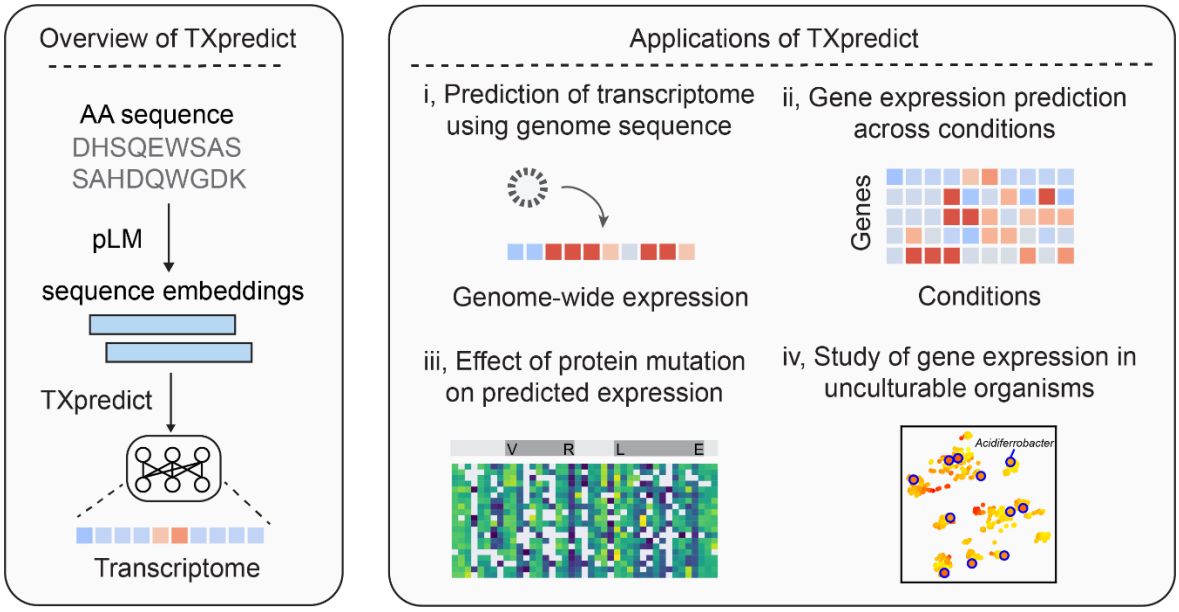
Predicting microbial transcriptome using annotated genome sequence

Guohao Fu[1, 6], Yujing Yan[2, 6], Ziyue Zhao[3], Ye Chen[4, *] and Bin Shao[5, 6, *]

[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] Qiushi School, Beijing Institute of Technology, Beijing 100081, P. R. China
[4] Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China
[5] Advanced Research Institute of Multidisciplinary Sciences, Beijing Institute of Technology, Beijing 100081, P. R. China
[6] These authors contribute equally to this work.

* To whom correspondence should be addressed. Email: chenye.h.ai@gmail.com, shaobinlx@gmail.com

GRAPHICAL ABSTRACT



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 uses the protein-coding sequences derived from the genome as the model's input, 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) as the ground-truth target, 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 expression profiles for 9.8 million genes. The models developed in this study are publicly available: https://github.com/lingxusb/TXpredict

MATERIAL AND METHODS

Training dataset

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

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

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

Processing of Nanopore sequencing datasets

We downloaded and processed nanopore sequencing datasets for three bacteria strains: *Pseudomonas aeruginosa* UCBPP-PA14 (PMID: 34780302)(26), *Haloferax volcanii* DS2 (PMID: 37192814)(27), and *Escherichia coli* str. K-12 substr. MG1655 (PMID: 39011882; 39453698; 34906997)(28–30). The analysis was performed with Miniconda3, minimap2 (v2.24) for read alignment, samtools (v1.16) for processing alignment files, and featureCounts (from the Subread package v2.0.6) for read counting. Sequence Read Archive (SRA) files for each sample (e.g., SRR27228856) were downloaded from the NCBI SRA database using the prefetch command. SRA files were converted to FASTQ format using fasterq-dump. Sequencing reads were aligned to the reference genome using minimap2 with the -ax map-ont parameter, optimized for Oxford Nanopore long-read data. The resulting SAM files were converted to BAM format, sorted, and indexed using samtools. Alignment quality was assessed using samtools flagstat to generate mapping statistics. Read counts for each gene were generated from the sorted BAM files using featureCounts with the gene annotation file (GTF) as a reference. The resulting count tables were then processed using custom awk

scripts to map gene identifiers to official gene names, producing a final gene expression matrix for each sample. To account for biases in gene length and sequencing depth, we normalized the gene counts as previously described. Genes not detected in any replicate were excluded from further analysis.

Transcriptome prediction model

We developed a Transformer-based model to predict gene expression levels from sequence features. The transformer encoder utilizes a multi-head self-attention mechanism (num_heads=4) with a single encoder layer. The encoder processes input features through a linear transformation into a hidden dimension of 128, followed by positional encoding. We implemented sinusoidal positional encoding to capture sequential information. The output of the transformer layer was then fed into two fully connected layers (dimension 512 and 128) with ReLU activations. The final linear layer produces a single output value representing the predicted gene expression level.

Model training and evaluation

We performed cross-species evaluations for bacteria, archaea and fungi separately 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 our model's 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 transcriptionally 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(31). 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 was 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(32), 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(33) 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 four nucleotide channels (A, T, C, G) to obtain single-base resolution importance scores. For visualization purposes, we aggregated these scores into 5-bp bins by summing the scores within each bin.

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

For the RNA-seq experiment, *Pseudomonas* sp. WBC-3 was cultured under three different conditions: PNP1 (minimal medium with 0.01% glycerol), PNP2 (minimal medium with 0.04% PNP), and PNP3 (minimal medium with 0.01% glycerol and 0.04% PNP). Each condition was prepared by diluting a 2× stock of yeast extract minimal medium supplemented with either 0.02% glycerol, 0.08% PNP, or both, resulting in the desired final concentrations. WBC-3

strain was first grown overnight in LB medium, then diluted 1:100 into 4 mL of each experimental medium and incubated overnight at 30°C with shaking at 200 rpm. Cultures were harvested at OD600 between 0.5 and 1.0 by centrifugation for 2 minutes, ensuring that the total collected biomass did not exceed 1×10^9 cells. Total RNA was extracted using the RNAprep Pure Bacteria Kit (TIANGEN, DP430), followed by DNase I treatment (TIANGEN, RT411) to remove genomic DNA contamination. RNA quality and concentration were confirmed to meet sequencing requirements. Raw sequencing reads were aligned and processed using standard RNA-seq pipelines. Read counts were quantified using StringTie with the following command: stringtie -p 8 -G GCF_035758525.1_ASM3575852v1_genomic.gtf -l sample_id -o sample.gtf sample.bam, using the *Pseudomonas* sp. WBC-3 reference genome (GCF_035758525.1).

