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.

Introduction

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 validated the model's performance in the non-model organism *Pseudomonas sp.* WBC-3 by conducting a series of RNA-seq experiments. By leveraging large-scale measurements from Deep Mutational Scanning (DMS) assays, we found that TXpredict's predictions of mutant sequence expression correlate with experimentally measured protein stability for specific proteins. 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

53 expression profiles for 9.8 million genes. The models developed in this study are publicly available:
54 https://github.com/lingxusb/TXpredict

55

56 **Results**

57 **Prediction of transcriptome based on the genome sequence**

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

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

79 In our initial evaluation, TXpredict performed poorly on the archaeon *N. maritimus*, with a Spearman
80 correlation of 0.23 in the leave-one-genome-out evaluation (Supplementary Fig. 2). Upon closer
81 examination of this dataset, we found that fewer than 40% of the reads were successfully mapped to
82 reference genome, demonstrating the potential of our model to detect potential problems in RNA-seq

83 data. In addition, we examined whether protein sequence similarity affects model performance. For
84 bacterium *B. hinzii*, we used BLASTP[26] to compute sequence similarity between its proteins and those in
85 the other bacterial strains. The model was trained on other species and tested on rare proteins from *B.*
86 *hinzii* that were not present in the training data, showing a correlation of 0.46 for them (Supplementary
87 Fig. 3). While this is lower than the correlation for all proteins (Fig. 1c), it suggests that the model
88 captures expression-related features beyond sequence similarity.

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

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

101

102 **Modeling context-dependent gene expression**

103 We next extended our framework to model condition-dependent gene expression (Supplementary Fig.
104 6). We collected RNA-seq profiles from 13 bacterial strains across 4.6k unique conditions from
105 iModulonDB, resulting in a total of 12.0M measurements (Supplementary Table 2). The model
106 integrated one-hot encoded 5'UTR sequences with protein embeddings as input features. For each gene,
107 we calculated the correlation between measured and predicted expression values across all conditions.
108 Our TXpredict model achieved an average correlation coefficient of 0.52 across all bacterial strains. For
109 *E. coli*, the distributions of Pearson's and Spearman's correlation coefficients show similar patterns, with
110 41% and 38% of genes having correlation coefficients above 0.5.

111 To further interpret the model's predictions, we calculated input gradients[17] to identify sequence motifs
112 contributing to gene expression variation across conditions (Supplementary Fig. 6). In *E. coli*, we

113 observed a peak in the input × gradient overlap with the -10 box of the sigmaS-dependent promoter of
114 the *yadV* gene[27]. Also input gradients increased after the transcriptional start site, suggesting potential
115 contributions from transcribed sequences. In a non-model organism *P. putida*, we identified a peak
116 overlapping with a Zur binding motif for the *PP_0120* gene, consistent with its role as a Zur iModulon
117 gene[1]. These findings illustrate our model's ability to capture condition-specific gene expression changes
118 in an interpretable manner.

119

120 **TXpredict captures protein mutation effects**

121 Deep Mutational Scanning (DMS) assays have been widely used to map protein fitness landscapes by
122 large-scale measurement of phenotypic effects of protein mutants[28,29]. We leveraged the
123 comprehensive benchmarks provided by ProteinGym[30] to investigate the gene expression effects
124 associated with these mutants (Fig. 2a). We selected yeast protein datasets, as our fungal model
125 achieved the highest accuracy in leave-one-genome-out evaluations. We observed strong correlations
126 between predicted gene expression and experimentally measured mutant effects for specific proteins,
127 such as THO1 and MYO3 (Fig. 2b and c), suggesting a potential link between their expression levels and
128 the functional outcomes. We then compared TXpredict predictions against zero-shot predictions from
129 several machine learning-based protein models[31–38] (Fig. 2d). TXpredict's performance was only
130 moderately lower than two inverse-folding models (ESM-IF1[36] and ProteinMPNN[34]) and outperformed
131 the ESM2 model, despite the fact that TXpredict utilizes ESM2 embeddings as input. TXpredict also
132 showed weaker correlations with predictions from MSA transformer[32] and TranceptEVE[35]. A detailed
133 residue-level analysis for THO1 revealed that TXpredict successfully captures the negative impact of an
134 arginine substitution at residue 19 within the SAP domain (Fig. 2e). We found lower correlations
135 between TXpredict predictions and mutation effects for bacterial datasets, possibly due to our model's
136 reduced accuracy for these genomes (Supplementary Fig. 7). Interestingly, for the yeast protein GAL4,
137 predictions derived from partial mutant sequences correlated more strongly with experimental
138 measurements than predictions from the full-length sequences (Fig. 2f).

139

140 **Application of TXpredict to new genomes**

141 We applied the trained TXpredict model to predict transcriptomes for new genomes including 2,169
142 bacterial, 423 archaeal and 93 fungal species (Fig. 3a and b). Collectively, these new genomes span 1,744

genera, of which 82% have no prior RNA-seq data according to publicly available data in the NCBI GEO database (Fig. 3 c and d). To explore the relationship between predicted gene expression and protein function, we used DeepFRI[39] to predict Gene Ontology (GO) terms for the new genes. Among the GO terms with the highest predicted expression levels were structural constituents of ribosome (GO:0003735) (Fig. 3e, Supplementary Fig. 8). In contrast, GO terms with the lowest predicted expression included proteins with DNA-binding activity (GO:0003677), consistent with their function as transcriptional regulators. Collectively, our model predicted the expression of up to 9.8M genes across these genomes, which were compiled into a database named TXpredictDB. The predicted gene expressions in different domains followed similar distributions (Fig. 3f). When visualizing protein embeddings of all proteins, we observed a clear separation of bacterial, archaeal and fungal genes (Fig. 3g, Supplementary Fig. 9). 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 explore the transcriptional landscape of understudied species, we focused on genera without prior RNA-seq data and grouped genes based on their annotated Gene Ontology (GO) terms (Supplementary Fig. 10). There was a positive correlation between the predicted expression levels of tRNA ligases and ribosomal proteins across genera (Fig. 3h, Spearman correlation 0.52). *Actinoplanes* had the lowest predicted expression for ribosomal proteins, in accord with their known slow growth rate[40]. The genus *Candidatus* includes a symbiotic bacterium found in carpenter ants (*Blochmanniella*) and showed low expression of both tRNA ligases and ribosomal proteins. This pattern likely reflects its reduced need for protein turnover and genome stasis[41], in contrast to free-living bacteria. We further visualized the embeddings of tRNA ligases to examine their sequence variations (Fig. 3i). tRNA ligases from *Candidatus* formed distinct low-expression clusters, whereas those from *Acidiferrobacter* were distributed across multiple clusters despite similarly low expression levels, suggesting divergent evolutionary constraints associated with their ecological niche.

We also clustered protein sequences using MMseqs2[42] and predicted gene expression within each high-similarity cluster using the TXpredict model. Interestingly, several clusters displayed non-unimodal expression distributions, with conserved amino acid differences separating low- and high-expression variants (Fig. 3j). For example, in cluster of the PTS IIA component homologs, low-expression variants shared a V116K substitution (Supplementary Fig. 11), which could reduce translation efficiency and mRNA stability due to the presence of constitutive lysine residues[43,44].

174 We further analyzed the predicted gene expression profiles of biosynthetic gene clusters (BGCs) across
175 bacterial species in TXpredictDB, which were identified using antiSMASH[45] (n = 11,400). Although BGCs
176 are typically lowly expressed, distinct patterns were observed across different cluster types. For
177 example, the proteusin BGC showed high expression of core biosynthetic genes, whereas the ectoine
178 BGC exhibited elevated expression of regulatory genes (Supplementary Fig. 12). These reference
179 expression profiles in non-model organisms may offer valuable insights into the functional organization
180 of BGCs.

181

182 **Discussion**

183 We present TXpredict, a transcriptome prediction model that can generalize to a wide range of
184 microbes. For a microbial genome with 4.6k genes, transcriptome prediction can be completed in
185 approximately 22 minutes using a Google Colab notebook (Methods), demonstrating the model's
186 efficiency and accessibility. Our approach is not without limitations: while it effectively leverages protein
187 embeddings for expression prediction, it does not account for the effects of cis-regulatory elements.
188 Moreover, our model's performance is constrained by the limited training profiles, especially for
189 archaeal species. Lastly, we provide a single reference transcriptome per genome in TXpredictDB, which
190 does not capture gene expression dynamics. Despite these limitations, TXpredict has the potential to
191 complement experimental approaches and advance the study of the vast number of sequenced yet
192 poorly characterized microbes. For example, identifying conserved expressed genes in pathogens may
193 enable the discovery of novel drug targets[46]. Furthermore, predicting gene expression across genomes
194 can provide insight into how genome plasticity contributes to bacterial pathogenicity[47]. Our model may
195 also suggest biologically plausible mRNA expression levels as starting points for heterologous gene
196 expression in synthetic biology applications. These efforts could further benefit from integrating single-
197 cell RNA-seq data [5,48,49] and computational tools for mRNA synthesis and stability prediction[50–52], paving
198 the way for a deeper understanding of microbial transcriptional landscapes.

