Predicting microbial transcriptome using genome sequence

Guohao Fu[1, 5], Yujing Yan[2, 5], Ye Chen[3 *], Bin Shao[4, 5 *]

[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] Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China

[4] Advanced Research Institute of Multidisciplinary Sciences, Beijing Institute of Technology, Beijing 100081, P. R. China

[5] These authors contribute equally to this work.

*Correspondence should be addressed to Ye Chen (chenye.h.ai@gmail.com), 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 and 0.62 in predicting gene expression for new bacterial and fungal genomes. We further extend this framework to predict transcriptomes for 2,685 additional microbial genomes spanning 1,744 genera, 82% of which remain uncharacterized at the transcriptional level. Our analysis highlights conserved and divergent transcriptional programs across understudied genera, providing a powerful resource for uncovering microbial adaptation strategies and metabolic potential across the tree of life.

For a genome-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 has 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 large-scale, cross-species studies of microbial gene expression across the tree of life.

Current machine learning approaches for predicting gene expression have primarily focused on analyzing untranslated regions (UTRs)[11–16] or large genomic windows surrounding the gene of interest [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 it leverages a protein language model, ESM2[24], to extract predictive features. Using transcriptome data from 22 bacterial, 9 archaeal and 9 fungal species obtained from the NCBI GEO database[25], we trained a transformer-based model that achieves a Spearman correlation up to 0.70 in leave-one-genome-out evaluations. We then validated the model's performance in the non-model organism *Pseudomonas sp.* WBC-3 by conducting a series of RNA-seq experiments. Additionally, we trained the model to account for condition-dependent gene expression using data from 13 bacterial species across 4.6k unique conditions and demonstrated that it can identify sequence determinants underlying these variations. Lastly, we applied the trained TXpredict models to predict transcriptomes for 2,685 additional genomes, the majority of them have never been transcriptionally characterized. This resulted in the creation of TXpredictDB, a database containing expression profiles for 9.8 million genes. The models developed in this study are publicly available: https://github.com/lingxusb/TXpredict

54 We first collected microbial RNA-seq datasets from the NCBI GEO datasets and selected the dataset with
55 the highest number of samples for each microbial species. Read counts were normalized with the same
56 pipeline (Methods). Taken together, our dataset includes 22 bacterial, 9 archaeal and 9 fungal species
57 and a total of 24.8M gene expression measurements (Supplementary Table 1). For each microbial
58 species, gene expression was averaged across all samples to construct a standard expression profile.
59 Protein embeddings for each gene were computed using the ESM2-650M model and combined with
60 basic protein statistics to serve as the input features for TXpredict (Fig. 1a, Methods). For model
61 evaluation, we employed a leave-one-genome-out cross-validation strategy. For bacterial species, the
62 model was trained on genes from 21 species and tested on the held-out species (Fig. 1b). We followed
63 the same approach for archaeal and fungal species. Model performance was evaluated using Spearman
64 correlation coefficients between predicted and measured expression values for the held-out genome.

65 We found our model achieved a mean Spearman correlation coefficient of 0.53 for the bacterial species
66 (Fig. 1c), and performed well for non-model organisms, with correlation coefficients of 0.64, 0.62 and
67 0.62 for *B. hinzii, B. thetaiotaomicron*, and *C. beijerinckii*. It is worth noting that transcriptome
68 predictions for these bacteria were based solely on the computational analysis of annotated genome
69 sequences, providing an alternative to labor-intensive RNA-seq experiments. The model's performance
70 in predicting gene expression variation decreased modestly as the variation calculations depend on the
71 diversity of experimental conditions (Supplementary Fig. 1). Despite having a limited training set for
72 fungal species, our TXpredict model obtained a mean correlation of 0.62 and reached a correlation
73 coefficient up to 0.70 for non-model organisms such as *O. oligospora* (Fig. 1c). For archaea, TXpredict
74 showed a mean Spearman correlation of 0.44, partly due to the limited availability of training data (Fig.
75 1b). The model demonstrated broad applicability to extremophiles from distinct natural niches (ranging
76 from Dead Sea to volcanic springs), further highlighting the robustness of our method.

77 In our initial evaluation, TXpredict performed poorly on the archaeon *N. maritimus*, with a Spearman
78 correlation of 0.23 in the leave-one-genome-out evaluation (Supplementary Fig. 2). Upon closer
79 examination of this dataset, we found that fewer than 40% of the reads were successfully mapped to
80 reference genome, demonstrating the potential of our model to detect potential problems in RNA-seq
81 data. In addition, we examined whether protein sequence similarity affects model performance. For
82 bacterium *B. hinzii*, we used BLASTP[26] to compute sequence similarity between its proteins and those in
83 the other bacterial strains. The model was trained on other species and tested on rare proteins from *B.*
84 *hinzii* that were not present in the training data, showing a correlation of 0.46 for these rare proteins

85 (Supplementary Fig. 3). While this is lower than the correlation for all proteins, it suggests that the
86 model captures expression-related features beyond sequence similarity.

87 We next explored TXpredict's cross-taxa performance by training separate models on bacterial,
88 archaeal, or fungal species and then testing them on other domains (Fig. 1d). Interestingly, we still found
89 our model showed moderate performance on cross-taxa predictions, suggesting that they captured
90 general principles of gene expression regulation. The archaeal model achieved a higher mean
91 correlation for eukaryotic species than for archaea, indicating that the gene expression is more
92 constrained in eukaryotes. At the species level, models trained on other taxa generally performed worse
93 than those trained on the same taxon (Fig. 1e and Supplementary Fig. 4).

94 To validate the prediction of the TXpredict model, we conducted new RNA-seq experiments on
95 *Pseudomonas sp.* WBC-3 and compared the predicted expression levels with the measured values. The
96 Spearman correlation between measured and predicted expression was 0.40 for all proteins and 0.50
97 for proteins with GO term annotations (Fig. 1f). Notably, our model showed high correlations (0.55) for
98 certain GO categories including catalytic activity, and also for operons like the flagellar genes (Fig. 1g and
99 Supplementary Fig. 5).

100 We next extended our framework to model condition-dependent gene expression (Fig. 1h). We
101 collected RNA-seq profiles from 13 bacterial strains across 4.6k unique conditions from iModulonDB,
102 resulting in a total of 12.0M measurements (Supplementary Table 2, Fig. 1i). The model integrated one-
103 hot encoded 5'UTR sequences with protein embeddings as input features. For each gene, we calculated
104 the correlation between measured and predicted expression values across all conditions. Our TXpredict
105 model achieved an average correlation coefficient of 0.52 across all bacterial strains (Fig. 1i). For *E. coli*,
106 the distributions of Pearson's and Spearman's correlation coefficients show similar patterns, with 41%
107 and 38% of genes having correlation coefficients above 0.5 (Fig. 1j).

108 To further interpret the model's predictions, we calculated input gradients[17] to identify sequence motifs
109 contributing to gene expression variation across conditions (Supplementary Fig. 6). In *E. coli*, we
110 observed a peak in the input × gradient overlap with the -10 box of the sigmaS-dependent promoter of
111 the *yadV* gene[27]. Also input gradients increased after the transcriptional start site, suggesting potential
112 contributions from transcribed sequences (Supplementary Fig. 5). In a non-model organism *P. putida*,
113 we identified a peak overlapping with a Zur binding motif for the *PP_0120* gene, consistent with its role

114 as a Zur iModulon gene[1]. These findings illustrate our model's ability to capture condition-specific gene
115 expression changes in an interpretable manner.

116 Lastly, we applied the trained TXpredict model to predict transcriptomes for new genomes including
117 2,169 bacterial, 423 archaeal and 93 fungal species (Fig. 2a and b). Collectively, these new genomes span
118 1,744 genera, of which 82% have no prior RNA-seq data according to publicly available data in the NCBI
119 GEO database. To explore the relationship between predicted gene expression and protein function, we
120 used DeepFRI[28] to predict Gene Ontology (GO) terms for the new genes (Supplementary Fig. 7). Among
121 the GO terms with the highest predicted expression levels were structural constituents of ribosome
122 (GO:0003735) (Fig. 2e, Supplementary Fig. 7). In contrast, GO terms with the lowest predicted
123 expression included proteins with DNA-binding activity (GO:0003677), consistent with their function as
124 transcriptional regulators. Collectively, our model predicted the expression of up to 9.8M genes across
125 these genomes, which were compiled into a database named TXpredictDB. The predicted gene
126 expressions in different domains followed similar distributions (Fig. 2f). When visualizing protein
127 embeddings of all proteins, we observed a clear separation of bacterial, archaeal and fungal genes (Fig.
128 2g, Supplementary Fig. 8). Small clusters of ribosomal genes located at the periphery of the plot and
129 showed high predicted expression, while transmembrane transporter proteins formed distinct low-
130 expression clusters.

