expression, while transmembrane transporter proteins formed distinct low-expression clusters. To investigate whether the predicted gene expression levels depend on gene functions, we examined the measured and predicted expression levels for genes in *B. hinzii* from the leave-one-genome-out test (Fig. 1h). We found that proteins with the same GO terms showed high variability in their expression and our model successfully captured this variation. This observation suggests that our model learns the intrinsic determinant of gene expression that is independent of protein function.

We next extended our framework to model condition-dependent gene expression (Fig. 2a). We collected RNA-seq profiles from 13 bacterial strains across 4.6k unique conditions from iModulonDB, resulting in a total of 12.0M measurements (Supplementary Table 2, Fig. 2b). The model integrated one-hot encoded 5'UTR sequences with protein embeddings as input features. For each gene, we calculated the correlation between measured and predicted expression values across all conditions. Our TXpredict model achieved an average correlation coefficient of 0.52 across all bacterial strains (Fig. 2b). For *E. coli*, the distributions of Pearson's and Spearman's correlation coefficients show similar patterns, with 41% and 38% of genes having correlation coefficients above 0.5 (Fig. 2c).

To further interpret the model's predictions, we calculated input gradients[17] to identify sequence motifs contributing to gene expression variation across conditions (Fig. 2d). In *E. coli*, we observed a peak in the input × gradient overlap with the -10 box of the sigmaS-dependent promoter of the *yadV* gene[26]. Also input gradients increased after the transcriptional start site, suggesting potential contributions from transcribed sequences (Fig. 2e). In a non-model organism *P. putida*, we identified a peak overlapping with a Zur binding motif for the *PP_0120* gene, consistent with its role as a Zur iModulon gene[1]. These findings highlight our model's ability to capture condition-specific gene expression changes in an interpretable manner.

Lastly, we utilized the trained TXpredict model to design condition-dependent promoters (Fig. 2f). The design process involved iterative rounds of mutation introduction, where we used the model to predict expression levels across conditions and selected promoters showing increased expression in the target condition relative to the reference condition. Through this optimization process, we successfully generated promoters with significantly higher predicted expression under *Fur* deletion (Fig. 2g, and Supplementary Fig. 5). Notably, the optimization efficiency depended on the initial expression difference between conditions rather than their overall expression correlation (Fig. 2h, and Supplementary Fig. 5).

119 In summary, we present TXpredict, a transcriptome prediction model that can generalize to a wide
120 range of microbes. For a microbial genome with 4.6k genes, transcriptome prediction can be completed
121 in approximately 22 minutes using a Google Colab notebook (Methods), demonstrating the model's
122 efficiency and accessibility. Our approach is not without limitations: while it effectively leverages protein
123 embeddings and 5'UTR sequences, it does not account for other regulatory elements or non-coding
124 RNAs that might influence gene expression. Moreover, our model's performance is constrained by the
125 limited publicly available RNA-seq datasets, especially for archaeal species. Despite these limitations,
126 TXpredict has the potential to advance the study of the vast number of sequenced yet poorly
127 characterized microbes. For example, identifying conserved expressed genes in pathogens may enable
128 the discovery of novel drug targets. Predicting the transcriptome of unculturable microbes also provides
129 critical insights into microbial adaptation to extreme environments and the evolution of gene regulation
130 across prokaryotes. These efforts could further benefit from integrating single-cell RNA-seq data [5,27,28]
131 and computational tools for mRNA synthesis and stability prediction[29–31], paving the way for a deeper
132 understanding of microbial transcriptional landscapes.