RESULTS

Prediction of transcriptome based on the genome sequence

We developed two distinct models that address different aspects of gene expression prediction. The transcriptome model predicts each gene's mean expression across diverse experimental conditions, providing a baseline expression profile. The condition-dependent model incorporates experimental context to predict how individual genes respond to different conditions. For the transcriptome model, we first collected microbial RNA-seq datasets from the NCBI GEO datasets and selected the dataset with the highest number of samples for each microbial species. Read counts were normalized with the same pipeline (Methods). Taken together, our dataset includes 22 bacterial, 9 archaeal and 9 fungal species and a total of 24.8M gene expression measurements (Supplementary Table 1). For each microbial species, gene expression was averaged across all samples to construct a standard expression profile. We found that this standard expression profile is robust to differences in condition coverage, since mean expression profiles from independent GEO datasets of the same species are strongly correlated (Supplementary Fig. 1). Protein embeddings for each gene were computed using the ESM2-650M model and combined with basic protein statistics to serve as the input features for TXpredict (Fig. 1a, Methods). For model evaluation, we employed a leave-one-genome-out cross-validation strategy. For bacterial species, the

model was trained on genes from 21 species and tested on the held-out species (Fig. 1b). We followed the same approach for archaeal and fungal species.

We found our model achieved a mean Spearman correlation coefficient of 0.53 for the bacterial species (Fig. 1c), and performed well for non-model organisms, with correlation coefficients of 0.64, 0.62 and 0.62 for *B. hinzii*, *B. thetaiotaomicron*, and *C. beijerinckii*. It is worth noting that transcriptome predictions for these bacteria were based solely on the annotated genome sequences. The model's performance in predicting gene expression variation decreased modestly as the variation calculations depend on the diversity of experimental conditions (Supplementary Fig. 2). Despite having a limited training set for fungal species, our TXpredict model obtained a mean correlation of 0.62 and reached a Spearman correlation coefficient up to 0.70 for non-model organisms such as *O. oligospora* (Fig. 1c). For archaea, TXpredict showed a mean Spearman correlation of 0.44 (Fig. 1c) and demonstrated broad applicability to extremophiles from distinct natural niches (ranging from the Dead Sea to volcanic springs), further highlighting the robustness of our method.

The model performance for the test genome has a weak correlation with its evolutionary distance to the training genomes (Supplementary Fig. 3), while adding UTR features into model input didn't significantly improve model performance (Supplementary Fig. 4). Even though our model is trained on proteins shorter than 1500 amino acids, it interpolated well to proteins longer than 1500 AA, especially for yeast proteins (Supplementary Fig. 5). Lastly, we evaluated model performance within functional categories. Expression levels of test genes sharing the same GO annotation can still vary considerably. TXpredict captures this fine-scale variation, with within-group correlations only slightly lower than that for all genes. (Supplementary Fig. 6).

In our initial evaluation, TXpredict performed poorly on the archaeon *N. maritimus*, with a Spearman correlation of 0.23 in the leave-one-genome-out evaluation (Supplementary Fig. 7). Upon closer examination of this dataset, we found that fewer than 40% of the reads were successfully mapped to reference genome, demonstrating the potential of our model to detect potential problems in RNA-seq data.

To evaluate whether TXpredict merely memorizes training data or learns generalizable patterns, we examined whether protein sequence similarity affects model performance. For

bacterium *B. hinzii*, we used BLASTP(34) to compute sequence similarity between its proteins and those in the other bacterial strains. The model was trained on other species and tested on rare proteins from *B. hinzii* that were not present in the training data, showing a correlation of 0.46 for them (Supplementary Fig. 8). While this is lower than the correlation for all proteins (Fig. 1c), it suggests that the model captures expression-related features beyond sequence similarity. We also benchmarked the performance of our model with a BLASTP-based nearest-neighbor baseline. For each protein in the held-out test genome, we used BLASTP to identify the most similar sequence in the training set and transferred its expression value as the prediction. BLASTP identified close homologs (with significant matches) for >95% of test proteins. For these proteins, TXpredict achieves higher Spearman correlations than the nearest-neighbor baseline in 21 out of 22 bacterial genomes, 6 out of 9 archaeal genomes, and all fungal genomes. The mean Spearman correlation values for TXpredict are 0.53 (bacteria), 0.44 (archaea) and 0.62 (fungi), compared to the nearest-neighbor baseline values of 0.38, 0.37 and 0.39, respectively (Supplementary Fig. 9). For the small fraction of sequences lacking BLASTP matches to the training set, TXpredict maintains robust performance comparable to performance on sequences with homologs, suggesting that TXpredict learns meaningful biological relationships between sequence features and expression levels.

We next explored TXpredict's cross-taxa performance by training separate models on bacterial, archaeal, or fungal species and then testing them on other domains (Fig. 1d). Our model still showed moderate performance on cross-taxa predictions, indicating that they encoded general principles of gene expression regulation. Interestingly, the archaeal model achieved a higher correlation for eukaryotic species than on archaea. For cross-domain evaluation, the archaeal model was trained on the complete archaeal dataset before testing on other genomes. In contrast, for the archaeal leave-one-genome-out evaluation, the small number of total conditions means that removing a genome often removes unique experimental conditions, forcing the model to predict expression under unseen contexts and reducing generalization. At the species level, models trained on other taxa generally performed worse than those trained on the same taxon (Fig. 1e and Supplementary Fig. 10).

To validate the prediction of the TXpredict model, we conducted new RNA-seq experiments on *Pseudomonas sp*. WBC-3 and compared the predicted expression levels with the measured values. The Spearman correlation between measured and predicted expression was 0.40 for all proteins and 0.50 for proteins with GO annotations (Fig. 1f). Notably, our model showed high correlations (0.55) for certain GO categories including catalytic activity, and also for operons like the flagellar genes (Fig. 1g and Supplementary Fig. 11).

Robustness of the transcriptome prediction model

To evaluate the robustness of our model training process, we examined its performance against variations in the number of training species, the number of samples per species, and the sequencing platform. First, we conducted a systematic scaling analysis using bacterial datasets. For each held-out genome, we trained TXpredict on random subsets of 5, 10, and all 21 available species. The results show a modest increase in mean Spearman correlation with training species: 0.48 for 5 species, 0.50 for 10 species, and 0.53 when using all 21 training species (Supplementary Fig. 12). This scaling trend indicates that expanding the number of training species consistently improves cross-species prediction accuracy, although the magnitude of improvement diminishes with each additional species. Second, we examined whether the number of samples per species impacts performance by conducting a subsampling analysis. We randomly selected subsets of samples for each species, recalculated mean expression values, and performed leave-one-genome-out cross-validation. The results show that model performance remains stable across different sample sizes (Supplementary Fig. 13). Last, we processed and analyzed nanopore datasets including *E. coli, P. aeruginosa, and H. volcanii*, and compared these data with our Illumina-based datasets and model predictions. We observed a strong correlation in mean expression for the three strains across the two platforms (Supplementary Fig. 14). Our TXpredict model trained on Illumina data performs well for *E. coli* but poorer on *H. volcanii*, likely due to limited experimental diversity in the available nanopore dataset (n = 2).

To test whether a codon-aware approach could further improve predictions, we conducted a comparison between protein-based and RNA-based models. We used CodonTransformer(35), a transformer model that encodes DNA sequences while preserving

codon information. Specifically, we used CodonTransformer to generate embeddings from coding DNA sequences of bacterial genes, and we then benchmarked this transcriptome model using the same cross-validation tests as the ESM-2 model. We found protein-based models using ESM-2 embeddings consistently outperform codon-based models (Supplementary Fig. 15).