131 We focused on genera without prior RNA-seq data and grouped genes based on their annotated Gene
132 Ontology (GO) terms (Supplementary Fig. 9). There was a positive correlation between the predicted
133 expression levels of tRNA ligases and ribosomal proteins across genera (Fig. 2h, Spearman correlation
134 0.52). *Actinoplanes* had the lowest predicted expression for ribosomal proteins, in accord with their
135 known slow growth rate[29]. The genus *Candidatus* includes a symbiotic bacterium found in carpenter ants
136 (*Blochmanniella*) and showed low expression of both tRNA ligases and ribosomal proteins. This pattern
137 likely reflects its reduced need for protein turnover and genome stasis[30], in contrast to free-living
138 bacteria. We further visualized the embeddings of tRNA ligases to explore their sequence variations (Fig.
139 2i). tRNA ligases from *Candidatus* formed distinct low-expression clusters, whereas those from
140 *Acidiferrobacter* were distributed across multiple clusters despite similarly low expression levels,
141 suggesting divergent evolutionary constraints associated with their ecological niche.

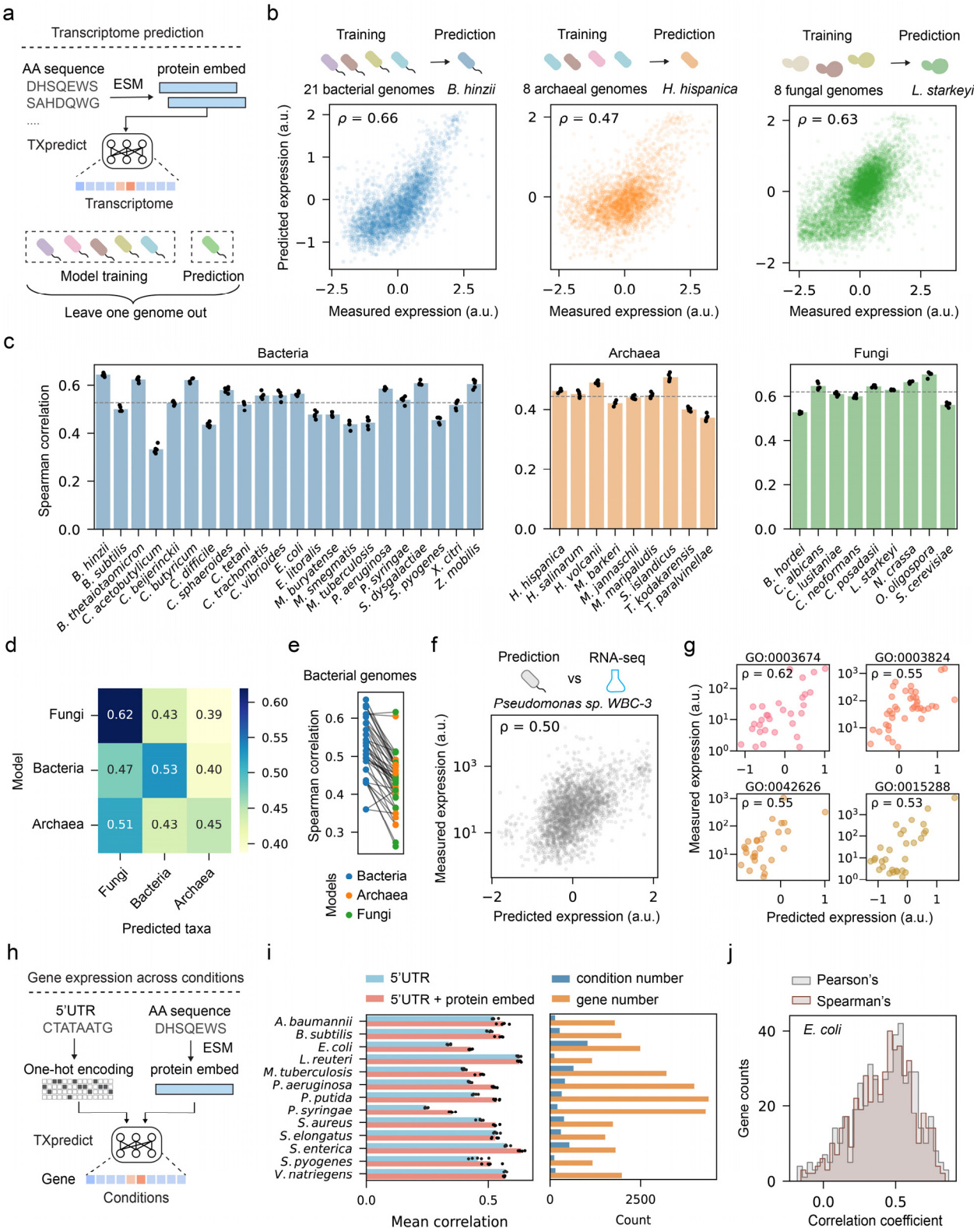
142 We also clustered protein sequences using MMseq2[31] and predicted gene expression within each high-
143 similarity cluster using the TXpredict model. Interestingly, several clusters displayed non-unimodal
144 expression distributions, with conserved amino acid differences separating low- and high-expression

145 variants (Fig. 2j). For example, in cluster of the PTS IIA component homologs, low-expression variants
146 shared a V116K substitution (Supplementary Fig. 10), which could reduce translation efficiency and
147 mRNA stability due to the presence of constitutive lysine residues[32,33].

148 We further analyzed the predicted gene expression profiles of biosynthetic gene clusters (BGCs) across
149 TXpredictDB, which were identified using antiSMASH[34] (n = 11,400). Although BGCs are typically lowly
150 expressed, distinct patterns were observed across different cluster types. For example, the proteusin
151 BGC showed high expression of core biosynthetic genes, whereas the ectoine BGC exhibited elevated
152 expression of regulatory genes (Supplementary Fig. 11). These reference expression profiles in non-
153 model organisms may offer valuable insights into the functional organization of BGCs.

154 In summary, we present TXpredict, a transcriptome prediction model that can generalize to a wide
155 range of microbes. For a microbial genome with 4.6k genes, transcriptome prediction can be completed
156 in approximately 22 minutes using a Google Colab notebook (Methods), demonstrating the model's
157 efficiency and accessibility. Our approach is not without limitations: while it effectively leverages protein
158 embeddings for expression prediction, it does not account for other regulatory elements or non-coding
159 RNAs. Moreover, our model's performance is constrained by the limited training profiles, especially for
160 archaeal species. Lastly, we provide a single reference transcriptome per genome in TXpredictDB, which
161 does not capture gene expression dynamics. Despite these limitations, TXpredict has the potential to
162 complement experimental approaches and advance the study of the vast number of sequenced yet
163 poorly characterized microbes. For example, identifying conserved expressed genes in pathogens may
164 enable the discovery of novel drug targets[35]. Furthermore, predicting gene expression across genomes
165 can provide insight into how genome plasticity contributes to bacterial pathogenicity[36]. Extending
166 transcriptome prediction to unculturable microbes also provides critical insights into microbial
167 adaptation to extreme environments and the evolution of gene regulation across prokaryotes[37,38]. These
168 efforts could further benefit from integrating single-cell RNA-seq data [5,39,40] and computational tools for
169 mRNA synthesis and stability prediction[41–43], paving the way for a deeper understanding of microbial
170 transcriptional landscapes.

Figures


a
Transcriptome prediction
AA sequence
DHSQEWS
SAHDQWG
ESM
protein embed
TXpredict
Transcriptome
Model training
Prediction
Leave one genome out
b
Training
Prediction
21 bacterial genomes
B. hinzii
8 archaeal genomes
H. hispanica
8 fungal genomes
L. starkeyi
ρ = 0.66
ρ = 0.47
ρ = 0.63
Predicted expression (a.u.)
Measured expression (a.u.)
c
Bacteria
Archaea
Fungi
Spearman correlation
B. hinzii
B. subtilis
B. thetaiotaomicron
C. acetobutylicum
C. beijerinckii
C. butyricum
C. difficile
C. sphaeroides
C. tetani
C. trachomatis
C. vibrioides
E. coli
E. litoralis
M. buryatense
M. smegmatis
M. tuberculosis
P. aeruginosa
P. syringae
S. dysgalactiae
S. pyogenes
X. citri
Z. mobilis
H. hispanica
H. salinarum
H. volcanii
M. barkeri
M. jannaschii
M. maripaludis
S. islandicus
T. kodakarensis
T. paralvinellae
B. hordei
C. albicans
C. lusitaniae
C. neoformans
C. posadasii
L. starkeyi
N. crassa
O. oligospora
S. cerevisiae
d
Model
Fungi
Bacteria
Archaea
0.62
0.43
0.39
0.47
0.53
0.40
0.51
0.43
0.45
Predicted taxa
e
Bacterial genomes
Spearman correlation
Models
Bacteria
Archaea
Fungi
f
Prediction
RNA-seq
vs
Pseudomonas sp. WBC-3
ρ = 0.50
Measured expression (a.u.)
Predicted expression (a.u.)
g
GO:0003674
GO:0003824
GO:0042626
GO:0015288
ρ = 0.62
ρ = 0.55
ρ = 0.55
ρ = 0.53
Measured expression (a.u.)
Predicted expression (a.u.)
h
Gene expression across conditions
5'UTR
CTATAATG
AA sequence
DHSQEWS
ESM
One-hot encoding
protein embed
TXpredict
Gene
Conditions
i
5'UTR
5'UTR + protein embed
condition number
gene number
A. baumannii
B. subtilis
E. coli
L. reuteri
M. tuberculosis
P. aeruginosa
P. putida
P. syringae
S. aureus
S. elongatus
S. enterica
S. pyogenes
V. natriegens
Mean correlation
Count
j
Pearson's
Spearman's
E. coli
Gene counts
Correlation coefficient