Figures

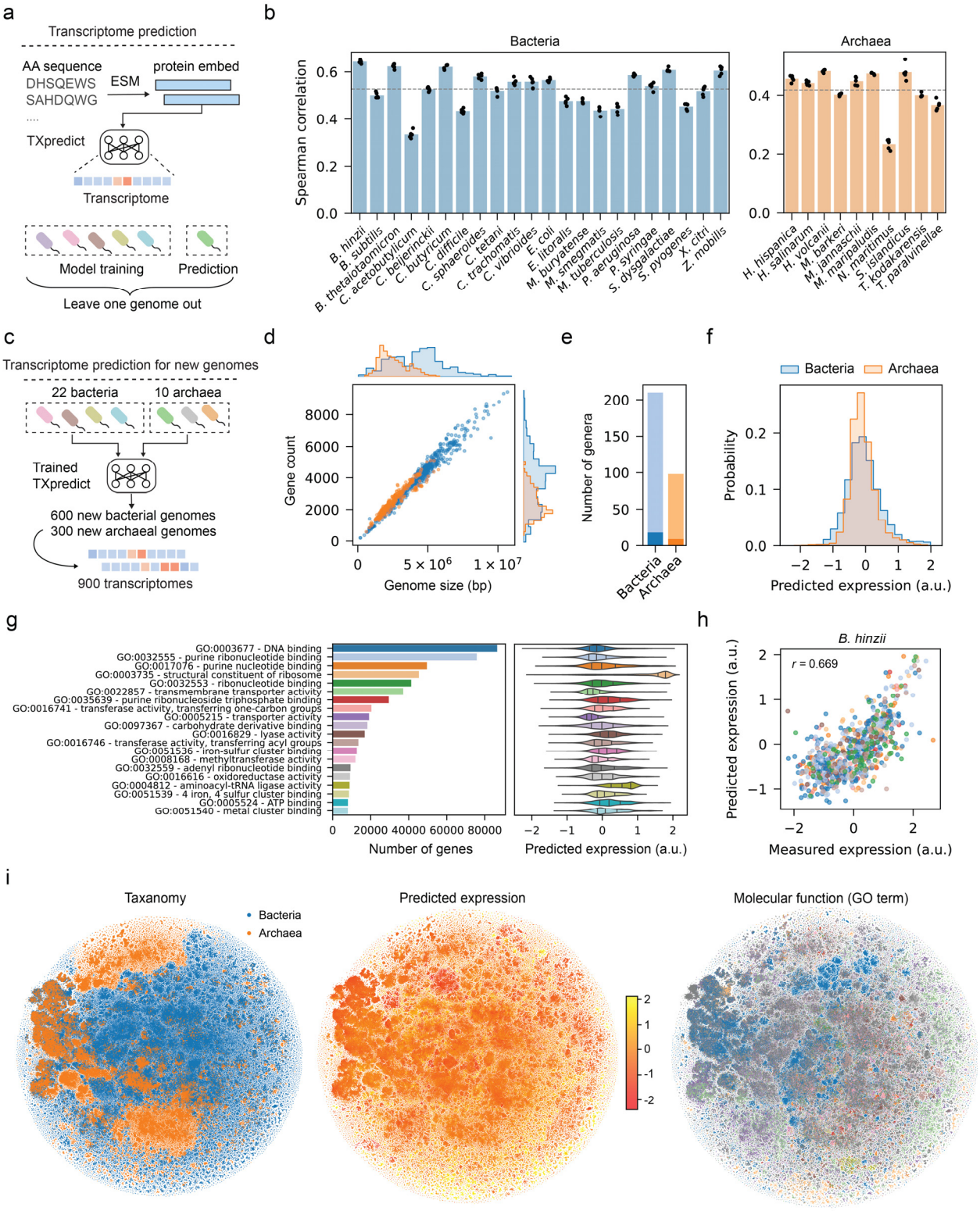


Figure 1. Prediction of transcriptome for new genome.

136 **a)** Schematic of TXpredict for transcriptome prediction. **b)** Spearman correlation coefficients of
137 predicted and measured gene expression in leave-one-genome-out scenarios. The TXpredict models
138 were separately trained for bacterial and archaeal species. Results from five rounds of model training
139 and testing are reported (n = 5). **c)** Transcriptome prediction for new genomes. **d)** Genome sizes and
140 gene counts for the new genomes (n = 900). **e)** Number of genera in the new genomes (light colors)
141 compared to training genomes (dark colors). **f)** Distributions of predicted gene expression levels for new
142 bacterial genomes and archaeal genomes (n = 2.34M and 0.77M). **g)** Top 20 GO terms of the genes in
143 the new genomes calculated by the DeepFRI model[25] (left) and their predicted gene expression levels
144 (right). **h)** Predicted and measured gene expression in the leave-one-genome-out test for *B. hinzii.* Genes
145 associated with the top 20 GO terms are shown and colored as in g). Spearman correlation coefficient
146 between measured and predicted expression values is indicated in the top left corner. **i)** t-SNE plots of
147 the protein embeddings calculated by the ESM2 model[24] (n = 3.11M). Each dot represents a gene in the
148 new genomes, colored by taxa (left), predicted gene expression (middle), and GO terms (right).