Figures

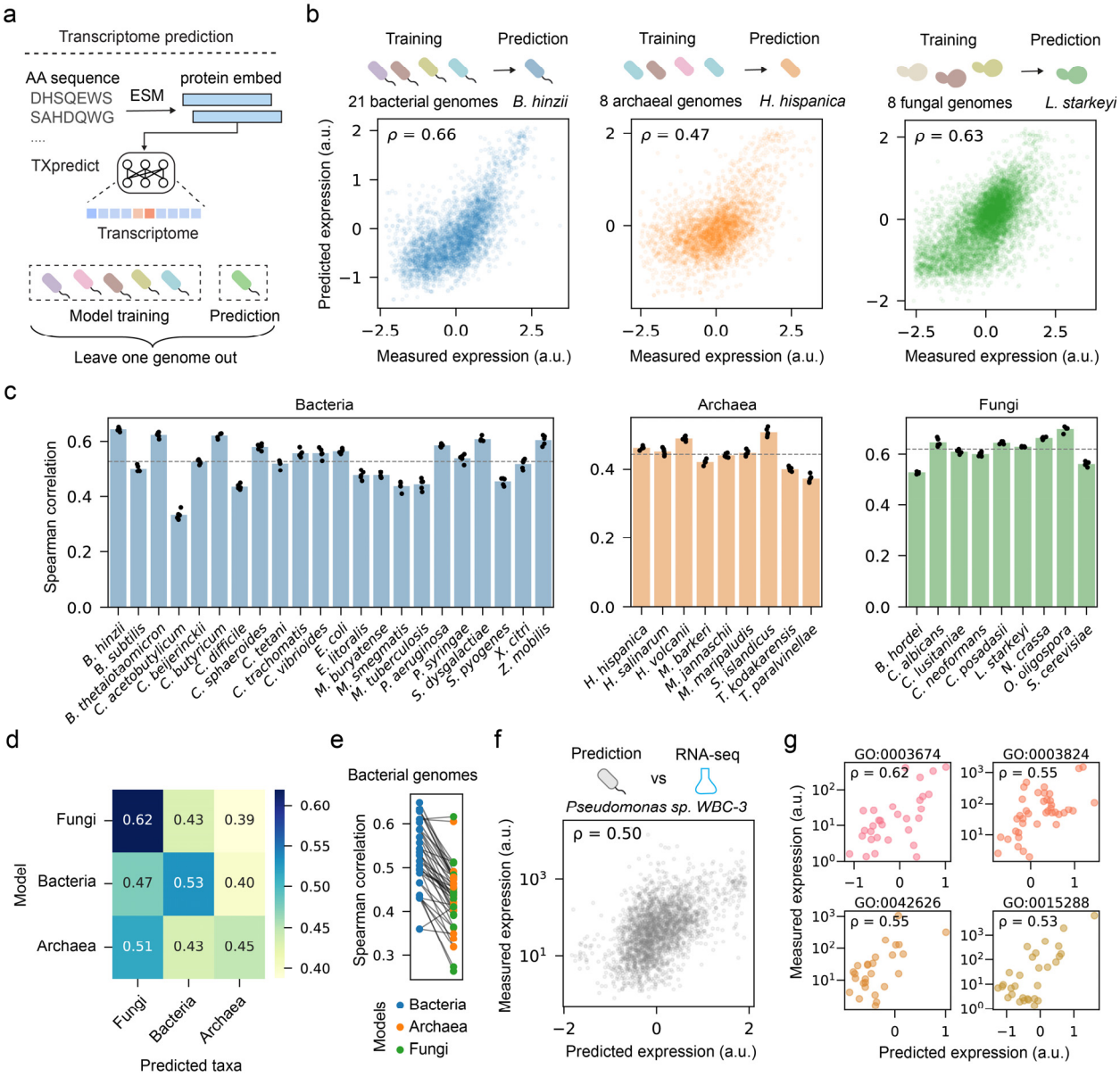


Figure 1. Prediction of transcriptome based on the genome sequence

a) Schematic of TXpredict for transcriptome prediction. **b)** Scatter plots showing predicted and measured gene expression in leave-one-genome-out evaluations for one species from bacteria, archaea, and fungi. Spearman correlation coefficients are reported in the top-left corner of each plot. **c)** Spearman correlation coefficients between predicted and measured gene expression in leave-one-genome-out evaluations. TXpredict models were separately trained for bacterial, archaeal, and fungal species. Results from five rounds of model training and testing are reported (n = 5). Dashed lines

208 indicate the mean correlations for each domain. **d)** Cross-domain prediction performance of TXpredict
209 models. Models were trained on all species from one domain and tested on species from the other
210 domains. Mean Spearman correlations across all test species are shown. Diagonal values represent
211 leave-one-genome-out performance within each domain. **e)** Predictions of archaeal and fungal models
212 on all bacterial species. Blue dots indicate leave-one-genome-out evaluations for bacteria. Predictions
213 for the same species were connected. **f)** Model predictions versus measured gene expression in new
214 RNA-seq experiment for *Pseudomonas sp.* WBC-3. Each dot represents one functional gene and the
215 Spearman correlation is shown. **g)** Predicted and measured gene expression for genes with different GO
216 annotations in *Pseudomonas sp.* WBC-3.

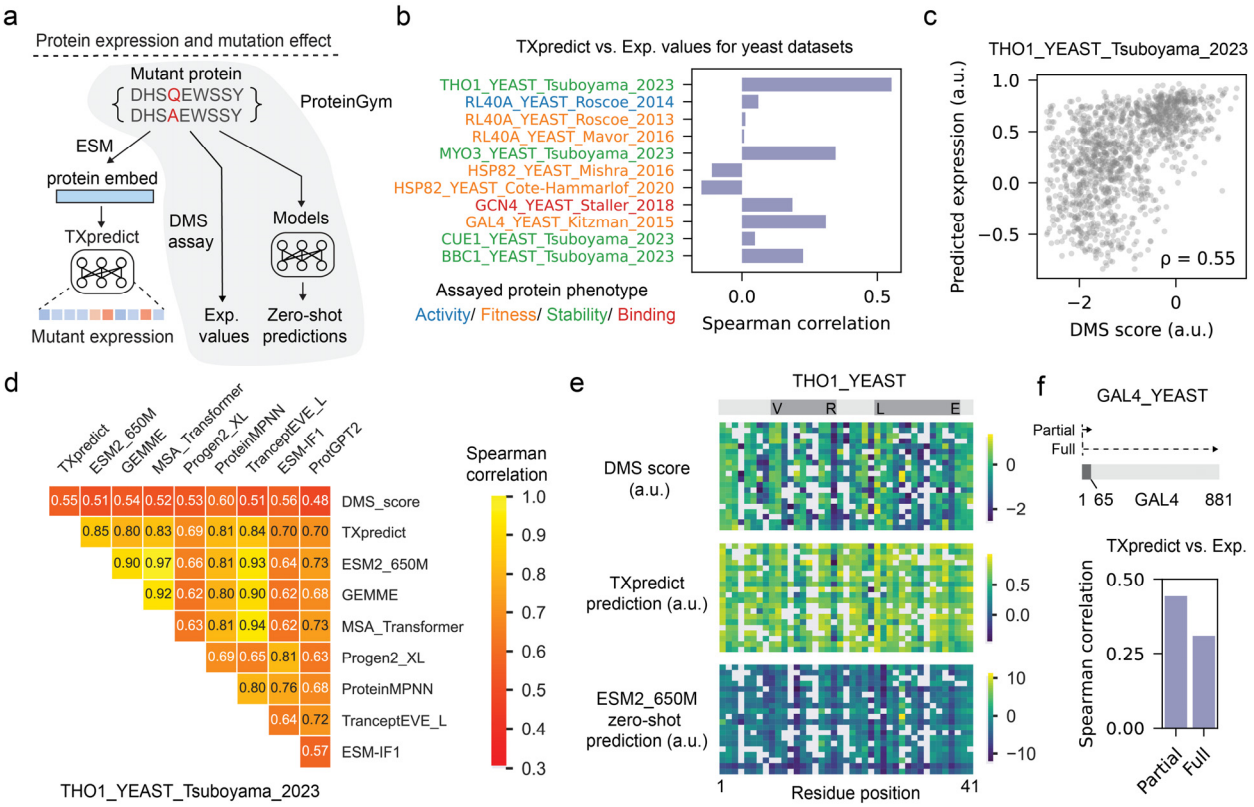


Figure 2. TXpredict captures protein mutation effects

a) Schematic showing relationships among TXpredict predictions, experimentally measured deep mutational scanning (DMS) scores, and zero-shot predictions from ProteinGym[30]. **b)** Spearman correlation between TXpredict predictions and experimental DMS scores in yeast datasets. **c)** Scattering plot showing predicted expression versus DMS scores for the THO1 protein mutants (n = 1,279). Spearman correlation coefficient is reported in the bottom-right corner. **d)** Cross-correlations of TXpredict, zero-shot model predictions, and DMS scores for THO1 mutants. **e)** Detailed comparison of DMS scores, TXpredict predictions, and ESM2 predictions across residue positions for single amino acid mutations in THO1. The alpha-helical domains are indicated by dark gray boxes. **f)** Correlation of TXpredict predictions for either a partial region or the full-length GAL4 protein with experimentally measured mutation effects. The mutated part of GAL4 is colored in dark gray.