173 **Figure 1. Prediction of transcriptome based on the genome sequence**

174 **a)** Schematic of TXpredict for transcriptome prediction. **b)** Scatter plots showing predicted and
175 measured gene expression in leave-one-genome-out evaluations for one species from bacteria, archaea,
176 and fungi. Spearman correlation coefficients are reported in the top-left corner of each plot. **c)**
177 Spearman correlation coefficients between predicted and measured gene expression for all species.
178 TXpredict models were separately trained for bacterial, archaeal, and fungal species. Results from five
179 rounds of model training and testing are reported (n = 5). Dashed lines indicate the mean correlations
180 for each domain. **d)** Cross-domain prediction performance of TXpredict models. Models were trained on
181 all species from one domain and tested on species from the other domains. Mean Spearman
182 correlations across all test species are shown. Diagonal values represent leave-one-genome-out
183 performance within each domain. **e)** Predictions of archaeal and fungal models on all bacterial species.
184 Blue dots indicate leave-one-genome-out evaluations for bacteria. Predictions for the same species
185 were connected. **f)** Model predictions versus measured gene expression in new RNA-seq experiment for
186 *Pseudomonas sp.* WBC-3. Each dot represents one functional gene and the Spearman correlation is
187 shown. **g)** Predicted and measured gene expression for genes with different GO annotations in
188 *Pseudomonas sp.* WBC-3. **h)** Schematic of TXpredict for condition-specific gene expression prediction. **i)**
189 Mean correlations for models trained separately on different bacterial species. Correlations from 5-fold
190 cross-validation tests are shown on the left (n = 5). Gene and condition counts for each species are
191 shown on the right. **j)** Distribution of Spearman and Pearson correlation coefficients for gene expression
192 across conditions in *E. coli* (n = 498).

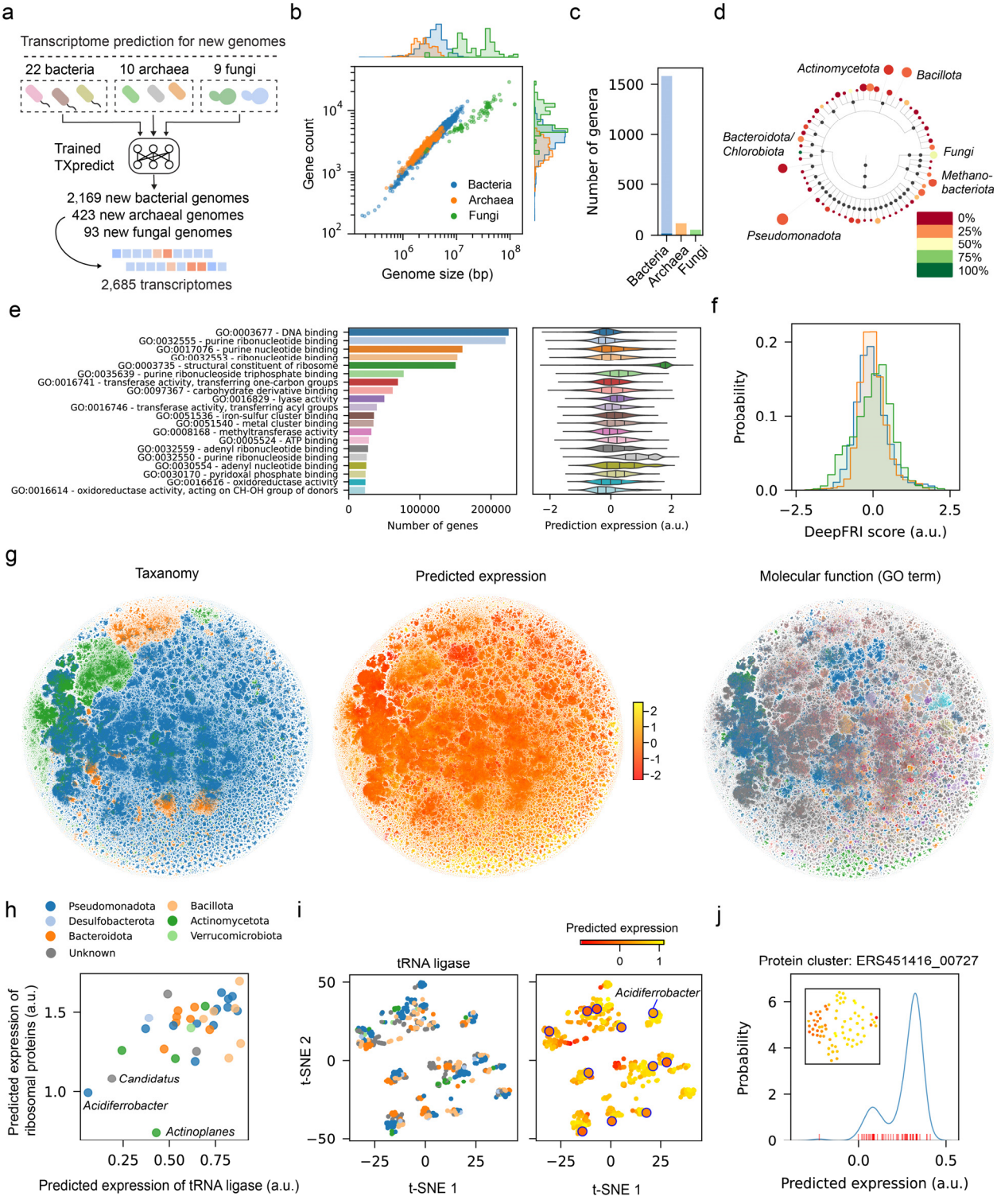


Figure 2. Application of TXpredict to new genomes

a) Transcriptome prediction for new genomes. **b)** Genome sizes and gene counts for the new genomes (n = 2,685). **c)** Number of genera in the new genomes (light colors) compared to training genomes (dark

197 colors). **d)** Polygenetic tree of newly added genomes. Each dot represents a phylum. The dot size is
198 proportional to the number of genera it contains and color indicates the proportion of genera with
199 existing RNA-seq data based on the NCBI GEO database. **e)** Top 20 GO terms of the genes in the new
200 genomes calculated by the DeepFRI model[28] (left) and their predicted gene expression levels (right). **f)**
201 Distributions of predicted gene expression levels across new bacterial (blue), archaeal (orange) and
202 fungal (green) genomes. **g)** t-SNE plots of protein embeddings generated by the ESM2 model[24] (n =
203 9.8M). Each dot represents a gene in the new genomes, colored by taxa (left), predicted gene expression
204 (middle), and GO terms (right). GO terms are colored as in panel (e). **h)** Predicted expression of tRNA
205 ligases and ribosomal proteins across understudied genera (n = 40), colored by phylum. **i)** t-SNE plots of
206 tRNA ligase embeddings across understudied genera, colored by phylum (left) and predicted expression
207 level (right). **j)** Histogram of predicted expression levels for proteins in a high-similarity cluster. Inset
208 shows the t-SNE plot of embeddings for these proteins, colored by predicted expression.

209 **Methods**

210 **Training dataset**

211 We collected RNA-seq data from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). For each
212 microbe species, we selected the dataset with the highest number of samples and used either raw count
213 data or RPKM (Reads Per Kilobase of transcript per Million mapped reads) values when available. For
214 raw count data, we calculated the RPKM for each gene, followed by log (1 + p) and z-score
215 normalization. For RPKM values, we applied log (1 + p) transformation and z-score normalization. In
216 total, we gathered data from 22 bacterial species (2,649 samples), 9 archaeal species (169 samples) and
217 9 fungal species (2,078 samples, Supplementary Table 1). For each species, we constructed a standard
218 gene expression profile by averaging the normalized expression values across all available samples.