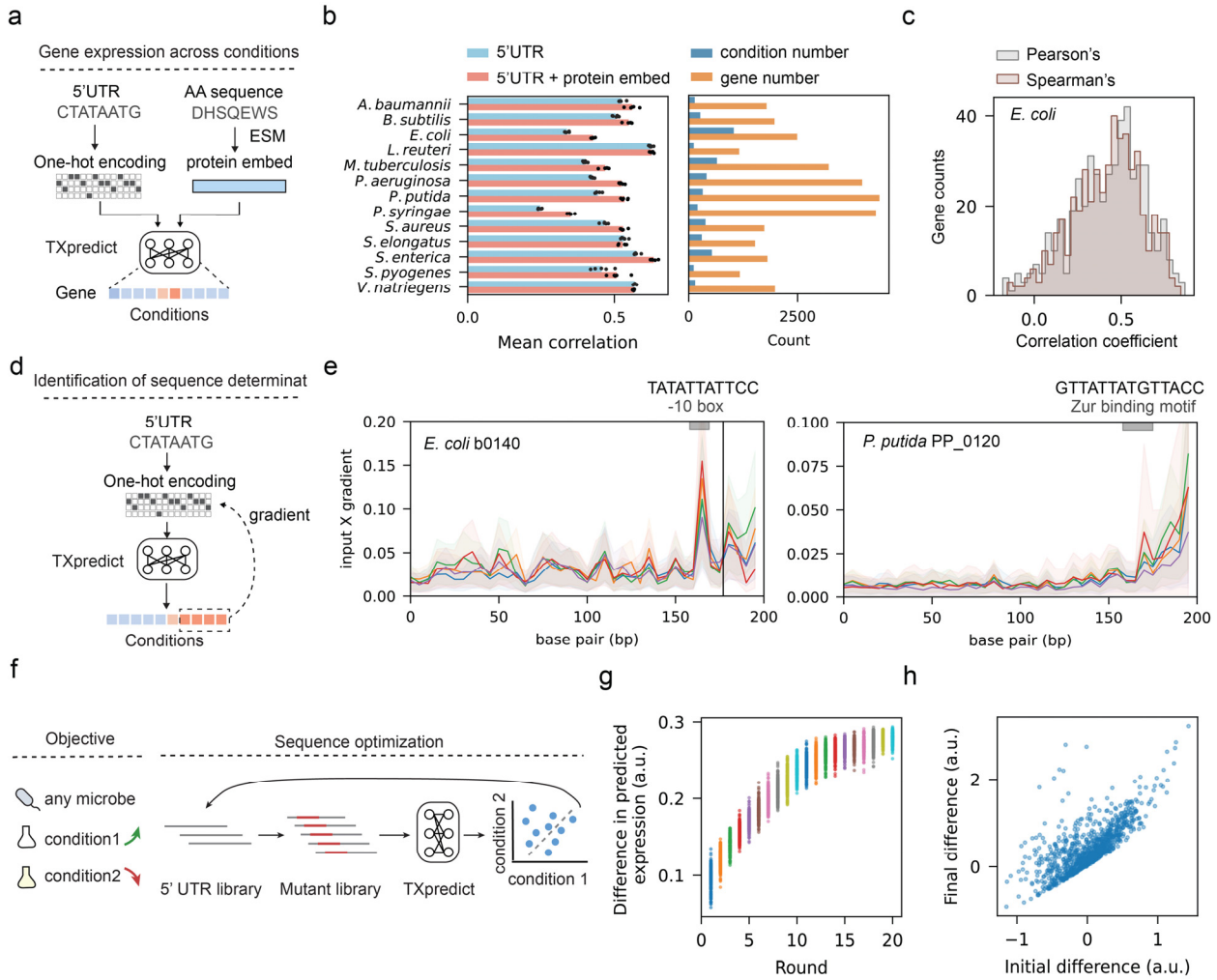


Figure 2. Prediction of condition-specific gene expression.

a) Schematic of TXpredict for condition-specific gene expression prediction. **b)** Mean correlations for models trained separately on different bacterial species. Correlations from 5-fold cross-validation tests are shown on the left (n = 5). Gene and condition counts for each species are shown on the right. **c)** Distribution of Spearman and Pearson correlation coefficients for gene expression across conditions in *E. coli* (n = 498). **d)** Calculation of input gradient for identifying sequence determinants of gene expression. **e)** Two examples of input × gradients and their correlation with known sequence motifs. **f)** Pipeline for the design of condition-specific promoters. **g)** Difference of predicted gene expression between two conditions for each round of optimization (n = 200 for each round). **h)** Difference of gene expression at the beginning and the end of the optimization. Each dot represents a randomly selected condition combination for *E. coli* (n = 1,000).

Methods

Training dataset

We collected RNA-seq data from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). For each microbe species, we selected the dataset with the highest number of samples and used either raw count data or RPKM (Reads Per Kilobase of transcript per Million mapped reads) values when available. For raw count data, we calculated the RPKM for each gene, followed by log (1 + p) and z-score normalization. For RPKM values, we applied log (1 + p) transformation and z-score normalization. In total, we gathered data from 22 bacterial species (2,649 samples) and 10 archaeal species (176 samples) (Supplementary Table 1). For each species, we constructed a standard gene expression profile by averaging the normalized expression values across all available samples.

To model gene expression across conditions, we also collected RNA-seq data from iModulonDB (https://imodulondb.org/index.html). We downloaded the processed datasets for each microbial species, resulting in a dataset that includes 13 bacterial species and a total of 4,619 conditions. For each microbial species, we only kept the genes with mean normalized gene expressions greater than -3 and a standard deviation less than 1 across conditions.

Protein embeddings for all annotated genes were calculated using the ESM2-650M model (*esm.pretrained.esm2_t33_650M_UR50D()*), following the instructions provided at https://github.com/facebookresearch/esm. Proteins longer than 1,500 amino acids (AAs) were excluded from this analysis. In addition to the protein embeddings, we also included basic protein sequence statistics in the model input: the normalized protein length (calculated by dividing each protein length by the longest protein length within each organism) and the amino acid composition (the proportion of each of the 20 amino acids). The final input dimension of the model is 1301.

Transcriptome prediction model

We developed a Transformer-based model to predict gene expression levels from sequence features. The transformer encoder utilizes a multi-head self-attention mechanism (num_heads=4) with a single encoder layer. The encoder processes input features through a linear transformation into a hidden dimension of 128, followed by positional encoding. We implemented sinusoidal positional encoding to capture sequential information. The output of the transformer layer was then fed into two fully

connected layers (dimension 512 and 128) with ReLU activations. The final linear layer produces a single output value representing the predicted gene expression level.

Model training and evaluation

We performed separate cross-species evaluations for bacteria and archaea to account for potential domain-specific expression patterns. For each domain, we employed a cross-validation strategy where models were trained on data from all but one microbial species and tested on the held-out species. This approach was repeated for each species in our dataset to assess generalization capability. The models were trained using the Adam optimizer with a learning rate of 0.0002 for 5 epoch with a batch size of 32. To ensure robust evaluation, we performed 5 independent training runs with different random seeds (42, 123, 456, 789, 1011). The mean squared error (MSE) loss function was used for optimization. Model performance was evaluated using both Spearman and Pearson correlation coefficients between predicted and measured expression levels. New bacteria and archaea genomes were downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/genbank/ and we randomly selected annotated bacterial and archaea genomes for transcriptome predictions. To determine whether a genus has been transcriptional characterized, we searched the GEO database using the following query: "*genus_name [Organism] AND ("Expression profiling by array"[Filter])*". For model training and inference, we used 3090 Ti GPU (24GB) and the software packages PyTorch (version 2.1.1) and transformer (version 4.28.1).

Conditional gene expression model

The conditional gene expression model shares the same transformer encoder architecture with the transcriptome prediction model, and trained separately for each species. The UTR branch processes one-hot encoded sequences from -100 to +100 bp relative to the transcription start site. The transformer layer has 2 attention heads and a feedforward dimension of 128. The protein embedding branch utilizes 8 attention heads with a hidden dimension of 256. The outputs from both branches are concatenated together and processed through a fully connected layer (dimension: 4096) with ReLU activation. The output dimension of the model is the number of conditions for each bacterial species.