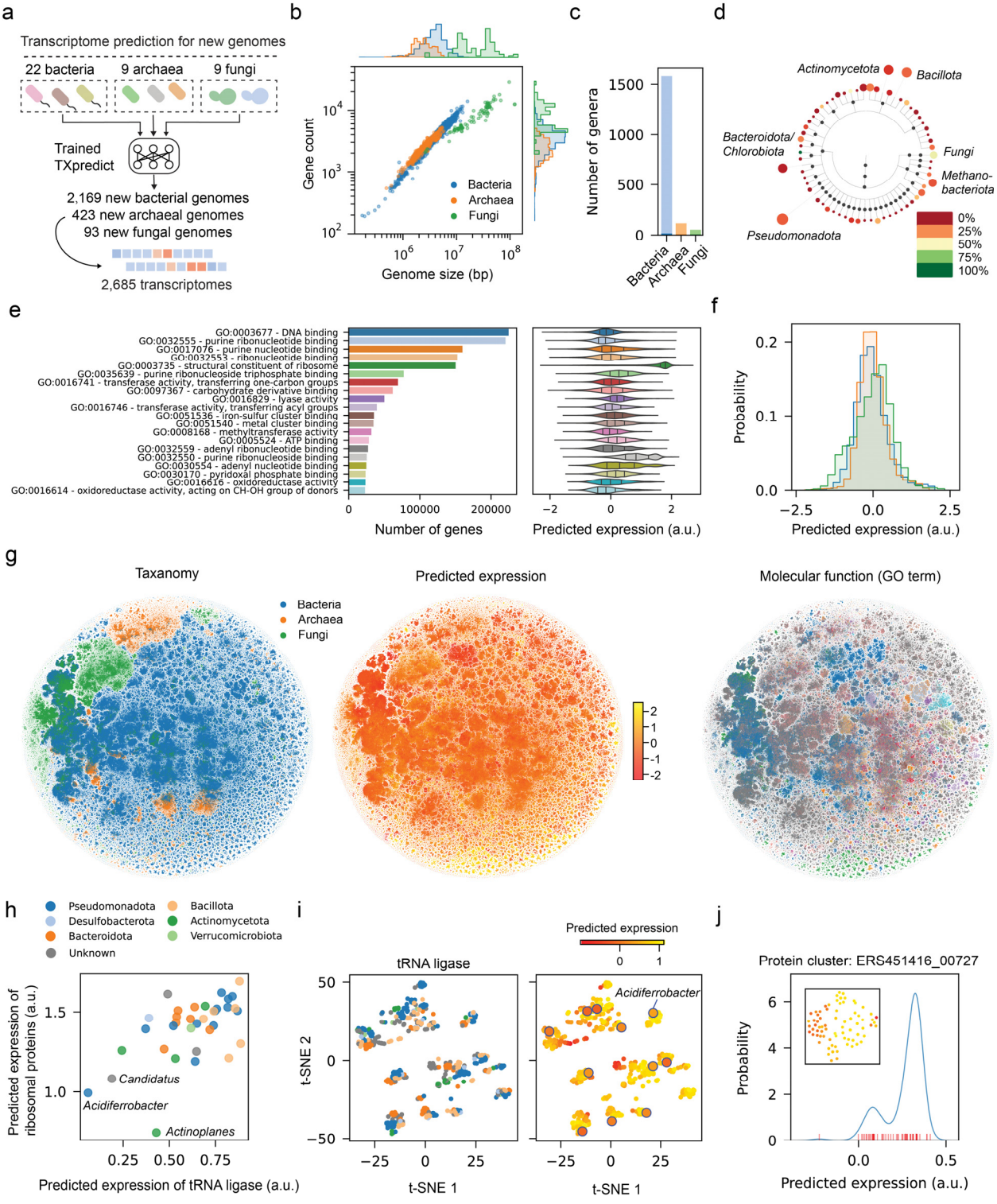


Figure 3. Application of TXpredict to new genomes

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

233 colors). **d)** Polygenetic tree of newly added genomes. Each dot represents a phylum. The dot size is
234 proportional to the number of genera it contains and color indicates the proportion of genera with
235 existing RNA-seq data based on the NCBI GEO database. **e)** Top 20 GO terms of the genes in the new
236 genomes calculated by the DeepFRI model[39] (left) and their predicted gene expression levels (right). **f)**
237 Distributions of predicted gene expression levels across new bacterial (blue), archaeal (orange) and
238 fungal (green) genomes. **g)** t-SNE plots of protein embeddings generated by the ESM2 model[24] (n =
239 9.8M). Each dot represents a gene in the new genomes, colored by taxa (left), predicted gene expression
240 (middle), and GO terms (right). GO terms are colored as in panel (e). **h)** Predicted expression of tRNA
241 ligases and ribosomal proteins across understudied genera (n = 40), colored by phylum. **i)** t-SNE plots of
242 tRNA ligase embeddings across understudied genera, colored by phylum (left) and predicted expression
243 level (right). **j)** Histogram of predicted expression levels for proteins in a high-similarity cluster. Inset
244 shows the t-SNE plot of embeddings for these proteins, colored by predicted expression.

245 **Methods**

246 **Training dataset**

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

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

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

267

268 **Transcriptome prediction model**

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

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

276

277 **Model training and evaluation**

278 We performed cross-species evaluations for bacteria, archaea and fungi separately to account for
279 potential domain-specific expression patterns. For each domain, we employed a cross-validation
280 strategy where models were trained on data from all but one microbial species and tested on the held-
281 out species. This approach was repeated for each species in our dataset to assess our model's
282 generalization capability. The models were trained using the Adam optimizer with a learning rate of
283 0.0002 for 5 epoch with a batch size of 32. To ensure robust evaluation, we performed 5 independent
284 training runs with different random seeds (42, 123, 456, 789, 1011). The mean squared error (MSE) loss
285 function was used for optimization. Model performance was evaluated using both Spearman and
286 Pearson correlation coefficients between predicted and measured expression levels.

287 New bacterial, archaeal and fungal genomes were downloaded from NCBI Genbank
288 (https://ftp.ncbi.nlm.nih.gov/genomes/genbank/). To maximize genus-level representation for
289 transcriptome prediction, we selected 2,169 bacterial, 423 archaeal, and 93 fungal genomes spanning a
290 total of 1,744 genera. Gene expression for these genomes was predicted using TXpredict models,
291 forming the basis of the TXpredictDB. To determine whether a genus has been transcriptional
292 characterized, we searched the GEO database using the following query: "*genus_name [Organism] AND*
293 *("Expression profiling by array"[Filter])*". We annotated taxonomy of these genomes by querying the
294 GTDB database[53]. For model training and inference, we used 3090 Ti GPU (24GB) and the software
295 packages PyTorch (version 2.1.1) and transformer (version 4.28.1).