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

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

231

232 **Transcriptome prediction model**

233 We developed a Transformer-based model to predict gene expression levels from sequence features.
234 The transformer encoder utilizes a multi-head self-attention mechanism (num_heads=4) with a single
235 encoder layer. The encoder processes input features through a linear transformation into a hidden
236 dimension of 128, followed by positional encoding. We implemented sinusoidal positional encoding to
237 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 bacterial, archaeal and fungal genomes were downloaded from NCBI Genbank (https://ftp.ncbi.nlm.nih.gov/genomes/genbank/). To maximize genus-level representation for transcriptome prediction, we selected 2,169 bacterial, 423 archaeal, and 93 fungal genomes spanning a total of 1,744 genera. Gene expression for these genomes was predicted using TXpredict models, forming the basis of the TXpredictDB. 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])*". We annotated taxonomy of these genomes by querying the GTDB database[44]. 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. Pearson correlation coefficients were calculated between predicted and measured expression levels across all conditions for genes in the test set.

Analysis of protein sequences

To visualize the high-dimensional embeddings in a two-dimensional space, we implemented t-SNE (t-Distributed Stochastic Neighbor Embedding) using the *opentsne* package. We initialized the embedding using PCA and constructed a uniform affinity matrix with 10 nearest neighbors. The optimization process consisted of three phases: an early exaggeration phase of 125 iterations with an exaggeration factor of 12 and momentum of 0.5, followed by a gradual exaggeration annealing phase where the exaggeration factor was linearly decreased from 12 to 1 over 125 iterations with momentum of 0.8. The final optimization phase ran for 2000 iterations without exaggeration. Function prediction for these proteins was carried out using the default deepFRI model[28], as available on GitHub (https://github.com/flatironinstitute/DeepFRI). We used a score cutoff of 0.5 which was reported to be significant in the original publication. To cluster the protein sequences in TXpredictDB, we used the MMseqs2[31] package with default settings and a minimum sequence identity threshold of 0.9.

Calculation of input gradient

To investigate the contribution of individual nucleotides in the UTR sequence to gene expression prediction across conditions, we implemented a gradient-based analysis. Using a sequence-only model variant where the protein embedding branch was masked, we calculated the gradient of the model output with respect to the input DNA sequence. For each target gene and condition, we performed a forward pass with the one-hot encoded UTR sequence while setting all other inputs to zero. We then computed the gradients through backpropagation and multiplied them element-wise with the input sequence (Gradient × Input method)[17]. The absolute values of these products were summed across the

295 four nucleotide channels (A, T, C, G) to obtain single-base resolution importance scores. For visualization
296 purposes, we aggregated these scores into 5-bp bins by summing the scores within each bin.

297

298 **RNA-seq experiment of the *Pseudomonas* sp. WBC-3 strain**

299 For the RNA-seq experiment, *Pseudomonas* sp. WBC-3 was cultured under three different conditions to:
300 PNP1 (minimal medium with 0.01% glycerol), PNP2 (minimal medium with 0.04% PNP), and PNP3
301 (minimal medium with 0.01% glycerol and 0.04% PNP). Each condition was prepared by diluting a 2×
302 stock of yeast extract minimal medium supplemented with either 0.02% glycerol, 0.08% PNP, or both,
303 resulting in the desired final concentrations. WBC-3 strain was first grown overnight in LB medium, then
304 diluted 1:100 into 4 mL of each experimental medium and incubated overnight at 30°C with shaking at
305 200 rpm. Cultures were harvested at OD600 between 0.5 and 1.0 by centrifugation for 2 minutes,
306 ensuring that the total collected biomass did not exceed 1×10^9 cells. Total RNA was extracted using the
307 RNAprep Pure Bacteria Kit (TIANGEN, DP430), followed by DNase I treatment (TIANGEN, RT411) to
308 remove genomic DNA contamination. RNA quality and concentration were confirmed to meet
309 sequencing requirements. Raw sequencing reads were aligned and processed using standard RNA-seq
310 pipelines. Read counts were quantified using StringTie with the following command: stringtie -p 8 -G
311 GCF_035758525.1_ASM3575852v1_genomic.gtf -l sample_id -o sample.gtf sample.bam, using the
312 *Pseudomonas* sp. WBC-3 reference genome (GCF_035758525.1).

313

314 **Code availability**

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

317 **References**

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

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

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

324 4. Shin, J., Rychel, K. & Palsson, B. O. Systems biology of competency in Vibrio
325 natriegens is revealed by applying novel data analytics to the transcriptome. *Cell Rep* **42**,
326 (2023).

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

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

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

333 8. Barnum, T. P. *et al.* Predicting microbial growth conditions from amino acid composition. *bioRxiv*
334 2023–2024 (2024).

335 9. Ramoneda, J. *et al.* Building a genome-based understanding of bacterial pH preferences. *Sci Adv*
336 **9**, eadf8998 (2024).

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

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

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

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

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

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

350 16. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature*
351 (2022) doi:10.1038/s41586-022-04506-6.

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

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

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

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

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

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

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

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

368 25. Clough, E. & Barrett, T. The gene expression omnibus database. *Statistical Genomics: Methods*
369 *and Protocols* 93–110 (2016).

370 26. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

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

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

376 29. Couch, J. N. Actinoplanes, a new genus of the Actinomycetales. *Journal of the Elisha Mitchell*
377 *Scientific Society* **66**, 87–92 (1950).

378 30. Degnan, P. H., Lazarus, A. B. & Wernegreen, J. J. Genome sequence of Blochmannia
379 pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects.
380 *Genome Res* **15**, 1023–1033 (2005).

381 31. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
382 analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

383 32. Radhakrishnan, A. & Green, R. Connections underlying translation and mRNA stability. *J Mol Biol*
384 **428**, 3558–3564 (2016).

385 33. Meydan, S. & Guydosh, N. R. Disome and Trisome Profiling Reveal Genome-wide Targets of
386 Ribosome Quality Control. *Mol Cell* **79**, 588-602.e6 (2020).

387 34. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical
388 structures and visualisation. *Nucleic Acids Res* **51**, W46–W50 (2023).

389 35. Wang, C., Hong, Z., Rongfen, Z., Clement Kin-Ming, T. & and Deng, S. Genomic Characterization
390 and Antimicrobial Resistance of ESBL-Producing, Escherichia coli Isolates in Suzhou, China. *Infect*
391 *Drug Resist* **18**, 1049–1057 (2025).

392 36. Krishnakant Kushwaha, S. *et al.* Comprehensive blueprint of Salmonella genomic plasticity
393 identifies hotspots for pathogenicity genes. *PLoS Biol* **22**, e3002746- (2024).

394 37. Shu, W.-S. & Huang, L.-N. Microbial diversity in extreme environments. *Nat Rev Microbiol* **20**,
395 219–235 (2022).

396 38. Caro-Astorga, J. *et al.* Polyextremophile engineering: a review of organisms that push the limits of
397 life. *Front Microbiol* **15**, (2024).

398 39. Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.-E. & Vogel, J. Single-cell RNA-sequencing
399 reports growth-condition-specific global transcriptomes of individual bacteria. *Nat Microbiol* **5**,
400 1202–1206 (2020).

401 40. Kuchina, A. *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science (1979)*
402 **371**, eaba5257 (2021).

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

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

408 43. Xia, Y. *et al.* Language model generates cis-regulatory elements across prokaryotes. *bioRxiv*
409 2011–2024 (2024).

410 44. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a
411 phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic*
412 *Acids Res* **50**, D785–D794 (2022).

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

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

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

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

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

424 50. D, S. M. *et al.* Arabinose-Induced Catabolite Repression as a Mechanism for Pentose Hierarchy
425 Control in Clostridium acetobutylicum ATCC 824. *mSystems* **3**, 10.1128/msystems.00064-18
426 (2018).

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

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

432 53. Espah Borujeni, A., Zhang, J., Doosthosseini, H., Nielsen, A. A. K. & Voigt, C. A. Genetic circuit
433 characterization by inferring RNA polymerase movement and ribosome usage. *Nat Commun* **11**,
434 5001 (2020).

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

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

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

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

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

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

448 60. Kachroo, P. *et al.* Integrated analysis of population genomics, transcriptomics and virulence
449 provides novel insights into Streptococcus pyogenes pathogenesis. *Nat Genet* **51**, 548–559
450 (2019).

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

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

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

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

459 65. D, T. B., Cem, M., B, N. T., Q, L. S. & F, H. J. Growth Kinetics, Carbon Isotope Fractionation, and
460 Gene Expression in the Hyperthermophile Methanocaldococcus jannaschii during Hydrogen-
461 Limited Growth and Interspecies Hydrogen Transfer. *Appl Environ Microbiol* **85**, e00180-19
462 (2019).