For each bacterial species, we trained the model for 10 epochs using the Adam optimizer with a learning rate of 0.0002 and a batch size of 32. Model performance was evaluated using 5-fold cross-validation.

219 Pearson correlation coefficients were calculated between predicted and measured expression levels
220 across all conditions for genes in the test set.

221

222 **Analysis of protein sequences**

223 To visualize the high-dimensional embeddings in a two-dimensional space, we implemented t-SNE (t-
224 Distributed Stochastic Neighbor Embedding) using the *opentsne* package. We initialized the embedding
225 using PCA and constructed a uniform affinity matrix with 10 nearest neighbors. The optimization process
226 consisted of three phases: an early exaggeration phase of 125 iterations with an exaggeration factor of
227 12 and momentum of 0.5, followed by a gradual exaggeration annealing phase where the exaggeration
228 factor was linearly decreased from 12 to 1 over 125 iterations with momentum of 0.8. The final
229 optimization phase ran for 2000 iterations without exaggeration. Function prediction for these proteins
230 was carried out using the default deepFRI model[25], as available on GitHub
231 (https://github.com/flatironinstitute/DeepFRI). We used a score cutoff of 0.5 which was reported to be
232 significant in the original publication.

233

234 **Calculation of input gradient**

235 To investigate the contribution of individual nucleotides in the UTR sequence to gene expression
236 prediction across conditions, we implemented a gradient-based analysis. Using a sequence-only model
237 variant where the protein embedding branch was masked, we calculated the gradient of the model
238 output with respect to the input DNA sequence. For each target gene and condition, we performed a
239 forward pass with the one-hot encoded UTR sequence while setting all other inputs to zero. We then
240 computed the gradients through backpropagation and multiplied them element-wise with the input
241 sequence (Gradient × Input method)[17]. The absolute values of these products were summed across the
242 four nucleotide channels (A, T, C, G) to obtain single-base resolution importance scores. For visualization
243 purposes, we aggregated these scores into 5-bp bins by summing the scores within each bin.

244

245 **Condition specific promoter design**

We implemented an iterative sequence optimization algorithm to design UTR sequences with desired expression patterns. We started with the Anderson promoter J23101 (https://parts.igem.org/Part:BBa_J23101). In each iteration, we generated 200 candidate sequences by randomly mutating 20 nucleotides within the target region (positions 0-180) of each sequence. These sequences were encoded using one-hot encoding and evaluated using our trained model. We calculated a differential expression score by subtracting the predicted expression levels between two target conditions. The top 10 sequences with the highest differential expression scores were selected as templates for the next iteration, with each template generating 20 new variants through random mutations. This process was repeated for 20 iterations to optimize the sequences for condition-specific expression patterns.

Code availability

Our trained model and inference codes are available from GitHub: https://github.com/lingxusb/TXpredict

References

1. Lim, H. G. *et al.* Machine-learning from Pseudomonas putida KT2440 transcriptomes reveals its transcriptional regulatory network. *Metab Eng* **72**, 297–310 (2022).

2. Patel, A. *et al.* Proteome allocation is linked to transcriptional regulation through a modularized transcriptome. *Nat Commun* **15**, 5234 (2024).

3. Ryan, D. *et al.* An expanded transcriptome atlas for Bacteroides thetaiotaomicron reveals a small RNA that modulates tetracycline sensitivity. *Nat Microbiol* **9**, 1130–1144 (2024).

4. Shin, J., Rychel, K. & Palsson, B. O. Systems biology of competency in Vibrio natriegens is revealed by applying novel data analytics to the transcriptome. *Cell Rep* **42**, (2023).

5. Ma, P. *et al.* Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. *Cell* **186**, 877–891 (2023).

6. Ryu, M.-H. *et al.* Control of nitrogen fixation in bacteria that associate with cereals. *Nat Microbiol* **5**, 314–330 (2020).

7. G, L. K., D, S. A., Joshua, L., Junqi, Y. & Lonnie, C. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**, 10.1128/msystems.00055-18 (2018).

8. Barnum, T. P. *et al.* Predicting microbial growth conditions from amino acid composition. *bioRxiv* 2023–2024 (2024).

9. Ramoneda, J. *et al.* Building a genome-based understanding of bacterial pH preferences. *Sci Adv* **9**, eadf8998 (2024).

10. Price, M. N., Deutschbauer, A. M. & Arkin, A. P. Filling gaps in bacterial catabolic pathways with computation and high-throughput genetics. *PLoS Genet* **18**, e1010156- (2022).

11. Zrimec, J. *et al.* Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun* **11**, 6141 (2020).

12. Peleke, F. F., Zumkeller, S. M., Gültas, M., Schmitt, A. & Szymański, J. Deep learning the cis-regulatory code for gene expression in selected model plants. *Nat Commun* **15**, 3488 (2024).

13. Dudnyk, K., Cai, D., Shi, C., Xu, J. & Zhou, J. Sequence basis of transcription initiation in the human genome. *Science (1979)* **384**, eadj0116 (2024).

14. LaFleur, T. L., Hossain, A. & Salis, H. M. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat Commun* **13**, 5159 (2022).

15. Urtecho, G., Tripp, A. D., Insigne, K. D., Kim, H. & Kosuri, S. Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry* **58**, 1539–1551 (2019).

16. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature* (2022) doi:10.1038/s41586-022-04506-6.

295 17. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range
296 interactions. *Nat Methods* **18**, 1196–1203 (2021).

297 18. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with
298 convolutional neural networks. *Genome Res* **28**, 739–750 (2018).

299 19. Huang, C. *et al.* Personal transcriptome variation is poorly explained by current genomic deep
300 learning models. *Nat Genet* **55**, 2056–2059 (2023).