296

297 **Conditional gene expression model**

298 The conditional gene expression model shares the same transformer encoder architecture with the
299 transcriptome prediction model, and was trained separately for each species. The UTR branch processes
300 one-hot encoded sequences from -100 to +100 bp relative to the transcription start site. The
301 transformer layer has 2 attention heads and a feedforward dimension of 128. The protein embedding
302 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[39], 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[42] 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

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

333

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

335 For the RNA-seq experiment, *Pseudomonas* sp. WBC-3 was cultured under three different conditions:
336 PNP1 (minimal medium with 0.01% glycerol), PNP2 (minimal medium with 0.04% PNP), and PNP3
337 (minimal medium with 0.01% glycerol and 0.04% PNP). Each condition was prepared by diluting a 2×
338 stock of yeast extract minimal medium supplemented with either 0.02% glycerol, 0.08% PNP, or both,
339 resulting in the desired final concentrations. WBC-3 strain was first grown overnight in LB medium, then
340 diluted 1:100 into 4 mL of each experimental medium and incubated overnight at 30°C with shaking at
341 200 rpm. Cultures were harvested at OD600 between 0.5 and 1.0 by centrifugation for 2 minutes,
342 ensuring that the total collected biomass did not exceed 1×10^9 cells. Total RNA was extracted using the
343 RNAprep Pure Bacteria Kit (TIANGEN, DP430), followed by DNase I treatment (TIANGEN, RT411) to
344 remove genomic DNA contamination. RNA quality and concentration were confirmed to meet
345 sequencing requirements. Raw sequencing reads were aligned and processed using standard RNA-seq
346 pipelines. Read counts were quantified using StringTie with the following command: stringtie -p 8 -G
347 GCF_035758525.1_ASM3575852v1_genomic.gtf -l sample_id -o sample.gtf sample.bam, using the
348 *Pseudomonas* sp. WBC-3 reference genome (GCF_035758525.1).

349

350 **Code availability**

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

353 **References**

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

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

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

360 4. Shin, J., Rychel, K. & Palsson, B. O. Systems biology of competency in Vibrio
361 natriegens is revealed by applying novel data analytics to the transcriptome. *Cell Rep* **42**,
362 (2023).

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

365 6. Ryu, M.-H. *et al.* Control of nitrogen fixation in bacteria that associate with cereals. *Nat Microbiol*
366 **5**, 314–330 (2020).

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

369 8. Barnum, T. P. *et al.* Predicting microbial growth conditions from amino acid composition. *bioRxiv*
370 2023–2024 (2024).

371 9. Ramoneda, J. *et al.* Building a genome-based understanding of bacterial pH preferences. *Sci Adv*
372 **9**, eadf8998 (2024).

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

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

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

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

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

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

386 16. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature*
387 (2022) doi:10.1038/s41586-022-04506-6.

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

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

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

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

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

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

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

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

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

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

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

28. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat Methods* **12**, 203–206 (2015).

29. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).

30. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. in *Advances in Neural Information Processing Systems* (eds. Oh, A. et al.) vol. 36 64331–64379 (Curran Associates, Inc., 2023).

31. Laine, E., Karami, Y. & Carbone, A. GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Mol Biol Evol* **36**, 2604–2619 (2019).

32. Rao, R. M. *et al.* MSA Transformer. in *Proceedings of the 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) vol. 139 8844–8856 (PMLR, 2021).

33. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. in *Advances in Neural Information Processing Systems* (eds. Ranzato, M.,

423 Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 29287–29303 (Curran
424 Associates, Inc., 2021).

425 34. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN.
426 *Science (1979)* **378**, 49–56 (2022).

427 35. Notin, P. *et al.* TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein
428 Sequences for Improved Fitness Prediction. *bioRxiv* (2022).

429 36. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. in *Proceedings of the*
430 *39th International Conference on Machine Learning* (eds. Chaudhuri, K. et al.) vol. 162 8946–8970
431 (PMLR, 2022).

432 37. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: Exploring the
433 boundaries of protein language models. *Cell Syst* **14**, 968-978.e3 (2023).

434 38. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein
435 design. *Nat Commun* **13**, 4348 (2022).

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

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

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

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

445 43. Radhakrishnan, A. & Green, R. Connections underlying translation and mRNA stability. *J Mol Biol*
446 **428**, 3558–3564 (2016).

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

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

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

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

456 48. Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.-E. & Vogel, J. Single-cell RNA-sequencing
457 reports growth-condition-specific global transcriptomes of individual bacteria. *Nat Microbiol* **5**,
458 1202–1206 (2020).

459 49. Kuchina, A. *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science (1979)*
460 **371**, eaba5257 (2021).

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

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

466 52. Xia, Y. *et al.* Language model generates cis-regulatory elements across prokaryotes. *bioRxiv*
467 2011–2024 (2024).

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

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

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

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

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

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

482 59. D, S. M. *et al.* Arabinose-Induced Catabolite Repression as a Mechanism for Pentose Hierarchy
483 Control in Clostridium acetobutylicum ATCC 824. *mSystems* **3**, 10.1128/msystems.00064-18
484 (2018).

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

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

490 62. Espah Borujeni, A., Zhang, J., Doosthosseini, H., Nielsen, A. A. K. & Voigt, C. A. Genetic circuit
491 characterization by inferring RNA polymerase movement and ribosome usage. *Nat Commun* **11**,
492 5001 (2020).

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

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

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

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

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

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

506 69. Kachroo, P. *et al.* Integrated analysis of population genomics, transcriptomics and virulence
507 provides novel insights into Streptococcus pyogenes pathogenesis. *Nat Genet* **51**, 548–559
508 (2019).

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

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

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

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

517 74. D, T. B., Cem, M., B, N. T., Q, L. S. & F, H. J. Growth Kinetics, Carbon Isotope Fractionation, and
518 Gene Expression in the Hyperthermophile Methanocaldococcus jannaschii during Hydrogen-
519 Limited Growth and Interspecies Hydrogen Transfer. *Appl Environ Microbiol* **85**, e00180-19
520 (2019).

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

523 76. Spietz, R. L. *et al.* Investigating Abiotic and Biotic Mechanisms of Pyrite Reduction. *Front*
524 *Microbiol* **13**, (2022).

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

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

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

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

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

537 82. Hunsaker, E. W., Yu, C.-H. A. & Franz, K. J. Copper Availability Influences the Transcriptomic
538 Response of Candida albicans to Fluconazole Stress. *G3 Genes|Genomes|Genetics* **11**, jkab065
539 (2021).

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

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

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

547 86. Reo, Y. *et al.* Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a
548 Structured Interplay between Metabolism, Stress Response, and Infection. *mSphere* **7**, e00033-22
549 (2022).

550 87. Rajput, A. *et al.* Advanced transcriptomic analysis reveals the role of efflux pumps and media
551 composition in antibiotic responses of Pseudomonas aeruginosa. *Nucleic Acids Res* **50**, 9675–
552 9688 (2022).

553 88. Heera, B., Kevin, R., R, L. C., V, S. A. & O, P. B. Machine learning uncovers the Pseudomonas
554 syringae transcriptome in microbial communities and during infection. *mSystems* **8**, e00437-23
555 (2023).

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

558 90. Yuan, Y. *et al.* Machine learning reveals the transcriptional regulatory network and circadian
559 dynamics of Synechococcus elongatus PCC 7942. *Proceedings of the National Academy of*
560 *Sciences* **121**, e2410492121 (2024).

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

563

564 **Supplementary Table**

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

Bacteria				
Species	**GSE number**	**Reference genome**	**Gene#**	**Condition#**
Bacillus subtilis	GSE224332[54]	NZ_CP020102.1	4220	317
Bacteroides thetaiotaomicron	GSE251676[55]	GCF_000011065.1	4618	64
Bordetella hinzii	GSE181542	CP052845.1	4531	228
Caulobacter vibrioides	GSE246782[56]	NC_011916.1	3882	34
Cereibacter sphaeroides	GSE186600[57]	GCA_000012905.2	4232	17
Chlamydia trachomatis	GSE262652	GCA_000068585.1	857	30
Clostridioides difficile	GSE249810[58]	GCA_000009205.5	3560	24
Clostridium acetobutylicum	GSE107804[59]	GCA_000008765.1	3755	48
Clostridium beijerinckii	GSE150480[60]	GCF_000833105.2	5316	16
Clostridium butyricum	GSE150480[60]	GCF_001456065.2	3848	16
Clostridium tetani	GSE182408[61]	NC_004557.1	2562	39
Escherichia coli	GSE152664[62]	NC_010473.1	3260	9
Erythrobacter litoralis	GSE126532	GCA_001719165.1	3003	49
Methylotuvimicrobium buryatense	GSE162089[63]	NZ_CP035467.1	3927	56
Mycobacterium tuberculosis	GSE110508[64,65]	GCA_000195955.2	3957	87
Mycolicibacterium smegmatis	GSE69983[66]	GCA_000015005.1	6700	30
Pseudomonas aeruginosa	GSE251671	GCA_000014625.1	5659	858
Pseudomonas syringae	GSE103441[67]	GCA_000007805.1	5586	114
Streptococcus dysgalactiae	GSE272047[68]	GCA_029234115.1	1939	114
Streptococcus pyogenes	GSE113055[69]	CP000056.2	1608	442
Xanthomonas citri	GSE252662[70]	AE008923.1	4294	38
Zymomonas mobilis	GSE139939[71]	CP023715.1	1759	19
Summary			83,073	2,649