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

465 67. Spietz, R. L. *et al.* Investigating Abiotic and Biotic Mechanisms of Pyrite Reduction. *Front*
466 *Microbiol* **13**, (2022).

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

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

472 70. G, D. E., E, S. J., Alix, A., Patricia, O. & A, H. D. Balancing Positive and Negative Selection: In Vivo
473 Evolution of Candida lusitaniae MRR1. *mBio* **12**, 10.1128/mbio.03328-20 (2021).

474 71. Kang, Y. S. *et al.* Leveraging a new data resource to define the response of Cryptococcus
475 neoformans to environmental signals. *Genetics* **229**, iyae178 (2025).

476 72. Schnepf, V., Vlot, A. C., Kugler, K. & Hückelhoven, R. Barley susceptibility factor RACB modulates
477 transcript levels of signalling protein genes in compatible interaction with Blumeria graminis f.sp.
478 hordei. *Mol Plant Pathol* **19**, 393–404 (2018).

479 73. Hunsaker, E. W., Yu, C.-H. A. & Franz, K. J. Copper Availability Influences the Transcriptomic
480 Response of Candida albicans to Fluconazole Stress. *G3 Genes|Genomes|Genetics* **11**, jkab065
481 (2021).

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

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

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

489 77. Reo, Y. *et al.* Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a
490 Structured Interplay between Metabolism, Stress Response, and Infection. *mSphere* **7**, e00033-22
491 (2022).

492 78. Rajput, A. *et al.* Advanced transcriptomic analysis reveals the role of efflux pumps and media
493 composition in antibiotic responses of Pseudomonas aeruginosa. *Nucleic Acids Res* **50**, 9675–
494 9688 (2022).

495 79. Heera, B., Kevin, R., R, L. C., V, S. A. & O, P. B. Machine learning uncovers the Pseudomonas
496 syringae transcriptome in microbial communities and during infection. *mSystems* **8**, e00437-23
497 (2023).

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

500 81. Yuan, Y. *et al.* Machine learning reveals the transcriptional regulatory network and circadian
501 dynamics of Synechococcus elongatus PCC 7942. *Proceedings of the National Academy of*
502 *Sciences* **121**, e2410492121 (2024).

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

505

506 **Supplementary Table**

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

Bacteria				
Species	**GSE number**	**Reference genome**	**Gene#**	**Condition#**
Bacillus subtilis	GSE224332[45]	NZ_CP020102.1	4220	317
Bacteroides thetaiotaomicron	GSE251676[46]	GCF_000011065.1	4618	64
Bordetella hinzii	GSE181542	CP052845.1	4531	228
Caulobacter vibrioides	GSE246782[47]	NC_011916.1	3882	34
Cereibacter sphaeroides	GSE186600[48]	GCA_000012905.2	4232	17
Chlamydia trachomatis	GSE262652	GCA_000068585.1	857	30
Clostridioides difficile	GSE249810[49]	GCA_000009205.5	3560	24
Clostridium acetobutylicum	GSE107804[50]	GCA_000008765.1	3755	48
Clostridium beijerinckii	GSE150480[51]	GCF_000833105.2	5316	16
Clostridium butyricum	GSE150480[51]	GCF_001456065.2	3848	16
Clostridium tetani	GSE182408[52]	NC_004557.1	2562	39
Escherichia coli	GSE152664[53]	NC_010473.1	3260	9
Erythrobacter litoralis	GSE126532	GCA_001719165.1	3003	49
Methylotuvimicrobium buryatense	GSE162089[54]	NZ_CP035467.1	3927	56
Mycobacterium tuberculosis	GSE110508[55,56]	GCA_000195955.2	3957	87
Mycolicibacterium smegmatis	GSE69983[57]	GCA_000015005.1	6700	30
Pseudomonas aeruginosa	GSE251671	GCA_000014625.1	5659	858
Pseudomonas syringae	GSE103441[58]	GCA_000007805.1	5586	114
Streptococcus dysgalactiae	GSE272047[59]	GCA_029234115.1	1939	114
Streptococcus pyogenes	GSE113055[60]	CP000056.2	1608	442
Xanthomonas citri	GSE252662[61]	AE008923.1	4294	38
Zymomonas mobilis	GSE139939[62]	CP023715.1	1759	19
Summary			83,073	2,649

508 **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[63]	GCF_000006805.1	2325	24
Haloferax volcanii	GSE248564[64]	GCF_000025685.1	3873	39
Methanocaldococcus jannaschii	GSE112986[65]	GCF_000091665.1	1724	13
Methanococcus maripaludis	GSE261517[66]	GCF_000011585.1	1730	12
Methanosarcina barkeri	GSE168895[67]	GCF_000970025.1	3338	12
Sulfolobus islandicus	GSE220819[68]	GCF_000189555.1	2682	18
Thermococcus kodakarensis	GSE186817	NC_006624.1	2214	4
Thermococcus paralvinellae	GSE101078[69]	GCF_000517445.1	1995	15
Summary			25,448	169

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

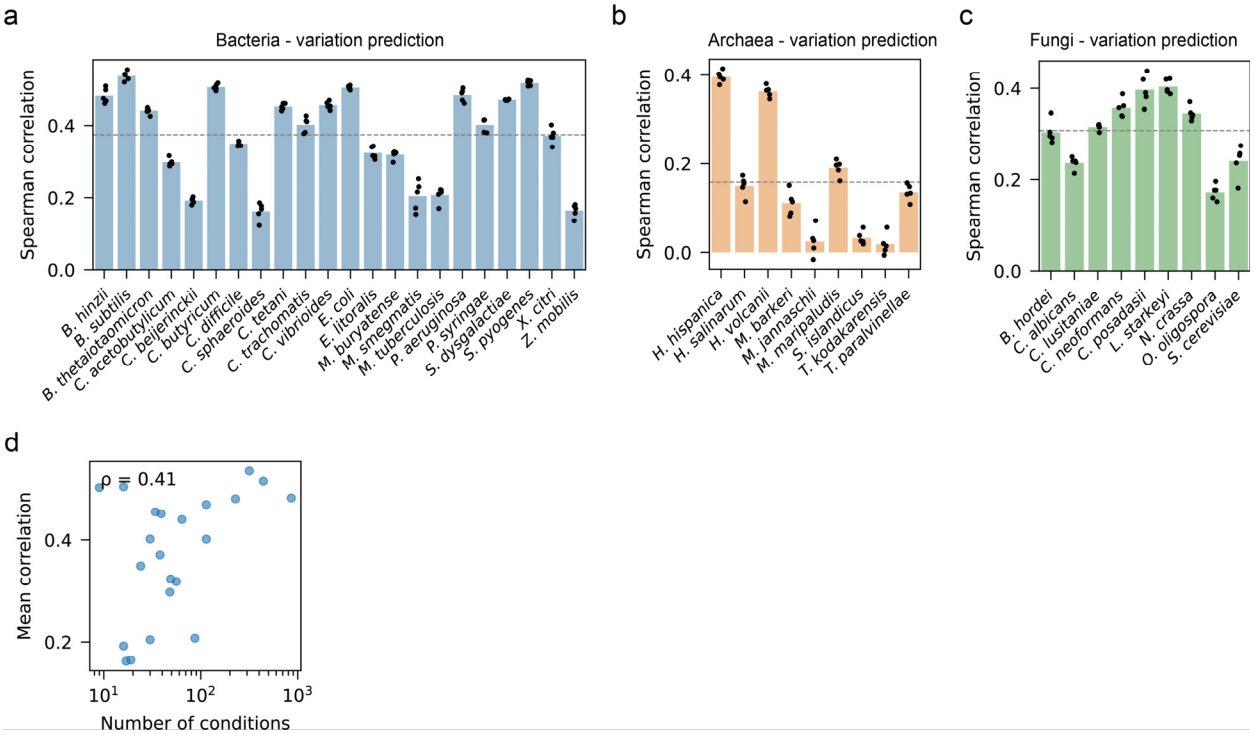
Fungi				
Species	**GSE number**	**Reference genome**	**Gene#**	**Condition#**
Clavispora lusitaniae	GSE162151[70]	ATCC42720_w_CBS_6936_MT	5851	54
Cryptococcus neoformans	GSE226255[71]	CP022321.1	6718	720
Saccharomyces cerevisiae	GSE234431	GCA_000146045.2	5498	897
Coccidioides posadasii	GSE279406	NC_089407.1	8103	144
Blumeria graminis	GSE69215[72]	GCA_000151065.3	6329	48
Lipomyces starkeyi	GSE276000	GCA_001661325.1	7928	36
Neurospora crassa	GSE181566	GCF_000182925.2	9532	80
Candida albicans	GSE159545[73]	GCF_000182965.3	5918	75
Orbilia oligospora	GSE245014	GCF_000225545.1	11195	24
Summary			67,072	2,078