301 20. Sasse, A. *et al.* Benchmarking of deep neural networks for predicting personal gene expression
302 from DNA sequence highlights shortcomings. *Nat Genet* **55**, 2060–2064 (2023).

303 21. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying Absolute Protein Synthesis Rates
304 Reveals Principles Underlying Allocation of Cellular Resources. *Cell* **157**, 624–635 (2014).

305 22. Lalanne, J.-B. *et al.* Evolutionary Convergence of Pathway-Specific Enzyme Expression
306 Stoichiometry. *Cell* **173**, 749-761.e38 (2018).

307 23. Klumpp, S., Scott, M., Pedersen, S. & Hwa, T. Molecular crowding limits translation and cell
308 growth. *Proceedings of the National Academy of Sciences* **110**, 16754–16759 (2013).

309 24. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language
310 model. *Science (1979)* **379**, 1123–1130 (2023).

311 25. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional
312 networks. *Nat Commun* **12**, 3168 (2021).

313 26. Raina, S., Missiakas, D., Baird, L., Kumar, S. & Georgopoulos, C. Identification and transcriptional
314 analysis of the Escherichia coli htrE operon which is homologous to pap and related pilin operons.
315 *J Bacteriol* **175**, 5009–5021 (1993).

316 27. Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.-E. & Vogel, J. Single-cell RNA-sequencing
317 reports growth-condition-specific global transcriptomes of individual bacteria. *Nat Microbiol* **5**,
318 1202–1206 (2020).

319 28. Kuchina, A. *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science (1979)*
320 **371**, eaba5257 (2021).

321 29. LaFleur, T. L., Hossain, A. & Salis, H. M. Automated model-predictive design of synthetic
322 promoters to control transcriptional profiles in bacteria. *Nat Commun* **13**, 5159 (2022).

323 30. Cetnar, D. P., Hossain, A., Vezeau, G. E. & Salis, H. M. Predicting synthetic mRNA stability using
324 massively parallel kinetic measurements, biophysical modeling, and machine learning. *Nat*
325 *Commun* **15**, 9601 (2024).

326 31. Xia, Y. *et al.* Language model generates cis-regulatory elements across prokaryotes. *bioRxiv*
327 2011–2024 (2024).

328 32. Jeckel, H. *et al.* Simultaneous spatiotemporal transcriptomics and microscopy of Bacillus subtilis
329 swarm development reveal cooperation across generations. *Nat Microbiol* **8**, 2378–2391 (2023).

330 33. Rachael, N., Swathi, P., S, P. F. & Peter, B. Metabolic changes associated with polysaccharide
331 utilization reduce susceptibility to some β-lactams in Bacteroides thetaiotaomicron. *mSphere* **9**,
332 e00103-24 (2024).

333 34. Kamat, A. *et al.* Widespread prevalence of a methylation-dependent switch to activate an
334 essential DNA damage response in bacteria. *PLoS Biol* **22**, e3002540- (2024).

335 35. Lakey, B. D. *et al.* The essential Rhodobacter sphaeroides CenKR two-component system
336 regulates cell division and envelope biosynthesis. *PLoS Genet* **18**, e1010270- (2022).

337 36. A, P. D. *et al.* Exogenous butyrate inhibits butyrogenic metabolism and alters virulence
338 phenotypes in Clostridioides difficile. *mBio* **15**, e02535-23 (2024).

339 37. D, S. M. *et al.* Arabinose-Induced Catabolite Repression as a Mechanism for Pentose Hierarchy
340 Control in Clostridium acetobutylicum ATCC 824. *mSystems* **3**, 10.1128/msystems.00064-18
341 (2018).

342 38. Rahimi, S. *et al.* Co-culturing Bacillus subtilis and wastewater microbial community in a bio-
343 electrochemical system enhances denitrification and butyrate formation. *Chemical Engineering*
344 *Journal* **397**, 125437 (2020).

345 39. Pennings, J. L. A. *et al.* Regulation of Clostridium tetani Neurotoxin Expression by Culture
346 Conditions. *Toxins (Basel)* **14**, 31 (2022).

347 40. Espah Borujeni, A., Zhang, J., Doosthosseini, H., Nielsen, A. A. K. & Voigt, C. A. Genetic circuit
348 characterization by inferring RNA polymerase movement and ribosome usage. *Nat Commun* **11**,
349 5001 (2020).

350 41. Wilson, E. H. *et al.* A Computational Framework for Identifying Promoter Sequences in Nonmodel
351 Organisms Using RNA-seq Data Sets. *ACS Synth Biol* **10**, 1394–1405 (2021).

352 42. Young, A. T. *et al.* Multi-omic regulatory networks capture downstream effects of kinase
353 inhibition in Mycobacterium tuberculosis. *NPJ Syst Biol Appl* **7**, 8 (2021).

354 43. Xavier, C. *et al.* Multisystem Analysis of Mycobacterium tuberculosis Reveals Kinase-Dependent
355 Remodeling of the Pathogen-Environment Interface. *mBio* **9**, 10.1128/mbio.02333-17 (2018).

356 44. Wu, M.-L. *et al.* Developmental transcriptome of resting cell formation in Mycobacterium
357 smegmatis. *BMC Genomics* **17**, 837 (2016).

358 45. Nobori, T. *et al.* Transcriptome landscape of a bacterial pathogen under plant immunity.
359 *Proceedings of the National Academy of Sciences* **115**, E3055–E3064 (2018).

360 46. M, E. J. *et al.* Integrative genomic, virulence, and transcriptomic analysis of emergent
361 Streptococcus dysgalactiae subspecies equisimilis (SDSE) emm type stG62647 isolates causing
362 human infections. *mBio* **15**, e02578-24 (2024).