Modeling context-dependent gene expression

We next extended our framework to model condition-dependent gene expression (Fig, 2). We collected RNA-seq profiles from 13 bacterial strains across 4.6k unique conditions from iModulonDB, resulting in a total of 12.0M measurements (Supplementary Table 2). The model integrated one-hot encoded 5'UTR sequences with protein embeddings as input features. For each gene, we calculated the correlation between measured and predicted expression values across all conditions. Our TXpredict model achieved an average correlation coefficient of 0.52 across all bacterial strains. For *E. coli*, the distributions of Pearson's and Spearman's correlation coefficients show similar patterns, with 41% and 38% of genes having correlation coefficients above 0.5. The poor model performance for *P. syringae* could be due to lack of condition diversity in this dataset (Supplementary Fig. 16). Our subsampling analysis reveals a modest increase in model performance when more conditions are added (Supplementary Fig. 17). Moreover, we compared our model's performance with a baseline nearest-neighbor model. For this baseline, we identified each test gene's most similar sequence in the training set via BLASTP and used that gene's expression profile as the prediction. This baseline shows an average Spearman correlation of 0.20 between predicted and measured expression across all genomes, which is much lower than that of TXpredict (Supplementary Fig. 18).

To further interpret the model's predictions, we calculated input gradients(17) to identify sequence motifs contributing to gene expression variation across conditions (Fig. 2d and e). In *E. coli*, we observed a peak in the input × gradient overlap with the -10 box of the sigmaS-dependent promoter of the *yadV* gene(36). Also input gradients increased after the transcriptional start site, suggesting potential contributions from transcribed sequences. In a non-model organism *P. putida*, we identified a peak overlapping with a Zur binding motif for

the *PP_0120* gene, consistent with its role as a Zur iModulon gene(1)(1). These findings illustrate our model's ability to capture condition-specific gene expression changes in an interpretable manner.

Relationship between protein expression and mutation effects

Deep Mutational Scanning (DMS) assays have been widely used to map protein fitness landscapes by large-scale measurement of phenotypic effects of protein mutants(37, 38). We leveraged the comprehensive benchmarks provided by ProteinGym(39) to evaluate whether predicted expression changes for protein variants correlate with experimentally measured functional scores from DMS assays (Fig. 3a). We selected yeast protein datasets, as our fungal model achieved the highest accuracy in leave-one-genome-out evaluations. We found that the correlation between protein expression and mutational effects is generally weak and protein-specific. For specific proteins, such as THO1 and MYO3, there are strong correlations between predicted expression and experimentally measured mutant effects (Fig. 3b and c). We then compared TXpredict predictions against zero-shot predictions from several machine learning-based protein models(40–47) (Fig. 3d). TXpredict's performance was only moderately lower than two inverse-folding models (ESM-IF1(45) and ProteinMPNN(43)) and outperformed the ESM2 model, despite the fact that TXpredict utilizes ESM2 embeddings as input. TXpredict also showed weaker correlations with predictions from MSA transformer(41) and TranceptEVE(44). A detailed residue-level analysis for THO1 revealed that TXpredict's expression prediction successfully captures the negative functional impact of an arginine substitution at residue 19 within the SAP domain (Fig. 3e). Interestingly, for the yeast protein GAL4, expression predictions derived from partial mutant sequences correlated more strongly with protein mutation effects than predictions from the full-length sequences (Fig. 3f).

Application of TXpredict to new genomes

We applied the trained TXpredict model to predict transcriptomes for new genomes including 2,169 bacterial, 423 archaeal and 93 fungal species (Fig. 4a 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. 4 c and d). To explore the relationship between predicted gene expression and protein function, we used DeepFRI(32)(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. 4e, Supplementary Fig. 19). 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. 4f). When visualizing protein embeddings of all proteins, we observed a clear separation of bacterial, archaeal and fungal genes (Fig. 4g, Supplementary Fig. 20). 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. 21). There was a positive correlation between the predicted expression levels of tRNA ligases and ribosomal proteins across genera (Fig. 4h, Spearman correlation 0.52). *Actinoplanes* had the lowest predicted expression for ribosomal proteins, in accord with their known slow growth rate(48). The endosymbiont Candidatus *Blochmannia*, found in carpenter ants, showed low expression of both tRNA ligases and ribosomal proteins. This pattern likely reflects its reduced need for protein turnover and genome stasis(49), in contrast to free-living bacteria. We further visualized the embeddings of tRNA ligases to examine their sequence variations (Fig. 4i). tRNA ligases from Ca. *Blochmannia* 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 respective ecological niches..

We also clustered protein sequences using MMseqs2(33) 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. 4j). For example, in cluster of the PTS IIA component homologs, low-expression variants shared a V116K substitution (Supplementary Fig. 22), which could reduce translation efficiency and mRNA stability due to the presence of constitutive lysine residues(50, 51).

We further analyzed the predicted gene expression profiles of biosynthetic gene clusters (BGCs) across bacterial species in TXpredictDB, which were identified using antiSMASH(52)(45) (n = 11,400). Although BGCs are typically lowly expressed, distinct patterns were observed across different cluster types. For example, the proteusin BGC showed high expression of core biosynthetic genes, whereas the ectoine BGC exhibited elevated expression of regulatory genes (Supplementary Fig. 23). These reference expression profiles in non-model organisms may offer valuable insights into the functional organization of BGCs.

DISCUSSION

We present TXpredict, a transcriptome prediction model that can generalize to a wide range of microbes. For a microbial genome with 4.6k genes, transcriptome prediction can be completed in approximately 22 minutes using a Google Colab notebook (Material and Methods), demonstrating the model's efficiency and accessibility. Our approach is not without limitations: while it effectively leverages protein embeddings for expression prediction, it does not account for the effects of cis-regulatory elements. Moreover, our model’s performance is constrained by the limited training profiles, especially for archaeal species. Lastly, we provide a single reference transcriptome per genome in TXpredictDB, which does not capture gene expression dynamics. Our condition-dependent model captures context-dependent expression patterns for unseen genes, but it does not extrapolate to entirely novel experimental conditions not represented in the training set.

Our training data for TXpredict relies on next-generation sequencing (NGS) data from GEO datasets, as they currently provide the most comprehensive coverage of bacterial species and experimental conditions available. Nanopore-based direct RNA sequencing offers

several advantages, including full-length transcript detection and the avoidance of amplification and fragmentation artifacts inherent to short-read methodologies(28–30, 53–57). While our model captures platform-independent biological signals, we anticipate that incorporating additional nanopore datasets as they become available will further improve the model accuracy.

Despite these limitations, we propose several potential directions to further improve our model's performance. First, employing more powerful pre-trained language models could produce richer embeddings that better capture evolutionary constraints relevant to expression prediction. Moreover, integrating multi-modal sequence inputs (protein, DNA, and codon-aware representations) may provide complementary signals at different levels. Third, for data-sparse species, structure- and homology-aware augmentation or transfer learning method could potentially reduce model bias and improve its generalization. Finally, including more training species and broader condition coverage would further enhance performance.

TXpredict has the potential to complement experimental approaches and advance the study of the vast number of sequenced yet poorly characterized microbes. For example, identifying conserved expressed genes in pathogens may enable the discovery of novel drug targets(58). Furthermore, predicting gene expression across genomes can provide insight into how genome plasticity contributes to bacterial pathogenicity(59). Our model may also suggest biologically plausible mRNA expression levels as starting points for heterologous gene expression in synthetic biology applications. These efforts could further benefit from integrating single-cell RNA-seq data(5, 60, 61) and computational tools for mRNA synthesis and stability prediction(62–64), paving the way for a deeper understanding of microbial transcriptional landscapes.