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

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

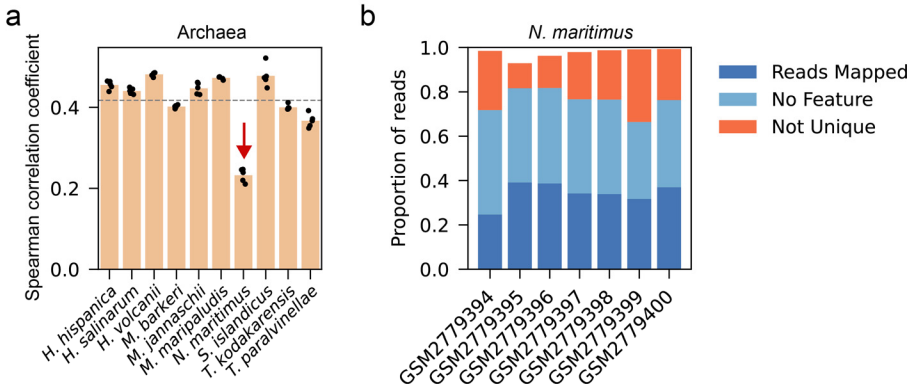
Species	**Reference**	**Reference genome**	**Gene#**	**Condition#**
Acinetobacter baumannii	Menon, et al.[74]	CP008706.1	1788	139
Bacillus subtilis	Sastry, et al.	NC_000964.3	1969	265
Escherichia coli	Lamoureux, et al.[75]	U00096.3	2490	1035
Limosilactobacillus reuteri	Josephs-Spaulding, et al.[76]	GCA_020785475.1	1159	117
Mycobacterium tuberculosis	Yoo, et al.[77]	AL123456.3	3214	647
Pseudomonas aeruginosa	Rajput, et al.[78]	NC_002516.2	3981	411
Pseudomonas putida	Lim, et al.[1]	AE015451.2	4379	321
Pseudomonas syringae	Bajpe, et al.[79]	AE016853.1	4296	202
Staphylococcus aureus	Poudel, et al.[80]	NC_010079.1	1736	385
Synechococcus elongatus	Yuan, et al.[81]	NC_007604.1	1522	300
Salmonella enterica	Yuan, et al.[81]	AE006468.2	1808	533
Streptococcus pyogenes	Hirose, et al.[82]	CP008776.1	1173	116
Vibrio natriegens	Shin, et al.[4]	GCA_001456255.1	1981	148
Summary			31,496	4,619

512 **Supplementary Figures**

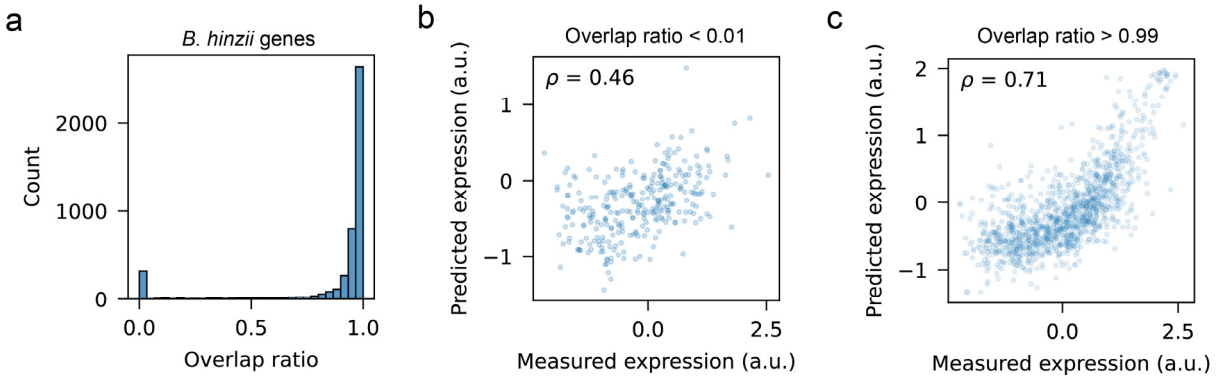


513

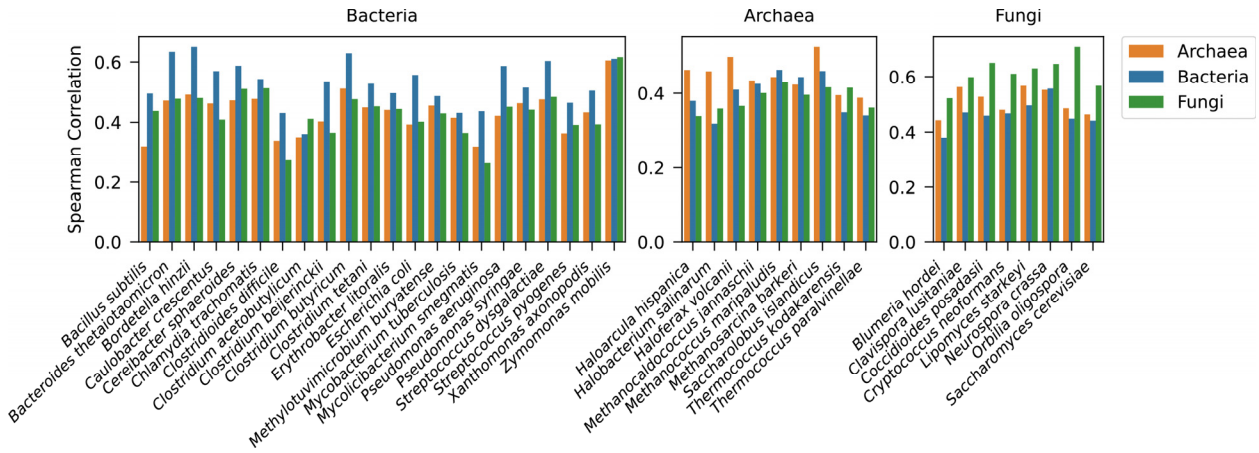
514 **<u>Supplementary Figure 1:</u> Prediction of gene expression variation.** Spearman correlations between
515 predicted and measured gene expression variations in leave-one-genome-out experiments for bacterial
516 **(a)**, archaeal **(b)**, and fungal **(c)** species. Results from five rounds of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



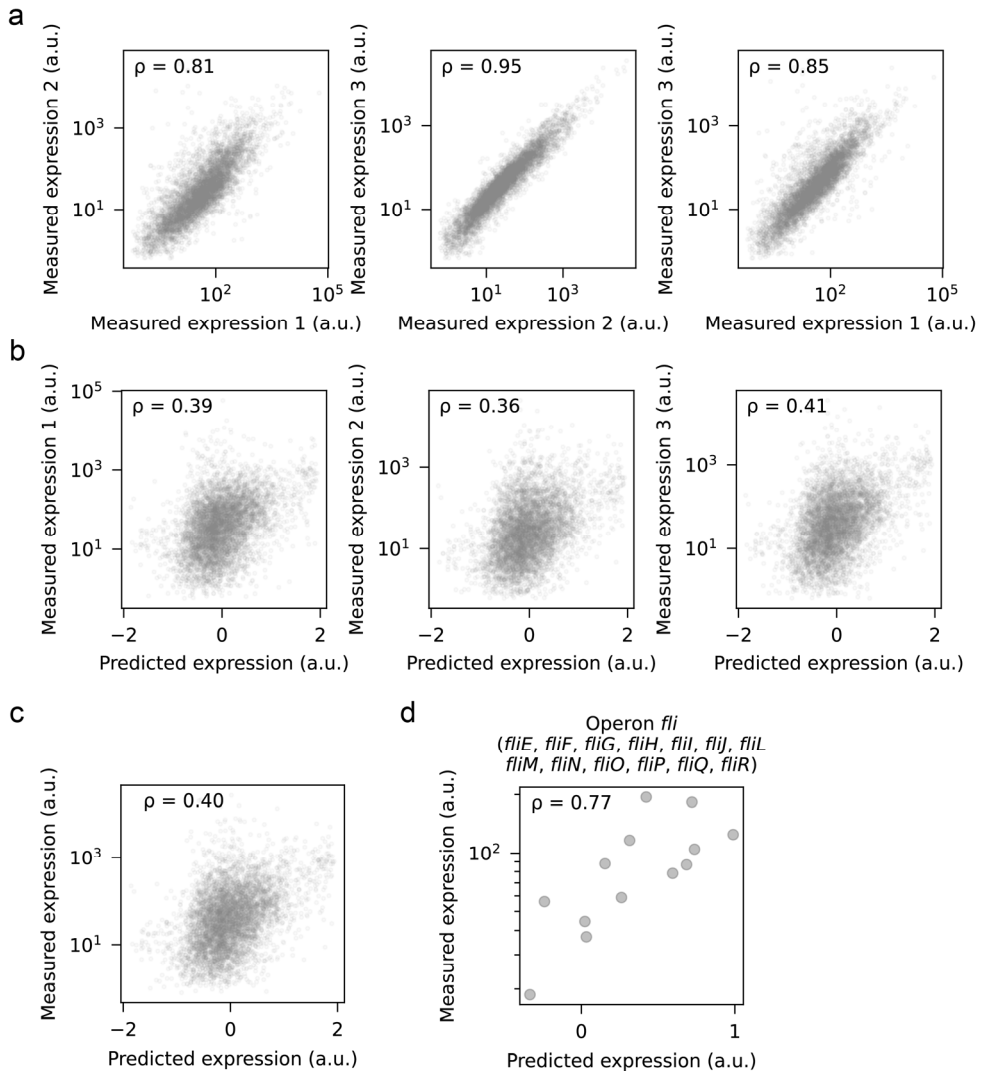
<u>Supplementary Figure 2:</u> Detection of issues related to RNA-seq dataset. a) Spearman correlation coefficients between predicted and measured gene expression for all archaeal species in the initial evaluation. Results from five rounds of model training and testing are shown (n = 5). **b)** Proportion of mapped/unmapped reads in the *N. maritimus* dataset (GSE103699).