363 47. Kachroo, P. *et al.* Integrated analysis of population genomics, transcriptomics and virulence
364 provides novel insights into Streptococcus pyogenes pathogenesis. *Nat Genet* **51**, 548–559
365 (2019).

366 48. Martim, D. B. *et al.* Resolving the metabolism of monolignols and other lignin-related aromatic
367 compounds in Xanthomonas citri. *Nat Commun* **15**, 7994 (2024).

368 49. M, V. J. *et al.* Genome-Scale Transcription-Translation Mapping Reveals Features of Zymomonas
369 mobilis Transcription Units and Promoters. *mSystems* **5**, 10.1128/msystems.00250-20 (2020).

370 50. Sakrikar, S. & Schmid, A. K. An archaeal histone-like protein regulates gene expression in
371 response to salt stress. *Nucleic Acids Res* **49**, 12732–12743 (2021).

372 51. Hackley, R. K. *et al.* TbsP and TrmB jointly regulate gapII to influence cell development
373 phenotypes in the archaeon Haloferax volcanii. *Mol Microbiol* **121**, 742–766 (2024).

374 52. D, T. B., Cem, M., B, N. T., Q, L. S. & F, H. J. Growth Kinetics, Carbon Isotope Fractionation, and
375 Gene Expression in the Hyperthermophile Methanocaldococcus jannaschii during Hydrogen-
376 Limited Growth and Interspecies Hydrogen Transfer. *Appl Environ Microbiol* **85**, e00180-19
377 (2019).

378 53. Payne, D. *et al.* Alternative sources of molybdenum for Methanococcus maripaludis and their
379 implication for the evolution of molybdoenzymes. *Commun Biol* **7**, 1337 (2024).

380 54. Spietz, R. L. *et al.* Investigating Abiotic and Biotic Mechanisms of Pyrite Reduction. *Front*
381 *Microbiol* **13**, (2022).

382 55. Yang, Y. *et al.* A novel RHH family transcription factor aCcr1 and its viral homologs dictate cell
383 cycle progression in archaea. *Nucleic Acids Res* **51**, 1707–1723 (2023).

384 56. Topçuoğlu, B. D., Meydan, C., Orellana, R. & Holden, J. F. Formate hydrogenlyase and formate
385 secretion ameliorate H2 inhibition in the hyperthermophilic archaeon Thermococcus
386 paralvinellae. *Environ Microbiol* **20**, 949–957 (2018).

387 57. D, M. N. *et al.* Independent component analysis reveals 49 independently modulated gene sets
388 within the global transcriptional regulatory architecture of multidrug-resistant Acinetobacter
389 baumannii. *mSystems* **9**, e00606-23 (2024).

390 58. Lamoureux, C. R. *et al.* A multi-scale expression and regulation knowledge base for Escherichia
391 coli. *Nucleic Acids Res* **51**, 10176–10193 (2023).

392 59. Jonathan, J.-S. *et al.* Reconstructing the transcriptional regulatory network of probiotic L. reuteri
393 is enabled by transcriptomics and machine learning. *mSystems* **9**, e01257-23 (2024).

394 60. Reo, Y. *et al.* Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a
395 Structured Interplay between Metabolism, Stress Response, and Infection. *mSphere* **7**, e00033-22
396 (2022).

397 61. Rajput, A. *et al.* Advanced transcriptomic analysis reveals the role of efflux pumps and media
398 composition in antibiotic responses of Pseudomonas aeruginosa. *Nucleic Acids Res* **50**, 9675–
399 9688 (2022).

400 62. Heera, B., Kevin, R., R, L. C., V, S. A. & O, P. B. Machine learning uncovers the Pseudomonas
401 syringae transcriptome in microbial communities and during infection. *mSystems* **8**, e00437-23
402 (2023).

403 63. Saugat, P. *et al.* Coordination of CcpA and CodY Regulators in Staphylococcus aureus USA300
404 Strains. *mSystems* **7**, e00480-22 (2022).

405 64. Yuan, Y. *et al.* Machine learning reveals the transcriptional regulatory network and circadian
406 dynamics of Synechococcus elongatus PCC 7942. *Proceedings of the National Academy of*
407 *Sciences* **121**, e2410492121 (2024).

408 65. Yujiro, H. *et al.* Elucidation of independently modulated genes in Streptococcus pyogenes reveals
409 carbon sources that control its expression of hemolytic toxins. *mSystems* **8**, e00247-23 (2023).

410

411 **Supplementary Table**

412 **Supplementary Table S1. Datasets used in the transcriptome prediction model**

Bacteria				
Species	**GSE number**	**Reference genome**	**Gene#**	**Condition#**
Bacillus subtilis	GSE224332[32]	NZ_CP020102.1	4220	317
Bacteroides thetaiotaomicron	GSE251676[33]	GCF_000011065.1	4618	64
Bordetella hinzii	GSE181542	CP052845.1	4531	228
Caulobacter vibrioides	GSE246782[34]	NC_011916.1	3882	34
Cereibacter sphaeroides	GSE186600[35]	GCA_000012905.2	4232	17
Chlamydia trachomatis	GSE262652	GCA_000068585.1	857	30
Clostridioides difficile	GSE249810[36]	GCA_000009205.5	3560	24
Clostridium acetobutylicum	GSE107804[37]	GCA_000008765.1	3755	48
Clostridium beijerinckii	GSE150480[38]	GCF_000833105.2	5316	16
Clostridium butyricum	GSE150480[38]	GCF_001456065.2	3848	16
Clostridium tetani	GSE182408[39]	NC_004557.1	2562	39
Escherichia coli	GSE152664[40]	NC_010473.1	3260	9
Erythrobacter litoralis	GSE126532	GCA_001719165.1	3003	49
Methylotuvimicrobium buryatense	GSE162089[41]	NZ_CP035467.1	3927	56
Mycobacterium tuberculosis	GSE110508[42,43]	GCA_000195955.2	3957	87
Mycolicibacterium smegmatis	GSE69983[44]	GCA_000015005.1	6700	30
Pseudomonas aeruginosa	GSE251671	GCA_000014625.1	5659	858
Pseudomonas syringae	GSE103441[45]	GCA_000007805.1	5586	114
Streptococcus dysgalactiae	GSE272047[46]	GCA_029234115.1	1939	114
Streptococcus pyogenes	GSE113055[47]	CP000056.2	1608	442
Xanthomonas citri	GSE252662[48]	AE008923.1	4294	38
Zymomonas mobilis	GSE139939[49]	CP023715.1	1759	19
Summary			83,073	2,649