DATA AVAILABILITY

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

REFERENCES

1. Lim,H.G., Rychel,K., Sastry,A. V., Bentley,G.J., Mueller,J., Schindel,H.S., Larsen,P.E., Laible,P.D., Guss,A.M., Niu,W., *et al.* (2022) Machine-learning from Pseudomonas putida KT2440 transcriptomes reveals its transcriptional regulatory network. *Metab Eng*, **72**, 297–310.
2. Patel,A., McGrosso,D., Hefner,Y., Campeau,A., Sastry,A. V., Maurya,S., Rychel,K., Gonzalez,D.J. and Palsson,B.O. (2024) Proteome allocation is linked to transcriptional regulation through a modularized transcriptome. *Nat Commun*, **15**, 5234.
3. Ryan,D., Bornet,E., Prezza,G., Alampalli,S.V., Franco de Carvalho,T., Felchle,H., Ebbecke,T., Hayward,R.J., Deutschbauer,A.M., Barquist,L., *et al.* (2024) An expanded transcriptome atlas for Bacteroides thetaiotaomicron reveals a small RNA that modulates tetracycline sensitivity. *Nat Microbiol*, **9**, 1130–1144.
4. Shin,J., Rychel,K. and Palsson,B.O. (2023) Systems biology of competency in Vibrio natriegens is revealed by applying novel data analytics to the transcriptome. *Cell Rep*, **42**, 112619.
5. Ma,P., Amemiya,H.M., He,L.L., Gandhi,S.J., Nicol,R., Bhattacharyya,R.P., Smillie,C.S. and Hung,D.T. (2023) Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. *Cell*, **186**, 877-891.e14.
6. Ryu,M.-H., Zhang,J., Toth,T., Khokhani,D., Geddes,B.A., Mus,F., Garcia-Costas,A., Peters,J.W., Poole,P.S., Ané,J.-M., *et al.* (2019) Control of nitrogen fixation in bacteria that associate with cereals. *Nat Microbiol*, **5**, 314–330.
7. Lloyd,K.G., Steen,A.D., Ladau,J., Yin,J. and Crosby,L. (2018) Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems*, **3**.
8. Barnum,T.P., Crits-Christoph,A., Molla,M., Carini,P., Lee,H.H. and Ostrov,N. (2024) Predicting microbial growth conditions from amino acid composition. 10.1101/2024.03.22.586313.
9. Ramoneda,J., Stallard-Olivera,E., Hoffert,M., Winfrey,C.C., Stadler,M., Niño-García,J.P. and Fierer,N. (2023) Building a genome-based understanding of bacterial pH preferences. *Sci Adv*, **9**.

10. Price,M.N., Deutschbauer,A.M. and Arkin,A.P. (2022) Filling gaps in bacterial catabolic pathways with computation and high-throughput genetics. *PLoS Genet*, **18**, e1010156.
11. Zrimec,J., Börlin,C.S., Buric,F., Muhammad,A.S., Chen,R., Siewers,V., Verendel,V., Nielsen,J., Töpel,M. and Zelezniak,A. (2020) Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun*, **11**, 6141.
12. Peleke,F.F., Zumkeller,S.M., Gültas,M., Schmitt,A. and Szymański,J. (2024) Deep learning the cis-regulatory code for gene expression in selected model plants. *Nat Commun*, **15**, 3488.
13. Dudnyk,K., Cai,D., Shi,C., Xu,J. and Zhou,J. (2024) Sequence basis of transcription initiation in the human genome. *Science (1979)*, **384**.
14. LaFleur,T.L., Hossain,A. and Salis,H.M. (2022) Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat Commun*, **13**, 5159.
15. Urtecho,G., Tripp,A.D., Insigne,K.D., Kim,H. and Kosuri,S. (2019) Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry*, **58**, 1539–1551.
16. Vaishnav,E.D., de Boer,C.G., Molinet,J., Yassour,M., Fan,L., Adiconis,X., Thompson,D.A., Levin,J.Z., Cubillos,F.A. and Regev,A. (2022) The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, **603**, 455–463.
17. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. and Kelley,D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*, **18**, 1196–1203.
18. Kelley,D.R., Reshef,Y.A., Bileschi,M., Belanger,D., McLean,C.Y. and Snoek,J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res*, **28**, 739–750.

19. Huang,C., Shuai,R.W., Baokar,P., Chung,R., Rastogi,R., Kathail,P. and Ioannidis,N.M. (2023) Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat Genet*, **55**, 2056–2059.
20. Sasse,A., Ng,B., Spiro,A.E., Tasaki,S., Bennett,D.A., Gaiteri,C., De Jager,P.L., Chikina,M. and Mostafavi,S. (2023) Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat Genet*, **55**, 2060–2064.
21. Li,G.-W., Burkhardt,D., Gross,C. and Weissman,J.S. (2014) Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell*, **157**, 624–635.
22. Lalanne,J.-B., Taggart,J.C., Guo,M.S., Herzel,L., Schieler,A. and Li,G.-W. (2018) Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell*, **173**, 749-761.e38.
23. Klumpp,S., Scott,M., Pedersen,S. and Hwa,T. (2013) Molecular crowding limits translation and cell growth. *Proceedings of the National Academy of Sciences*, **110**, 16754–16759.
24. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (1979)*, **379**, 1123–1130.
25. Clough,E. and Barrett,T. (2016) The Gene Expression Omnibus Database. In.pp. 93–110.
26. Pust,M.-M., Davenport,C.F., Wiehlmann,L. and Tümmler,B. (2022) Direct RNA Nanopore Sequencing of Pseudomonas aeruginosa Clone C Transcriptomes. *J Bacteriol*, **204**, e0041821.
27. Grünberger,F., Jüttner,M., Knüppel,R., Ferreira-Cerca,S. and Grohmann,D. (2023) Nanopore-based RNA sequencing deciphers the formation, processing, and modification steps of rRNA intermediates in archaea. *RNA*, **29**, 1255–1273.
28. Grünberger,F., Ferreira-Cerca,S. and Grohmann,D. (2022) Nanopore sequencing of RNA and cDNA molecules in Escherichia coli. *RNA*, **28**, 400–417.

29. Rodger,G., Lipworth,S., Barrett,L., Oakley,S., Crook,D.W., Eyre,D.W. and Stoesser,N. (2024) Comparison of direct cDNA and PCR-cDNA Nanopore sequencing of RNA from Escherichia coli isolates. *Microb Genom*, **10**.
30. Tan,L., Guo,Z., Shao,Y., Ye,L., Wang,M., Deng,X., Chen,S. and Li,R. (2024) Analysis of bacterial transcriptome and epitranscriptome using nanopore direct RNA sequencing. *Nucleic Acids Res*, **52**, 8746–8762.
31. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res*, **50**, D785–D794.
32. Gligorijević,V., Renfrew,P.D., Kosciolek,T., Leman,J.K., Berenberg,D., Vatanen,T., Chandler,C., Taylor,B.C., Fisk,I.M., Vlamakis,H., *et al.* (2021) Structure-based protein function prediction using graph convolutional networks. *Nat Commun*, **12**, 3168.
33. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, **35**, 1026–1028.
34. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
35. Fallahpour,A., Gureghian,V., Filion,G.J., Lindner,A.B. and Pandi,A. (2025) CodonTransformer: a multispecies codon optimizer using context-aware neural networks. *Nat Commun*, **16**, 3205.
36. Raina,S., Missiakas,D., Baird,L., Kumar,S. and Georgopoulos,C. (1993) Identification and transcriptional analysis of the Escherichia coli htrE operon which is homologous to pap and related pilin operons. *J Bacteriol*, **175**, 5009–5021.
37. Kitzman,J.O., Starita,L.M., Lo,R.S., Fields,S. and Shendure,J. (2015) Massively parallel single-amino-acid mutagenesis. *Nat Methods*, **12**, 203–206.
38. Tsuboyama,K., Dauparas,J., Chen,J., Laine,E., Mohseni Behbahani,Y., Weinstein,J.J., Mangan,N.M., Ovchinnikov,S. and Rocklin,G.J. (2023) Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, **620**, 434–444.