566 **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[72]	GCF_000006805.1	2325	24
Haloferax volcanii	GSE248564[73]	GCF_000025685.1	3873	39
Methanocaldococcus jannaschii	GSE112986[74]	GCF_000091665.1	1724	13
Methanococcus maripaludis	GSE261517[75]	GCF_000011585.1	1730	12
Methanosarcina barkeri	GSE168895[76]	GCF_000970025.1	3338	12
Sulfolobus islandicus	GSE220819[77]	GCF_000189555.1	2682	18
Thermococcus kodakarensis	GSE186817	NC_006624.1	2214	4
Thermococcus paralvinellae	GSE101078[78]	GCF_000517445.1	1995	15
Summary			25,448	169

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

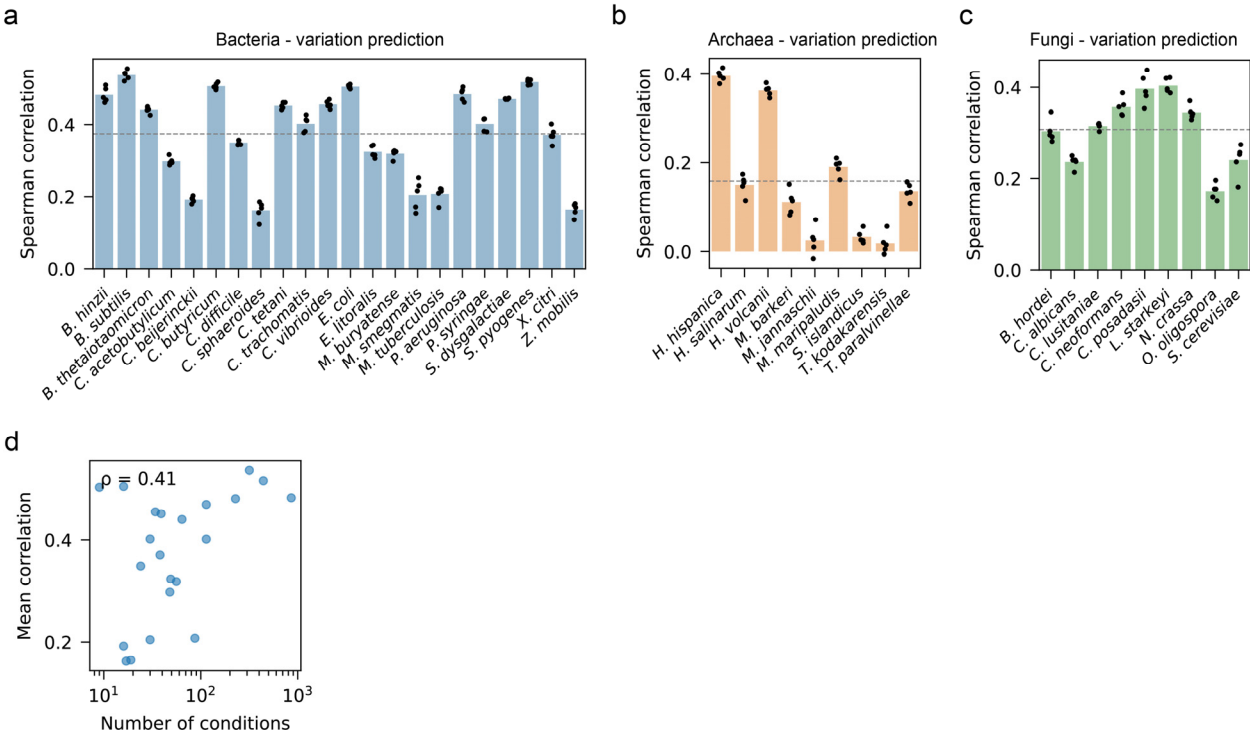
Fungi				
Species	**GSE number**	**Reference genome**	**Gene#**	**Condition#**
Clavispora lusitaniae	GSE162151[79]	ATCC42720_w_CBS_6936_MT	5851	54
Cryptococcus neoformans	GSE226255[80]	CP022321.1	6718	720
Saccharomyces cerevisiae	GSE234431	GCA_000146045.2	5498	897
Coccidioides posadasii	GSE279406	NC_089407.1	8103	144
Blumeria graminis	GSE69215[81]	GCA_000151065.3	6329	48
Lipomyces starkeyi	GSE276000	GCA_001661325.1	7928	36
Neurospora crassa	GSE181566	GCF_000182925.2	9532	80
Candida albicans	GSE159545[82]	GCF_000182965.3	5918	75
Orbilia oligospora	GSE245014	GCF_000225545.1	11195	24
Summary			67,072	2,078

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

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

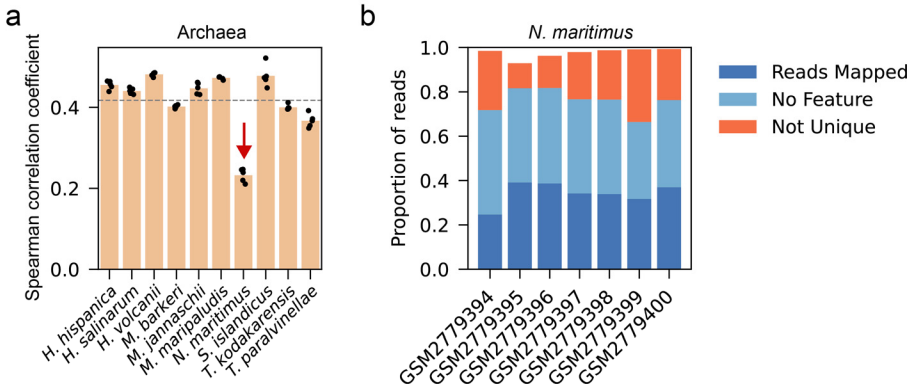
Species	**Reference**	**Reference genome**	**Gene#**	**Condition#**
Acinetobacter baumannii	Menon, et al.[83]	CP008706.1	1788	139
Bacillus subtilis	Sastry, et al.	NC_000964.3	1969	265
Escherichia coli	Lamoureux, et al.[84]	U00096.3	2490	1035
Limosilactobacillus reuteri	Josephs-Spaulding, et al.[85]	GCA_020785475.1	1159	117
Mycobacterium tuberculosis	Yoo, et al.[86]	AL123456.3	3214	647
Pseudomonas aeruginosa	Rajput, et al.[87]	NC_002516.2	3981	411
Pseudomonas putida	Lim, et al.[1]	AE015451.2	4379	321
Pseudomonas syringae	Bajpe, et al.[88]	AE016853.1	4296	202
Staphylococcus aureus	Poudel, et al.[89]	NC_010079.1	1736	385
Synechococcus elongatus	Yuan, et al.[90]	NC_007604.1	1522	300
Salmonella enterica	Yuan, et al.[90]	AE006468.2	1808	533
Streptococcus pyogenes	Hirose, et al.[91]	CP008776.1	1173	116
Vibrio natriegens	Shin, et al.[4]	GCA_001456255.1	1981	148
Summary			31,496	4,619