413 **Supplementary Table S1 - continue. Datasets used in the transcriptome prediction model**

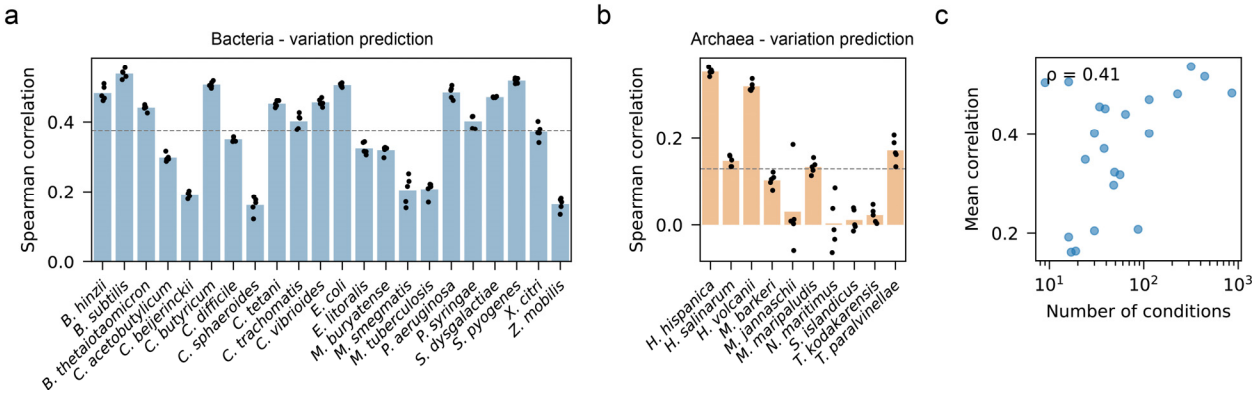
Archaea				
Species	**GSE number**	**Reference genome**	**Gene#**	**Condition#**
Haloarcula hispanica	GSE227031	GCF_000223905.1	3776	32
Halobacterium salinarum	GSE182493[50]	GCF_000006805.1	2325	24
Haloferax volcanii	GSE248564[51]	GCF_000025685.1	3873	39
Methanocaldococcus jannaschii	GSE112986[52]	GCF_000091665.1	1724	13
Methanococcus maripaludis	GSE261517[53]	GCF_000011585.1	1730	12
Methanosarcina barkeri	GSE168895[54]	GCF_000970025.1	3338	12
Nitrosopumilus maritimus	GSE103699	GCA_000018465.1	1791	7
Sulfolobus islandicus	GSE220819[55]	GCF_000189555.1	2682	18
Thermococcus kodakarensis	GSE186817	NC_006624.1	2214	4
Thermococcus paralvinellae	GSE101078[56]	GCF_000517445.1	1995	15
Summary			25,448	176

414 **Supplementary Table S2. Datasets used in the conditional gene expression model**

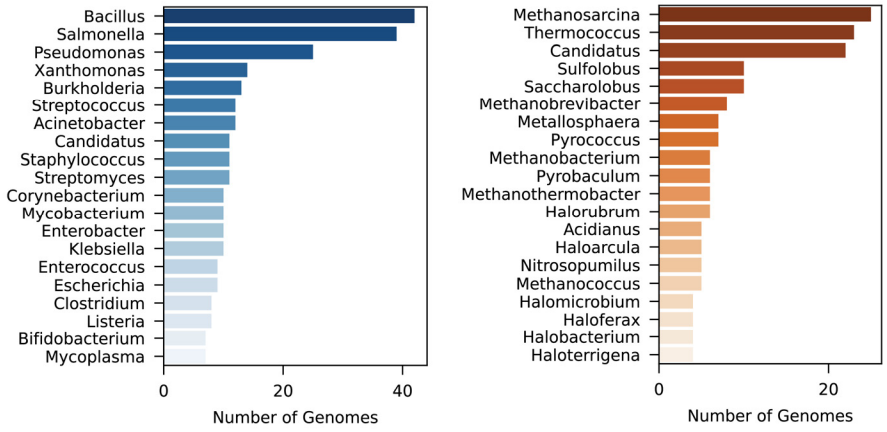
415 **(https://imodulondb.org/)**

Species	**Reference**	**Reference genome**	**Gene#**	**Condition#**
Acinetobacter baumannii	Menon, et al.[57]	CP008706.1	1788	139
Bacillus subtilis	Sastry, et al.	NC_000964.3	1969	265
Escherichia coli	Lamoureux, et al.[58]	U00096.3	2490	1035
Limosilactobacillus reuteri	Josephs-Spaulding, et al.[59]	GCA_020785475.1	1159	117
Mycobacterium tuberculosis	Yoo, et al.[60]	AL123456.3	3214	647
Pseudomonas aeruginosa	Rajput, et al.[61]	NC_002516.2	3981	411
Pseudomonas putida	Lim, et al.[1]	AE015451.2	4379	321
Pseudomonas syringae	Bajpe, et al.[62]	AE016853.1	4296	202
Staphylococcus aureus	Poudel, et al.[63]	NC_010079.1	1736	385
Synechococcus elongatus	Yuan, et al.[64]	NC_007604.1	1522	300
Salmonella enterica	Yuan, et al.[64]	AE006468.2	1808	533
Streptococcus pyogenes	Hirose, et al.[65]	CP008776.1	1173	116
Vibrio natriegens	Shin, et al.[4]	GCA_001456255.1	1981	148
Summary			31,496	4,619