39. Notin,P., Kollasch,A., Ritter,D., van Niekerk,L., Paul,S., Spinner,H., Rollins,N., Shaw,A., Orenbuch,R., Weitzman,R., *et al.* (2023) ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In Oh,A., Naumann,T., Globerson,A., Saenko,K., Hardt,M., Levine,S. (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vol. 36, pp. 64331–64379.
40. Laine,E., Karami,Y. and Carbone,A. (2019) GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects. *Mol Biol Evol*, **36**, 2604–2619.
41. Rao,R.M., Liu,J., Verkuil,R., Meier,J., Canny,J., Abbeel,P., Sercu,T. and Rives,A. (2021) MSA Transformer. In Meila,M., Zhang,T. (eds), *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, Vol. 139, pp. 8844–8856.
42. Meier,J., Rao,R., Verkuil,R., Liu,J., Sercu,T. and Rives,A. (2021) Language models enable zero-shot prediction of the effects of mutations on protein function. In Ranzato,M., Beygelzimer,A., Dauphin,Y., Liang,P.S., Vaughan,J.W. (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vol. 34, pp. 29287–29303.
43. Dauparas,J., Anishchenko,I., Bennett,N., Bai,H., Ragotte,R.J., Milles,L.F., Wicky,B.I.M., Courbet,A., de Haas,R.J., Bethel,N., *et al.* (2022) Robust deep learning–based protein sequence design using ProteinMPNN. *Science (1979)*, **378**, 49–56.
44. Notin,P., Niekerk,L. Van, Kollasch,A.W., Ritter,D., Gal,Y. and Marks,D.S. (2022) TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein Sequences for Improved Fitness Prediction. *bioRxiv*.
45. Hsu,C., Verkuil,R., Liu,J., Lin,Z., Hie,B., Sercu,T., Lerer,A. and Rives,A. (2022) Learning inverse folding from millions of predicted structures. In Chaudhuri,K., Jegelka,S., Song,L., Szepesvari,C., Niu,G., Sabato,S. (eds), *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, Vol. 162, pp. 8946–8970.
46. Nijkamp,E., Ruffolo,J.A., Weinstein,E.N., Naik,N. and Madani,A. (2023) ProGen2: Exploring the boundaries of protein language models. *Cell Syst*, **14**, 968-978.e3.

47. Ferruz,N., Schmidt,S. and Höcker,B. (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun*, **13**, 4348.
48. J. N. Couch (1950) Actinoplanes, a new genus of the Actinomycetales. *Journal of the Elisha Mitchell Scientific Society*, **66**, 87–92.
49. Degnan,P.H., Lazarus,A.B. and Wernegreen,J.J. (2005) Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res*, **15**, 1023–1033.
50. Radhakrishnan,A. and Green,R. (2016) Connections Underlying Translation and mRNA Stability. *J Mol Biol*, **428**, 3558–3564.
51. Meydan,S. and Guydosh,N.R. (2020) Disome and Trisome Profiling Reveal Genome-wide Targets of Ribosome Quality Control. *Mol Cell*, **79**, 588-602.e6.
52. Blin,K., Shaw,S., Augustijn,H.E., Reitz,Z.L., Biermann,F., Alanjary,M., Fetter,A., Terlouw,B.R., Metcalf,W.W., Helfrich,E.J.N., *et al.* (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res*, **51**, W46–W50.
53. Garalde,D.R., Snell,E.A., Jachimowicz,D., Sipos,B., Lloyd,J.H., Bruce,M., Pantic,N., Admassu,T., James,P., Warland,A., *et al.* (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods*, **15**, 201–206.
54. Jenjaroenpun,P., Wongsurawat,T., Wadley,T.D., Wassenaar,T.M., Liu,J., Dai,Q., Wanchai,V., Akel,N.S., Jamshidi-Parsian,A., Franco,A.T., *et al.* (2021) Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res*, **49**, e7.
55. Mattick,J.S.A., Bromley,R.E., Watson,K.J., Adkins,R.S., Holt,C.I., Lebov,J.F., Sparklin,B.C., Tyson,T.S., Rasko,D.A. and Dunning Hotopp,J.C. (2024) Deciphering transcript architectural complexity in bacteria and archaea. *mBio*, **15**, e0235924.
56. Grünberger,F., Jüttner,M., Knüppel,R., Ferreira-Cerca,S. and Grohmann,D. (2023) Nanopore-based RNA sequencing deciphers the formation, processing, and modification steps of rRNA intermediates in archaea. *RNA*, **29**, 1255–1273.
57. Pust,M.-M., Davenport,C.F., Wiehlmann,L. and Tümmler,B. (2022) Direct RNA Nanopore Sequencing of Pseudomonas aeruginosa Clone C Transcriptomes. *J Bacteriol*, **204**, e0041821.

58. Wang,C., Zhang,H., Zhao,R., Tsui,C.K.-M. and Deng,S. (2025) Genomic Characterization and Antimicrobial Resistance of ESBL-Producing, Escherichia coli Isolates in Suzhou, China. *Infect Drug Resist*, **Volume 18**, 1049–1057.
59. Krishnakant Kushwaha,S., Wu,Y., Leonardo Avila,H., Anand,A., Sicheritz-Pontén,T., Millard,A., Amol Marathe,S. and Nobrega,F.L. (2024) Comprehensive blueprint of Salmonella genomic plasticity identifies hotspots for pathogenicity genes. *PLoS Biol*, **22**, e3002746.
60. Imdahl,F., Vafadarnejad,E., Homberger,C., Saliba,A.-E. and Vogel,J. (2020) Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nat Microbiol*, **5**, 1202–1206.
61. Kuchina,A., Brettner,L.M., Paleologu,L., Roco,C.M., Rosenberg,A.B., Carignano,A., Kibler,R., Hirano,M., DePaolo,R.W. and Seelig,G. (2021) Microbial single-cell RNA sequencing by split-pool barcoding. *Science (1979)*, **371**.
62. LaFleur,T.L., Hossain,A. and Salis,H.M. (2022) Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat Commun*, **13**, 5159.
63. Cetnar,D.P., Hossain,A., Vezeau,G.E. and Salis,H.M. (2024) Predicting synthetic mRNA stability using massively parallel kinetic measurements, biophysical modeling, and machine learning. *Nat Commun*, **15**, 9601.
64. Xia,Y., Sun,J., Du,X., Liang,Z., Shi,W., Guo,S. and Huo,Y.-X. (2024) Language model generates *cis*- regulatory elements across prokaryotes. 10.1101/2024.11.07.622410.
65. Jeckel,H., Nosho,K., Neuhaus,K., Hastewell,A.D., Skinner,D.J., Saha,D., Netter,N., Paczia,N., Dunkel,J. and Drescher,K. (2023) Simultaneous spatiotemporal transcriptomics and microscopy of Bacillus subtilis swarm development reveal cooperation across generations. *Nat Microbiol*, **8**, 2378–2391.
66. Nilson,R., Penumutchu,S., Pagano,F.S. and Belenky,P. (2024) Metabolic changes associated with polysaccharide utilization reduce susceptibility to some β-lactams in *Bacteroides thetaiotaomicron*. *mSphere*, **9**.