570 **Supplementary Figures**

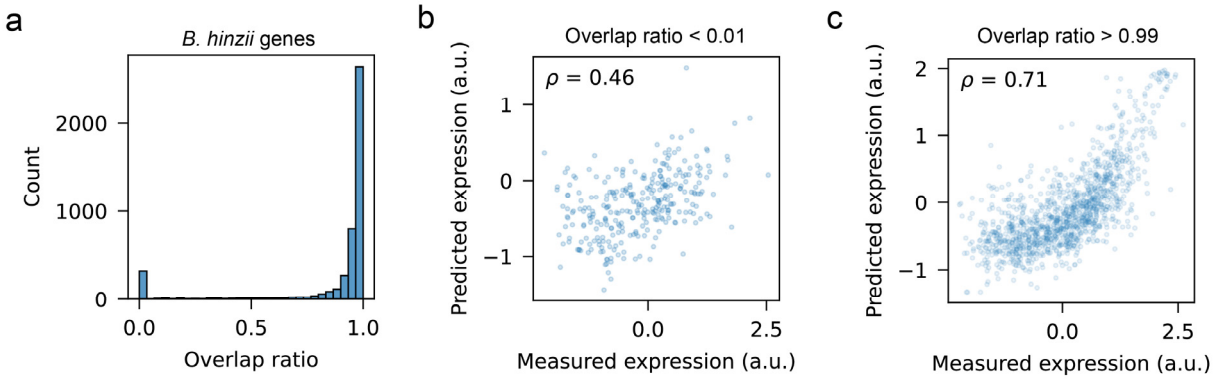


571

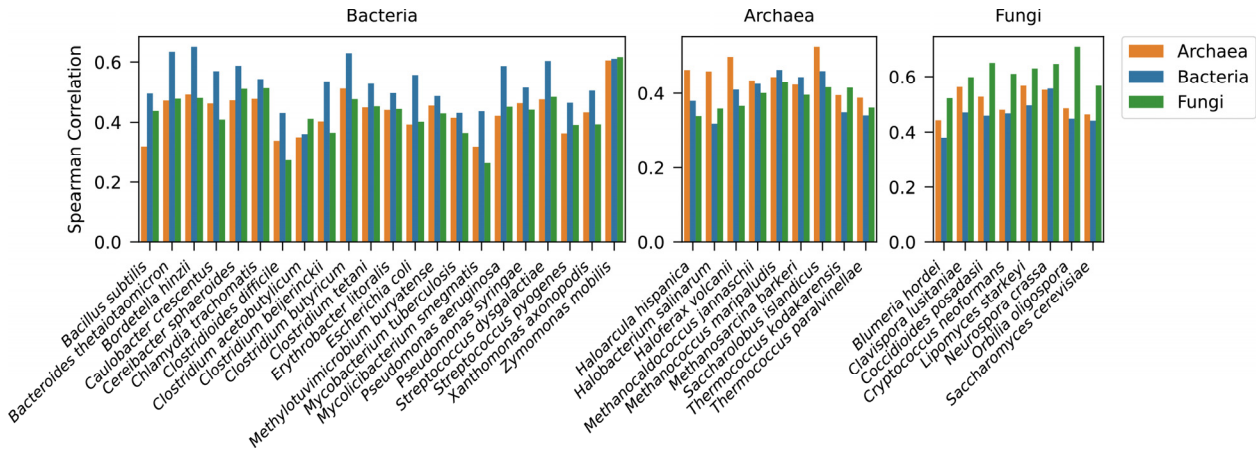
572 **<u>Supplementary Figure 1:</u> Prediction of gene expression variation.** Spearman correlations between
573 predicted and measured gene expression variations in leave-one-genome-out experiments for bacterial
574 **(a)**, archaeal **(b)**, and fungal **(c)** species. Results from five rounds of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



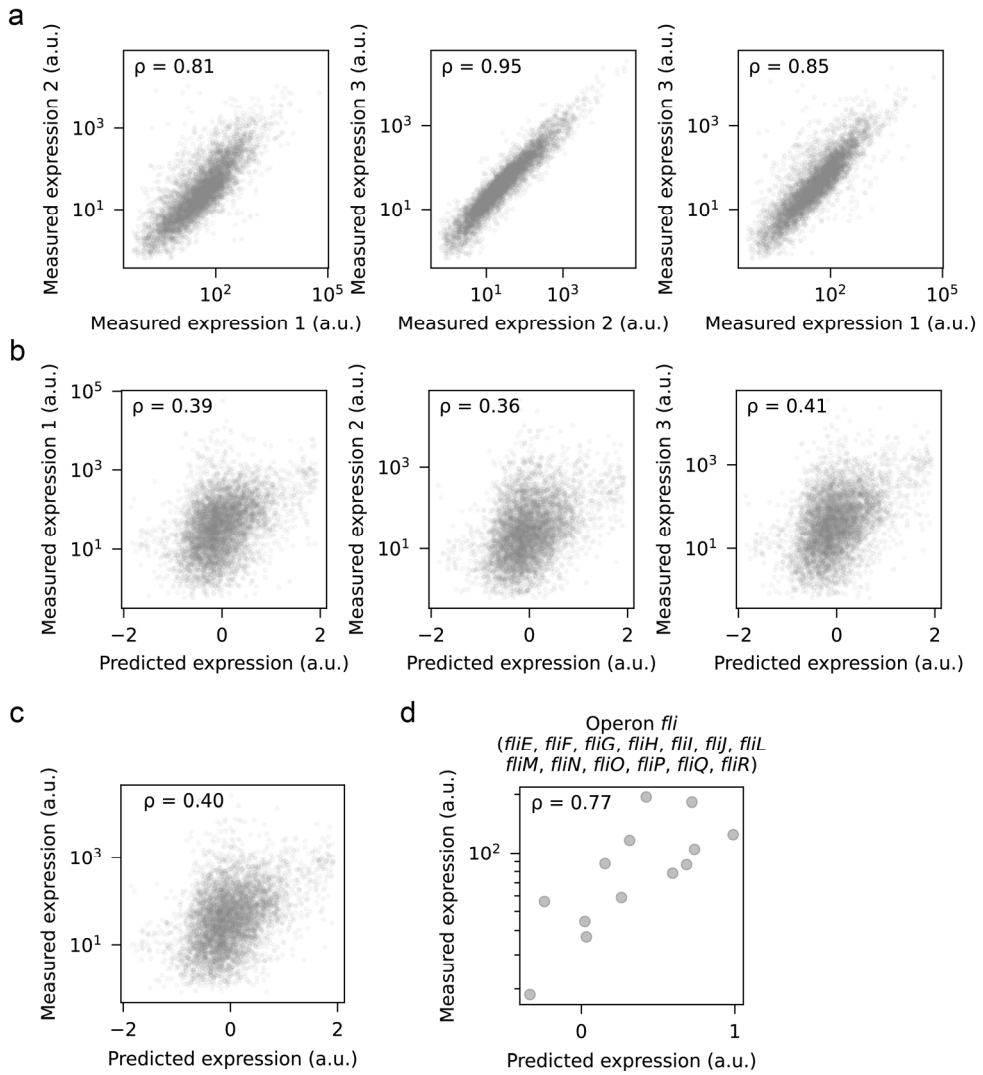
<u>Supplementary Figure 2:</u> Detection of issues related to RNA-seq dataset. a) Spearman correlation coefficients between predicted and measured gene expression for all archaeal species in the initial evaluation. Results from five rounds of model training and testing are shown (n = 5). **b)** Proportion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



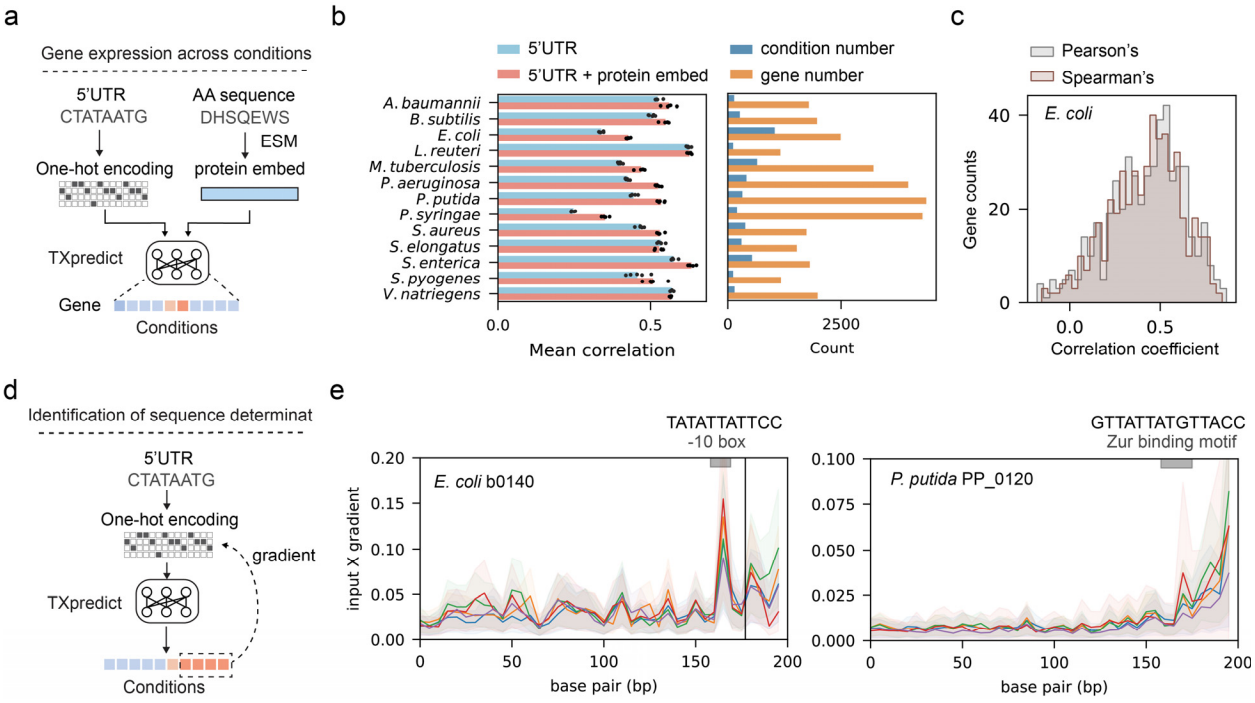
<u>Supplementary Figure 3:</u> Expression prediction for rare proteins. a) Histogram showing the distribution of overlap ratios between *B. hinzii* proteins and proteins in the training dataset. Overlap ratios were calculated as the aligned sequence length divided by the query length. **b)** Predicted 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).