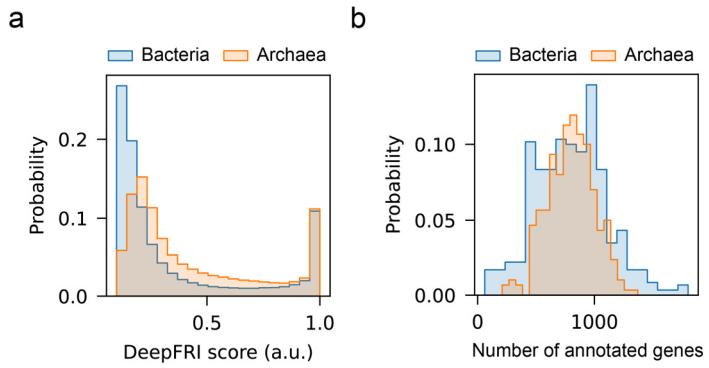
Supplementary Figures



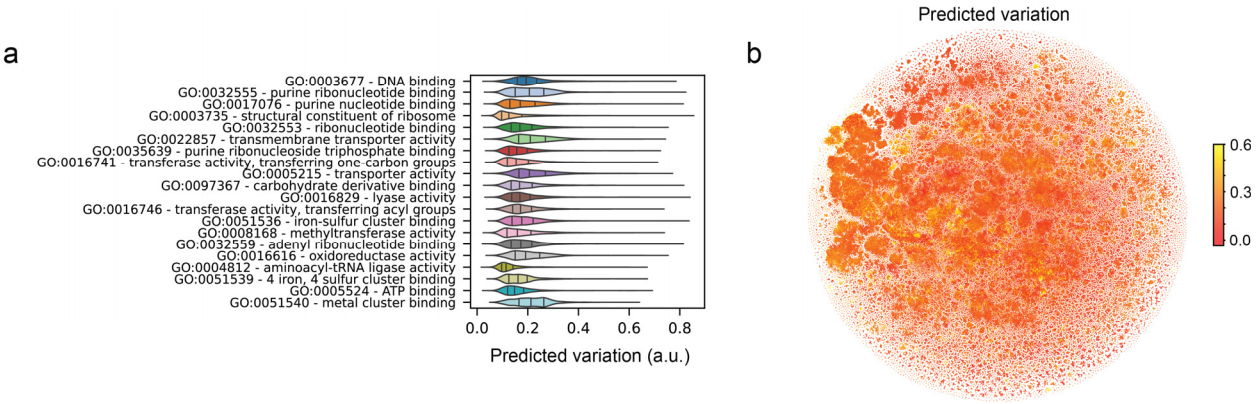
<u>Supplementary Figure 1:</u> Prediction of gene expression variation. a) Spearman correlations between predicted and measured gene expression variations in leave-one-genome-out experiments for bacterial species. Results from five rounds of model training and evaluation are shown (n = 5). **b)** Spearman correlations between predicted and measured gene expression variations in leave-one-genome-out experiments for archaeal species. **c)** Number of conditions and mean Spearman correlations for bacterial species (n = 22).



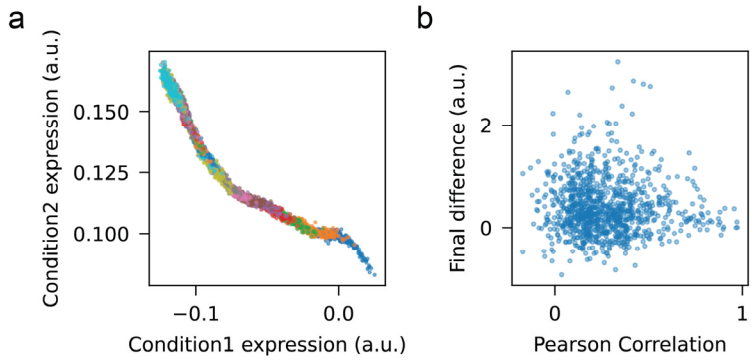
<u>Supplementary Figure 2:</u> Most frequent genera in the new genomes. The number of genomes for the most frequence genera is shown for bacteria (left) and archaea (right).



<u>Supplementary Figure 3:</u> DeepFRI results for the new genomes. a) DeepFRI scores for genes in the new bacterial and archaeal genomes (n = 2.34M and 0.77M). **b)** Number of genes with DeepFRI scores greater than 0.5 in the new genomes.



<u>Supplementary Figure 4:</u> Prediction of gene expression variation for the new genomes. a) Predicted gene expression variation for genes associated with the top 20 GO terms in the new genomes. **b)** Predicted gene expression variation visualized in t-SNE plots of protein embeddings (n = 3.11M).



435

436 **<u>Supplementary Figure 5:</u> Optimization of condition-specific promoters. a)** Predicted gene expression in
437 two conditions across different optimization rounds. Different colors represent sequences from
438 different optimization rounds (n = 200 for each round). **b)** The Pearson correlation between gene
439 expressions across two conditions and the difference in predicted expression for the optimized
440 sequences. Each dot represents a random condition combination for *E. coli* (n = 1,000).