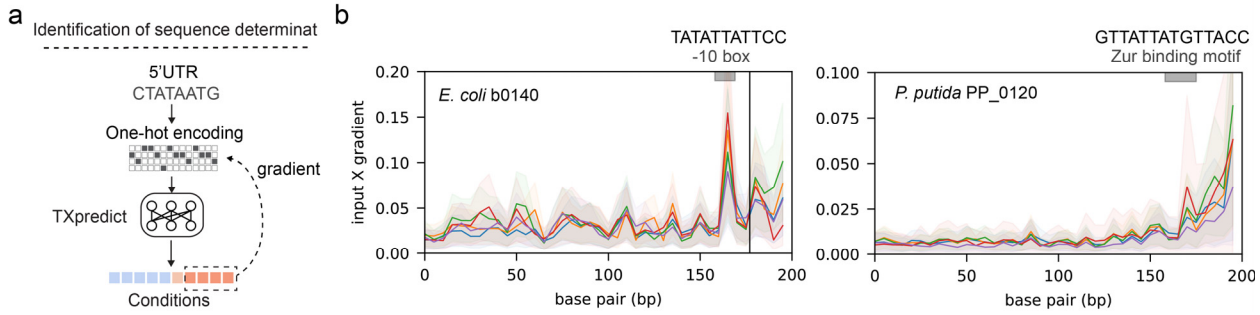
<u>Supplementary Figure 3:</u> Prediction of gene expression for rare proteins. a) Histogram showing the distribution of overlap ratios between *B. hinzii* proteins and proteins in the training sequences. Overlap ratios were calculated as the aligned sequence length divided by the query length. **b)** Predicted versus measured gene expression for proteins with low homology to the training set (overlap ratio < 0.01). **c)** Predicted and measured gene expressions for proteins with close homologs in the training dataset (overlap > 0.99).



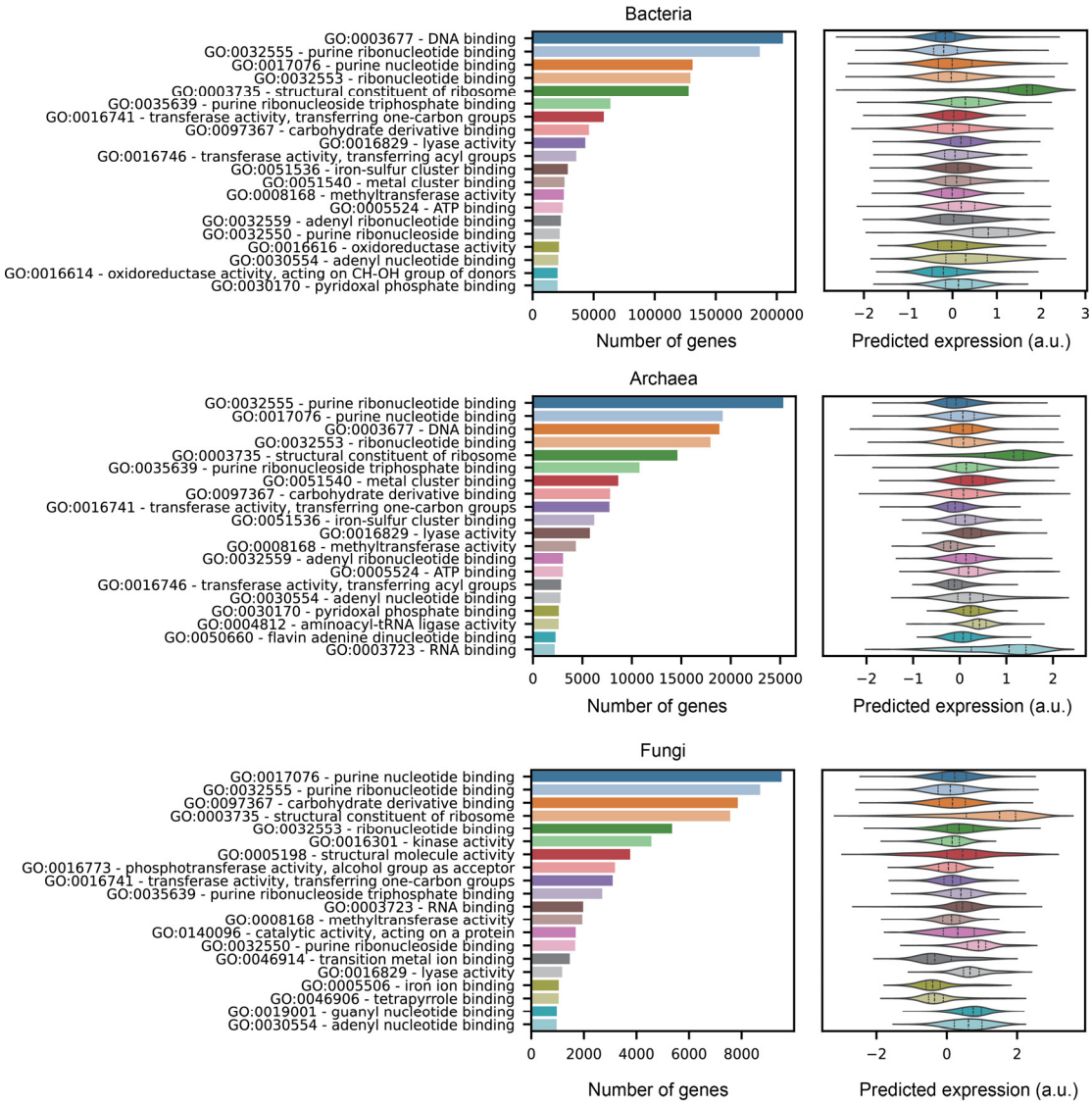
<u>Supplementary Figure 4:</u> Cross-domain prediction performance of TXpredict. TXpredict was trained on all species from a single domain and evaluated on species from the two other domains. Results from three domain-specific models are shown: bacteria (blue), archaea (orange), and fungi (green).



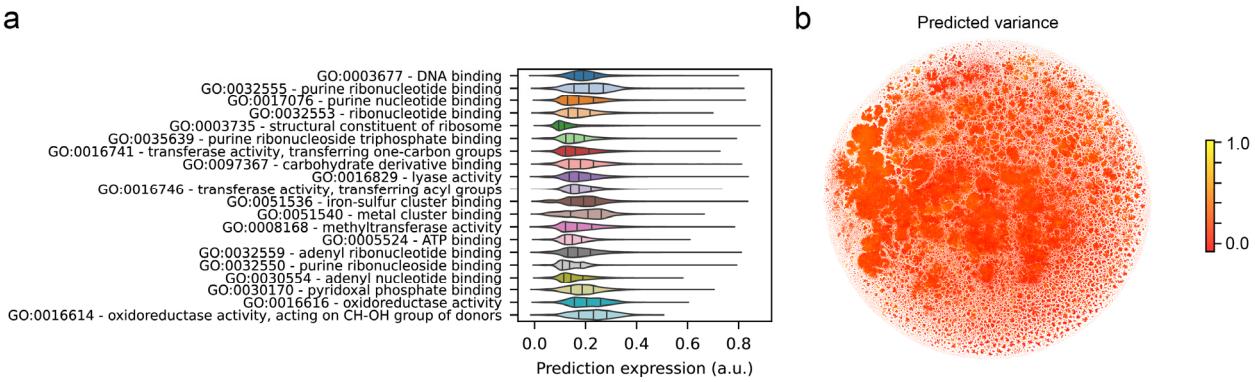
<u>Supplementary Figure 5:</u> Performance of TXpredict on *Pseudomonas sp.* WBC-3. a) Correlation of measured gene expression values of *Pseudomonas sp. WBC-3* across three experimental conditions. **b)** Predicted versus measured gene expression under each condition. **c)** Scatter plot comparing predicted and mean measured gene expression across three conditions for all genes. **d)** Predicted and measured gene expression levels for genes in the *fli* operon.