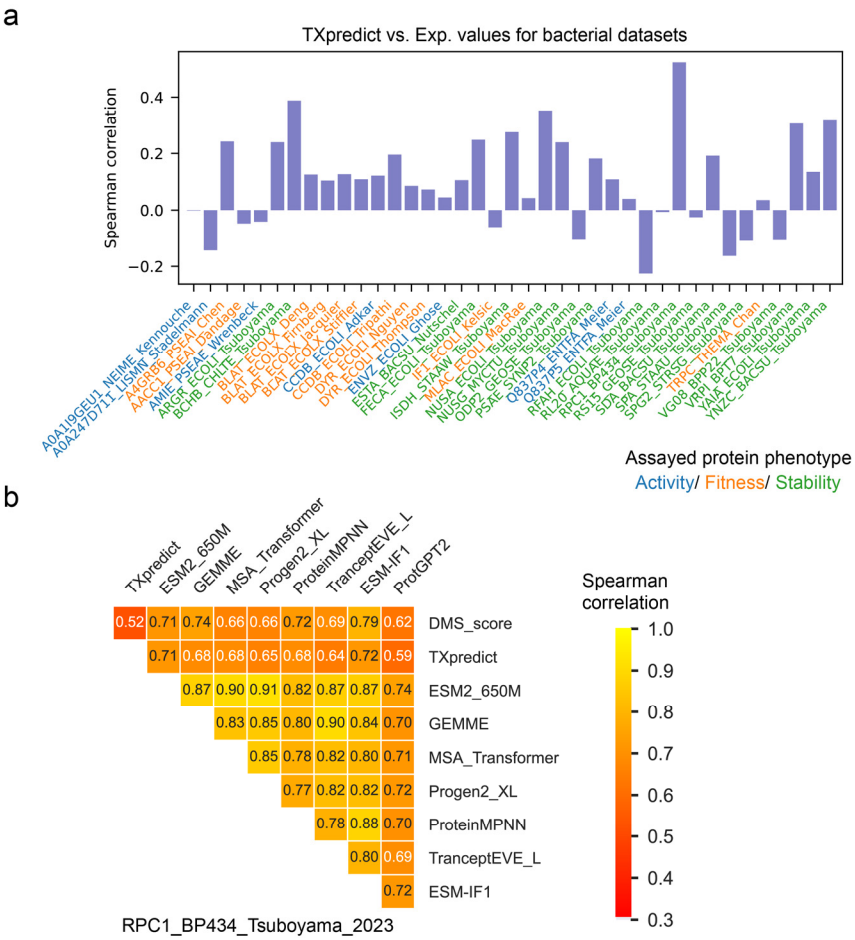
<u>Supplementary Figure 4:</u> Cross-domain prediction performance of TXpredict. TXpredict was trained on all species from a single domain and evaluated on species from the two other domains. Results from three domain-specific models are shown: bacteria (blue), archaea (orange), and fungi (green).



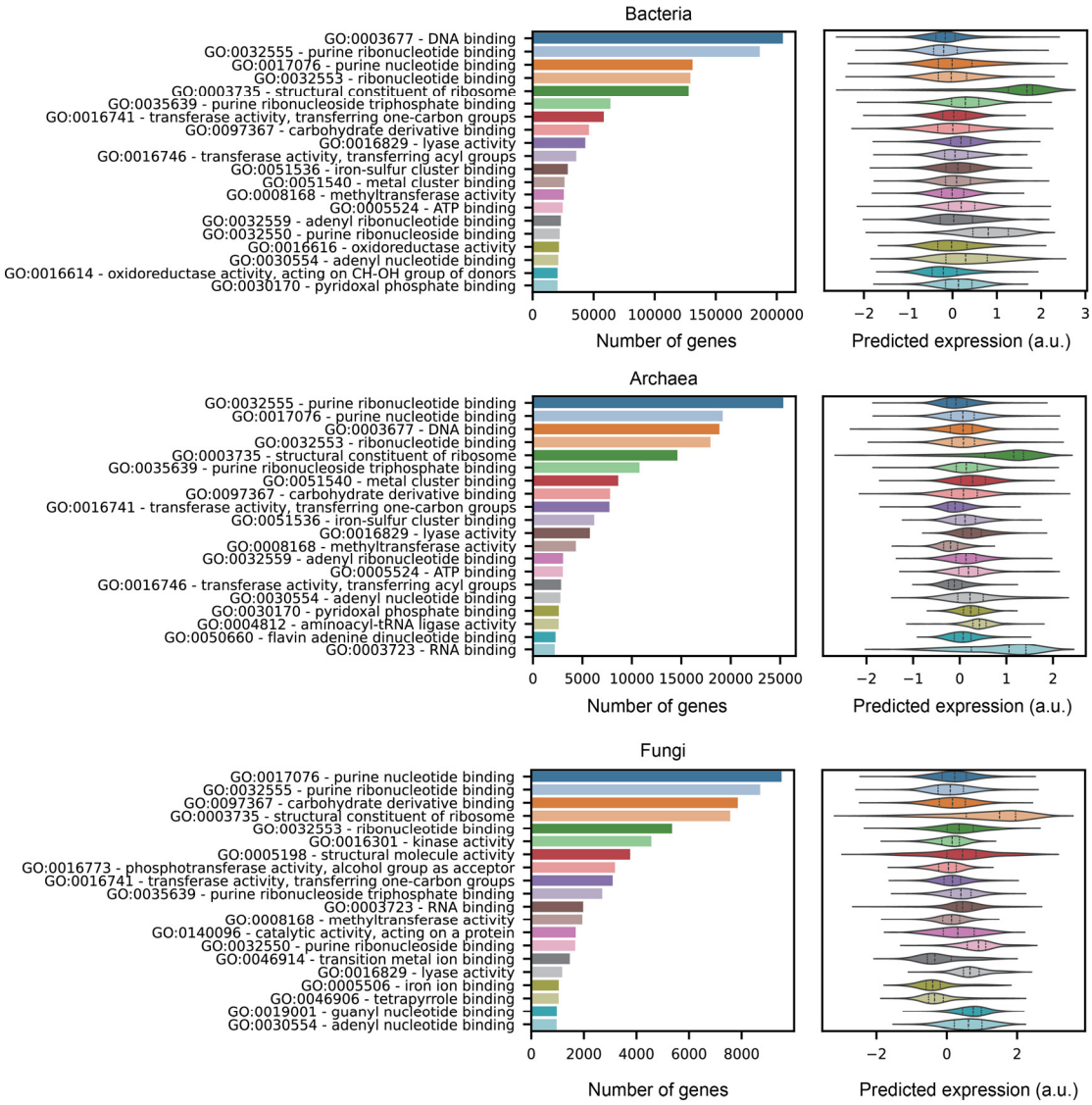
<u>Supplementary Figure 5:</u> Performance of TXpredict on *Pseudomonas sp.* WBC-3. a) Correlation of measured gene expression values of *Pseudomonas sp. WBC-3* across three experimental conditions. **b)** Predicted versus measured gene expression under each condition. **c)** Scatter plot comparing predicted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



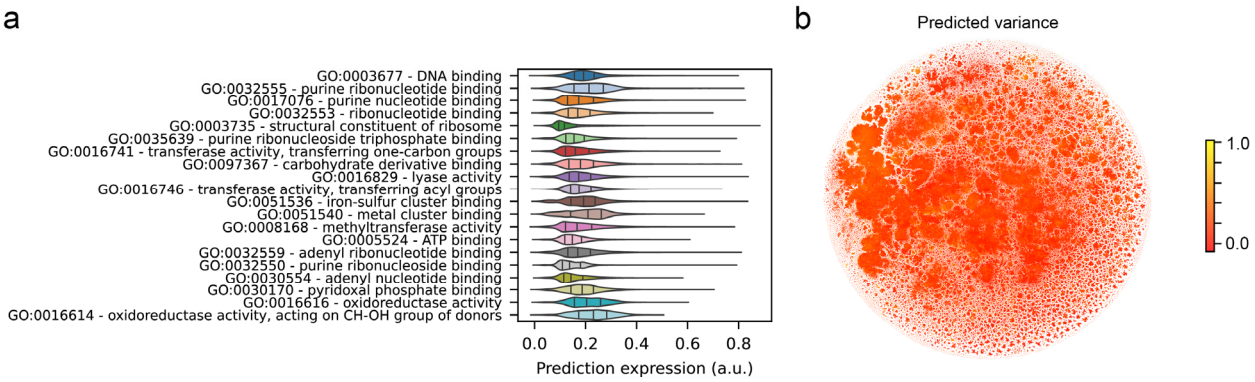
<u>Supplementary Figure 6:</u> Prediction of context-dependent gene expression. a) Schematic of TXpredict for condition-specific gene expression prediction. **b)** Mean correlations for models trained separately on different bacterial species. Correlations from 5-fold cross-validation tests are shown 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 gradients for identifying sequence determinants of condition-dependent gene expression. **e)** Two examples of input × gradients and their correlations with known sequence motifs.