67. Kamat,A., Tran,N.T., Sharda,M., Sontakke,N., Le,T.B.K. and Badrinarayanan,A. (2024) Widespread prevalence of a methylation-dependent switch to activate an essential DNA damage response in bacteria. *PLoS Biol*, **22**, e3002540.
68. Lakey,B.D., Myers,K.S., Alberge,F., Mettert,E.L., Kiley,P.J., Noguera,D.R. and Donohue,T.J. (2022) The essential Rhodobacter sphaeroides CenKR two-component system regulates cell division and envelope biosynthesis. *PLoS Genet*, **18**, e1010270.
69. Pensinger,D.A., Dobrila,H.A., Stevenson,D.M., Hryckowian,N.D., Amador-Noguez,D. and Hryckowian,A.J. (2024) Exogenous butyrate inhibits butyrogenic metabolism and alters virulence phenotypes in *Clostridioides difficile*. *mBio*, **15**.
70. Servinsky,M.D., Renberg,R.L., Perisin,M.A., Gerlach,E.S., Liu,S. and Sund,C.J. (2018) Arabinose-Induced Catabolite Repression as a Mechanism for Pentose Hierarchy Control in *Clostridium acetobutylicum* ATCC 824. *mSystems*, **3**.
71. Rahimi,S., Modin,O., Roshanzamir,F., Neissi,A., Saheb Alam,S., Seelbinder,B., Pandit,S., Shi,L. and Mijakovic,I. (2020) Co-culturing Bacillus subtilis and wastewater microbial community in a bio-electrochemical system enhances denitrification and butyrate formation. *Chemical Engineering Journal*, **397**, 125437.
72. Pennings,J.L.A., Abachin,E., Esson,R., Hodemaekers,H., Francotte,A., Claude,J.-B., Vanhee,C., Uhlrich,S. and Vandebriel,R.J. (2022) Regulation of Clostridium tetani Neurotoxin Expression by Culture Conditions. *Toxins (Basel)*, **14**, 31.
73. Espah Borujeni,A., Zhang,J., Doosthosseini,H., Nielsen,A.A.K. and Voigt,C.A. (2020) Genetic circuit characterization by inferring RNA polymerase movement and ribosome usage. *Nat Commun*, **11**, 5001.
74. Wilson,E.H., Groom,J.D., Sarfatis,M.C., Ford,S.M., Lidstrom,M.E. and Beck,D.A.C. (2021) A Computational Framework for Identifying Promoter Sequences in Nonmodel Organisms Using RNA-seq Data Sets. *ACS Synth Biol*, **10**, 1394–1405.
75. Young,A.T., Carette,X., Helmel,M., Steen,H., Husson,R.N., Quackenbush,J. and Platig,J. (2021) Multi-omic regulatory networks capture downstream effects of kinase inhibition in Mycobacterium tuberculosis. *NPJ Syst Biol Appl*, **7**, 8.

76. Carette,X., Platig,J., Young,D.C., Helmel,M., Young,A.T., Wang,Z., Potluri,L.-P., Moody,C.S., Zeng,J., Prisic,S., *et al.* (2018) Multisystem Analysis of *Mycobacterium tuberculosis* Reveals Kinase-Dependent Remodeling of the Pathogen-Environment Interface. *mBio*, **9**.
77. Wu,M.-L., Gengenbacher,M., Chung,J.C.S., Chen,S.L., Mollenkopf,H.-J., Kaufmann,S.H.E. and Dick,T. (2016) Developmental transcriptome of resting cell formation in Mycobacterium smegmatis. *BMC Genomics*, **17**, 837.
78. Nobori,T., Velásquez,A.C., Wu,J., Kvitko,B.H., Kremer,J.M., Wang,Y., He,S.Y. and Tsuda,K. (2018) Transcriptome landscape of a bacterial pathogen under plant immunity. *Proceedings of the National Academy of Sciences*, **115**.
79. Kachroo,P., Eraso,J.M., Beres,S.B., Olsen,R.J., Zhu,L., Nasser,W., Bernard,P.E., Cantu,C.C., Saavedra,M.O., Arredondo,M.J., *et al.* (2019) Integrated analysis of population genomics, transcriptomics and virulence provides novel insights into Streptococcus pyogenes pathogenesis. *Nat Genet*, **51**, 548–559.
80. Martim,D.B., Brilhante,A.J.V.C., Lima,A.R., Paixão,D.A.A., Martins-Junior,J., Kashiwagi,F.M., Wolf,L.D., Costa,M.S., Menezes,F.F., Prata,R., *et al.* (2024) Resolving the metabolism of monolignols and other lignin-related aromatic compounds in Xanthomonas citri. *Nat Commun*, **15**, 7994.
81. Vera,J.M., Ghosh,I.N., Zhang,Y., Hebert,A.S., Coon,J.J. and Landick,R. (2020) Genome-Scale Transcription-Translation Mapping Reveals Features of Zymomonas mobilis Transcription Units and Promoters. *mSystems*, **5**.
82. Sakrikar,S. and Schmid,A.K. (2021) An archaeal histone-like protein regulates gene expression in response to salt stress. *Nucleic Acids Res*, **49**, 12732–12743.
83. Hackley,R.K., Hwang,S., Herb,J.T., Bhanap,P., Lam,K., Vreugdenhil,A., Darnell,C.L., Pastor,M.M., Martin,J.H., Maupin‐Furlow,J.A., *et al.* (2024) <scp>TbsP</scp> and <scp>TrmB</scp> jointly regulate *gapII* to influence cell development phenotypes in the archaeon *Haloferax volcanii*. *Mol Microbiol*, **121**, 742–766.
84. Topçuoğlu,B.D., Meydan,C., Nguyen,T.B., Lang,S.Q. and Holden,J.F. (2019) Growth Kinetics, Carbon Isotope Fractionation, and Gene Expression in the Hyperthermophile *Methanocaldococcus jannaschii* during Hydrogen-Limited Growth and Interspecies Hydrogen Transfer. *Appl Environ Microbiol*, **85**.