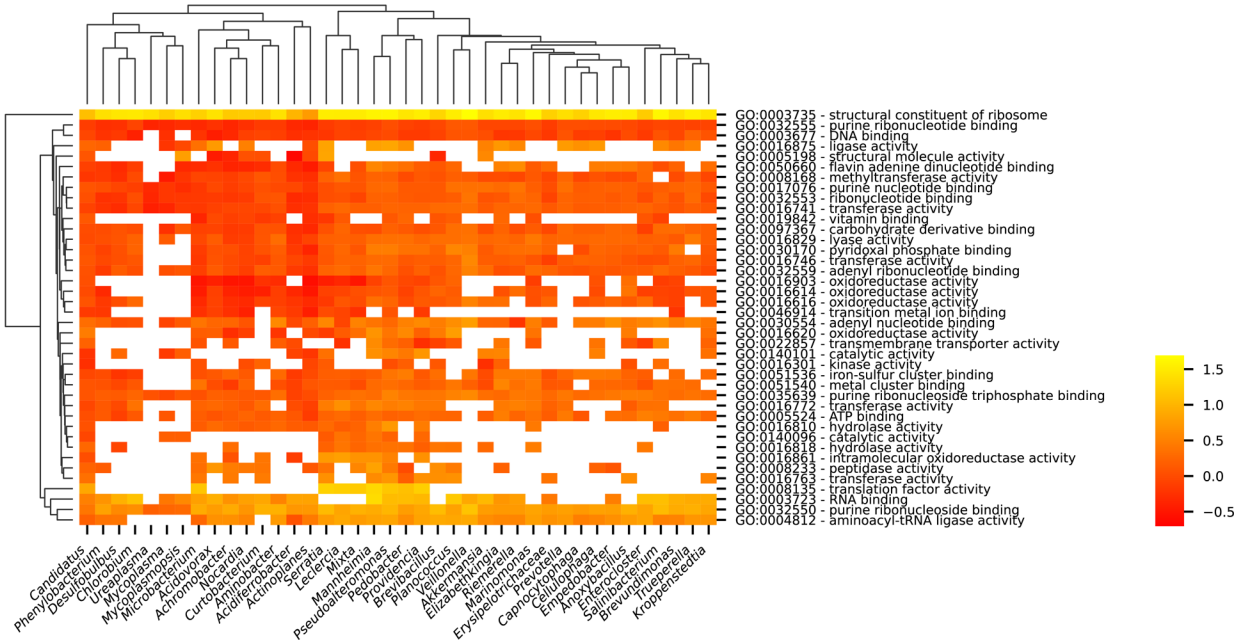
541 **<u>Supplementary Figure 6:</u> Sequence determinant of condition-dependent gene expression. a)**
542 Calculation of input gradients for identifying sequence determinants of condition-dependent gene
543 expression. **b)** Two examples of input × gradients and their correlations with known sequence motifs.



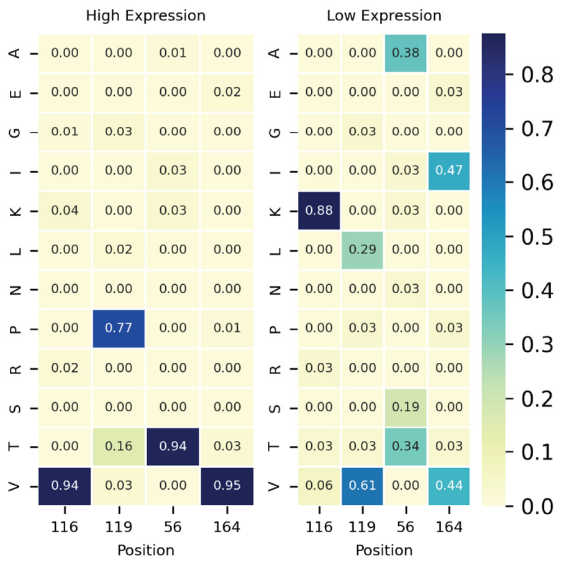
<u>Supplementary Figure 7:</u> Top GO terms and the predicted gene expressions in TXpredictDB. Top 20 Gene Ontology (GO) terms for genes from newly added bacterial, archaeal, and fungal genomes, as annotated by the DeepFRI model[28]. Predicted gene expression levels for these genes are shown alongside their associated GO terms.



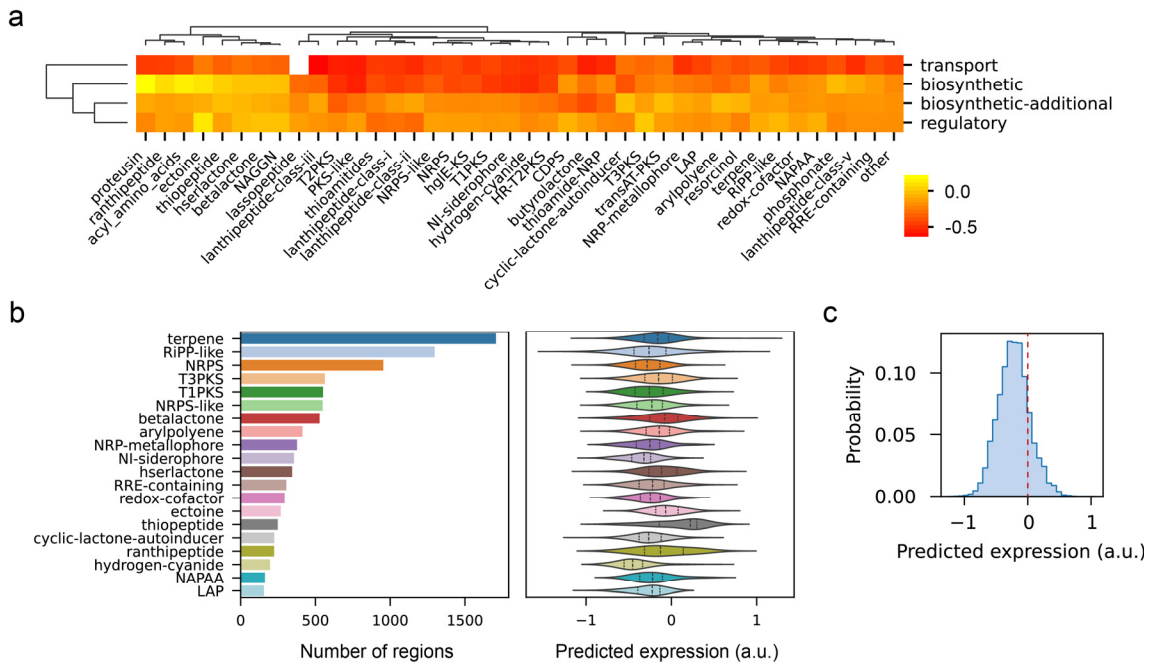
<u>Supplementary Figure 8:</u> Prediction of gene expression variation in newly added genomes. a) Predicted gene expression variation for genes associated with the top 20 GO terms in the newly added genomes. **b)** Predicted gene expression variation visualized in t-SNE plots of protein embeddings (n = 9.8M).



552 **<u>Supplementary Figure 9:</u> Prediction of gene expression for understudied genera.** Gene expression
553 values for the top 40 GO terms and 40 genera are shown. Only genus–GO term combinations with more
554 than 10 associated genes are included, with mean gene expression levels presented.



<u>Supplementary Figure 10:</u> Amino acid frequency differences between low and high expression genes within a protein cluster. TXpredict was used to predict gene expression for proteins in cluster ERS451416_00727, identified using MMseq2[31] at a 90% sequence identity threshold. Differences in amino acid frequencies between lowly and highly expressed genes within the cluster are shown.



559 **<u>Supplementary Figure 11:</u> Predicted gene expression across biosynthetic gene clusters (BGCs). a)**
560 Predicted gene expression levels for core biosynthetic, additional biosynthetic, regulatory, and transport
561 genes across different BGC types. **b)** Number of regions and predicted expression levels for the top 20
562 BGC types. **c)** Histogram of mean expression levels of BGCs (n = 11,400).