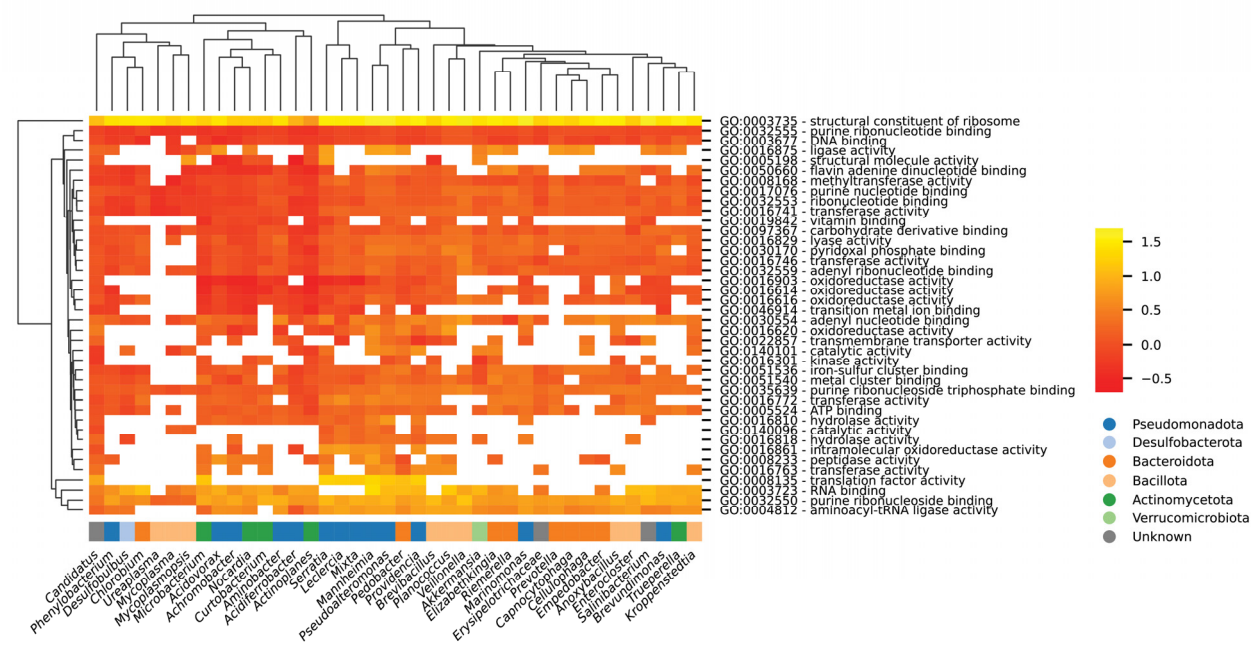
<u>Supplementary Figure 7:</u> Prediction of mutation effects in bacterial datasets. a**)** Correlation between TXpredict predictions and experimental DMS scores in bacterial datasets. **b)** Cross-correlations of TXpredict, zero-shot model predictions, and DMS scores for RPC1 mutants.



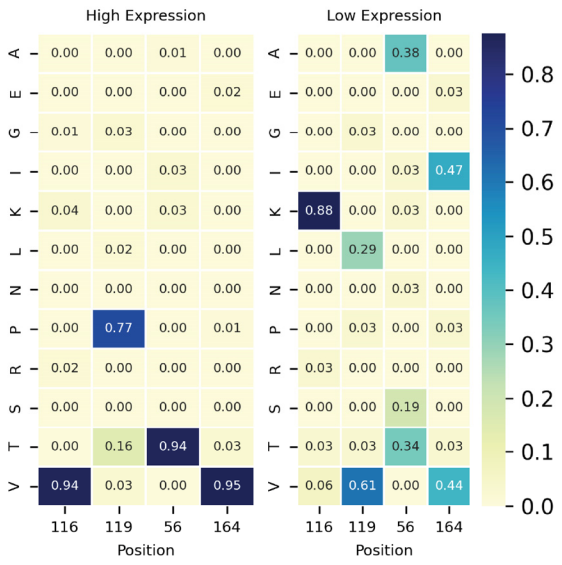
<u>Supplementary Figure 8:</u> Top GO terms and the predicted gene expressions in TXpredictDB. Top 20 Gene Ontology (GO) terms for genes from newly added bacterial, archaeal, and fungal genomes, as annotated by the DeepFRI model[39]. Predicted gene expression levels for these genes are shown alongside their associated GO terms.



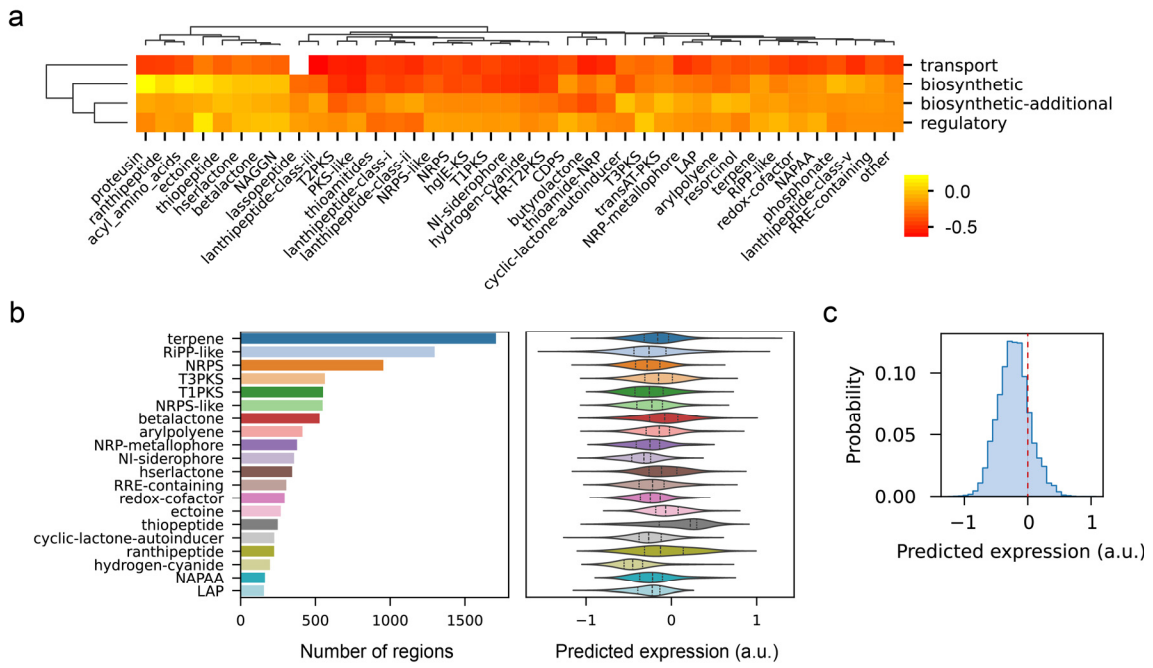
<u>Supplementary Figure 9:</u> Prediction of gene expression variation in newly added genomes. a) Predicted gene expression variation for genes associated with the top 20 GO terms in the newly added genomes. **b)** Predicted gene expression variation visualized in the t-SNE plot of protein embeddings (n = 9.8M).



615 **<u>Supplementary Figure 10:</u> Prediction of gene expression for understudied genera.** Gene expression
616 values for the top 40 GO terms and 40 genera are shown. Only genus–GO term combinations with more
617 than 10 associated genes are included, with mean gene expression levels presented. Genera are colored
618 by phylum.



<u>Supplementary Figure 11:</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 MMseqs2[42] at a 90% sequence identity threshold. Differences in amino acid frequencies between lowly and highly expressed genes within the cluster are shown.



<u>Supplementary Figure 12:</u> Predicted gene expression across biosynthetic gene clusters (BGCs). a) Predicted gene expression levels for core biosynthetic, additional biosynthetic, regulatory, and transport genes across different BGC types. **b)** Number of regions and predicted expression levels for the top 20 BGC types. **c)** Histogram of mean expression levels of BGCs (n = 11,400).