85. Payne,D., Keller,L.M., Larson,J., Bothner,B., Colman,D.R. and Boyd,E.S. (2024) Alternative sources of molybdenum for Methanococcus maripaludis and their implication for the evolution of molybdoenzymes. *Commun Biol*, **7**, 1337.
86. Spietz,R.L., Payne,D., Kulkarni,G., Metcalf,W.W., Roden,E.E. and Boyd,E.S. (2022) Investigating Abiotic and Biotic Mechanisms of Pyrite Reduction. *Front Microbiol*, **13**.
87. Yang,Y., Liu,J., Fu,X., Zhou,F., Zhang,S., Zhang,X., Huang,Q., Krupovic,M., She,Q., Ni,J., *et al.* (2023) A novel RHH family transcription factor aCcr1 and its viral homologs dictate cell cycle progression in archaea. *Nucleic Acids Res*, **51**, 1707–1723.
88. Topçuoğlu,B.D., Meydan,C., Orellana,R. and Holden,J.F. (2018) Formate hydrogenlyase and formate secretion ameliorate H_2 inhibition in the hyperthermophilic archaeon *Thermococcus paralvinellae*. *Environ Microbiol*, **20**, 949–957.
89. Demers,E.G., Stajich,J.E., Ashare,A., Occhipinti,P. and Hogan,D.A. (2021) Balancing Positive and Negative Selection: *In Vivo* Evolution of Candida lusitaniae *MRR1*. *mBio*, **12**.
90. Kang,Y.S., Jung,J., Brown,H.L., Mateusiak,C., Doering,T.L. and Brent,M.R. (2025) Leveraging a new data resource to define the response of *Cryptococcus neoformans* to environmental signals. *Genetics*, **229**.
91. Schnepf,V., Vlot,A.C., Kugler,K. and Hückelhoven,R. (2018) Barley susceptibility factor RACB modulates transcript levels of signalling protein genes in compatible interaction with *Blumeria graminis* f.sp. *hordei*. *Mol Plant Pathol*, **19**, 393–404.
92. Hunsaker,E.W., Yu,C.-H.A. and Franz,K.J. (2021) Copper Availability Influences the Transcriptomic Response of *Candida albicans* to Fluconazole Stress. *G3 Genes|Genomes|Genetics*, **11**.
93. Menon,N.D., Poudel,S., Sastry,A. V., Rychel,K., Szubin,R., Dillon,N., Tsunemoto,H., Hirose,Y., Nair,B.G., Kumar,G.B., *et al.* (2024) Independent component analysis reveals 49 independently modulated gene sets within the global transcriptional

regulatory architecture of multidrug-resistant *Acinetobacter baumannii*. *mSystems*, **9**.

94. Lamoureux,C.R., Decker,K.T., Sastry,A. V, Rychel,K., Gao,Y., McConn,J.L., Zielinski,D.C. and Palsson,B.O. (2023) A multi-scale expression and regulation knowledge base for *Escherichia coli*. *Nucleic Acids Res*, **51**, 10176–10193.
95. Josephs-Spaulding,J., Rajput,A., Hefner,Y., Szubin,R., Balasubramanian,A., Li,G., Zielinski,D.C., Jahn,L., Sommer,M., Phaneuf,P., *et al.* (2024) Reconstructing the transcriptional regulatory network of probiotic *L. reuteri* is enabled by transcriptomics and machine learning. *mSystems*, **9**.
96. Yoo,R., Rychel,K., Poudel,S., Al-bulushi,T., Yuan,Y., Chauhan,S., Lamoureux,C., Palsson,B.O. and Sastry,A. (2022) Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection. *mSphere*, **7**.
97. Rajput,A., Tsunemoto,H., Sastry,A. V, Szubin,R., Rychel,K., Chauhan,S.M., Pogliano,J. and Palsson,B.O. (2022) Advanced transcriptomic analysis reveals the role of efflux pumps and media composition in antibiotic responses of *Pseudomonas aeruginosa*. *Nucleic Acids Res*, **50**, 9675–9688.
98. Bajpe,H., Rychel,K., Lamoureux,C.R., Sastry,A. V. and Palsson,B.O. (2023) Machine learning uncovers the *Pseudomonas syringae* transcriptome in microbial communities and during infection. *mSystems*, **8**.
99. Poudel,S., Hefner,Y., Szubin,R., Sastry,A., Gao,Y., Nizet,V. and Palsson,B.O. (2022) Coordination of CcpA and CodY Regulators in Staphylococcus aureus USA300 Strains. *mSystems*, **7**.
100. Yuan,Y., Al Bulushi,T., Sastry,A. V., Sancar,C., Szubin,R., Golden,S.S. and Palsson,B.O. (2024) Machine learning reveals the transcriptional regulatory network and circadian dynamics of *Synechococcus elongatus* PCC 7942. *Proceedings of the National Academy of Sciences*, **121**.
101. Hirose,Y., Poudel,S., Sastry,A. V., Rychel,K., Lamoureux,C.R., Szubin,R., Zielinski,D.C., Lim,H.G., Menon,N.D., Bergsten,H., *et al.* (2023) Elucidation of independently modulated genes in *Streptococcus pyogenes* reveals carbon sources that control its expression of hemolytic toxins. *mSystems*, **8**.

FIGURES

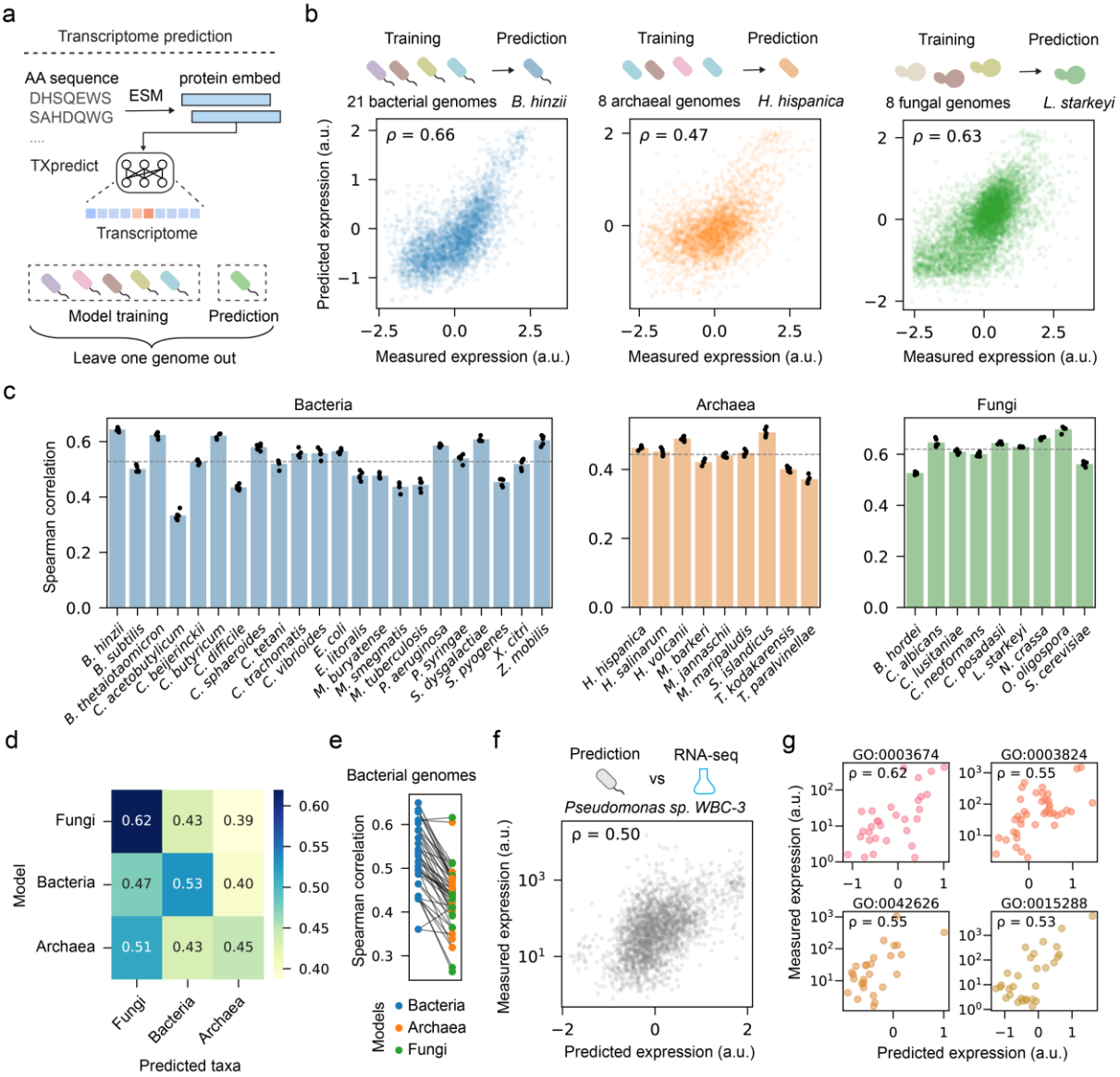


Figure 1. Prediction of transcriptome based on the genome sequence. **a)** Schematic of TXpredict for transcriptome prediction. **b)** Scatter plots showing predicted and measured gene expression in leave-one-genome-out evaluations for one species from bacteria, archaea, and fungi. Spearman correlation 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 indicate the mean correlations for each domain. **d)** Cross-domain prediction performance of TXpredict models. Models were trained on all

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

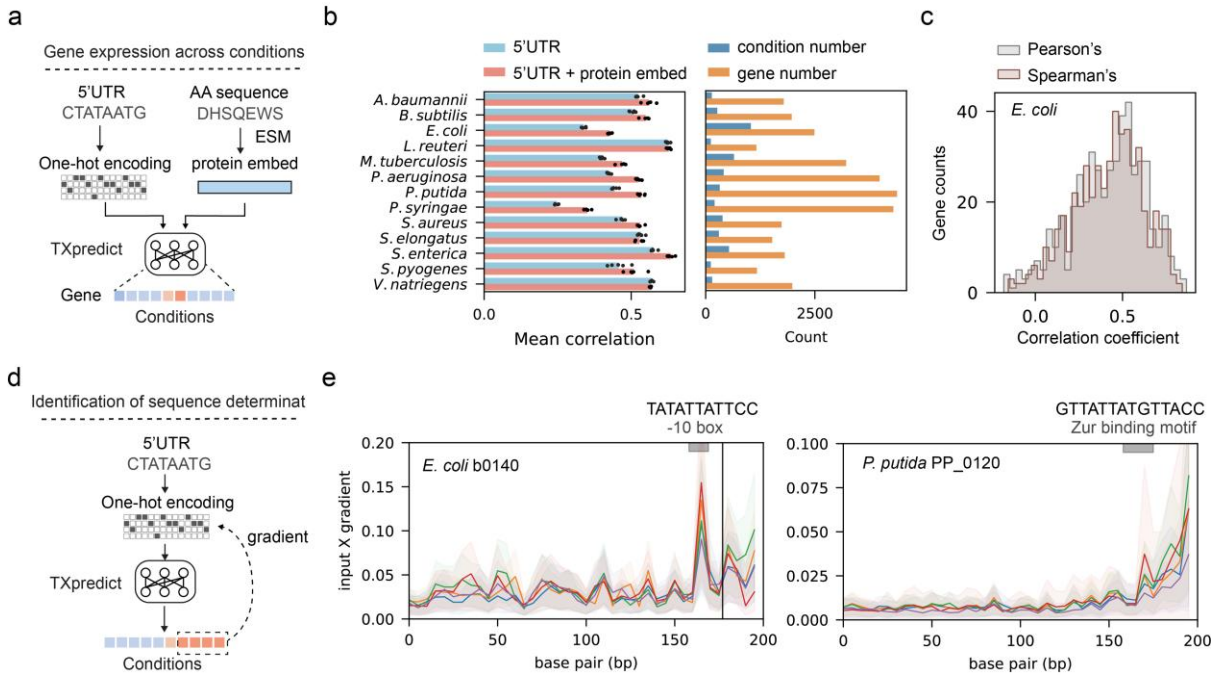


Figure 2. Prediction of context-dependent gene expression. **a)** Schematic of TXpredict for condition-specific gene expression prediction. **b)** Mean correlations for models trained separately on different bacterial species. Correlations from 5-fold cross-validation tests are shown on the left (n = 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.

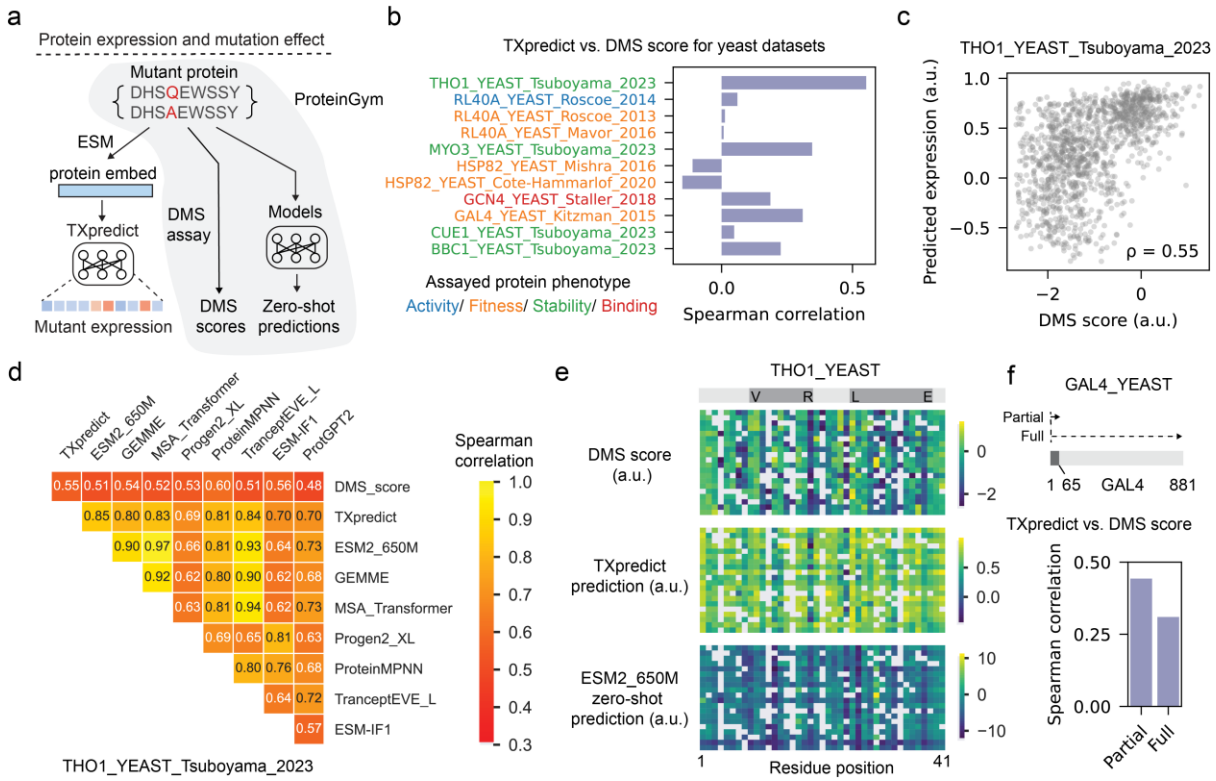


Figure 3. Relationship between protein expression and mutation effects. **a)** Schematic showing relationships among TXpredict predictions, experimentally measured deep mutational scanning (DMS) scores, and zero-shot predictions from ProteinGym30. **b)** Spearman correlation between TXpredict predictions 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.

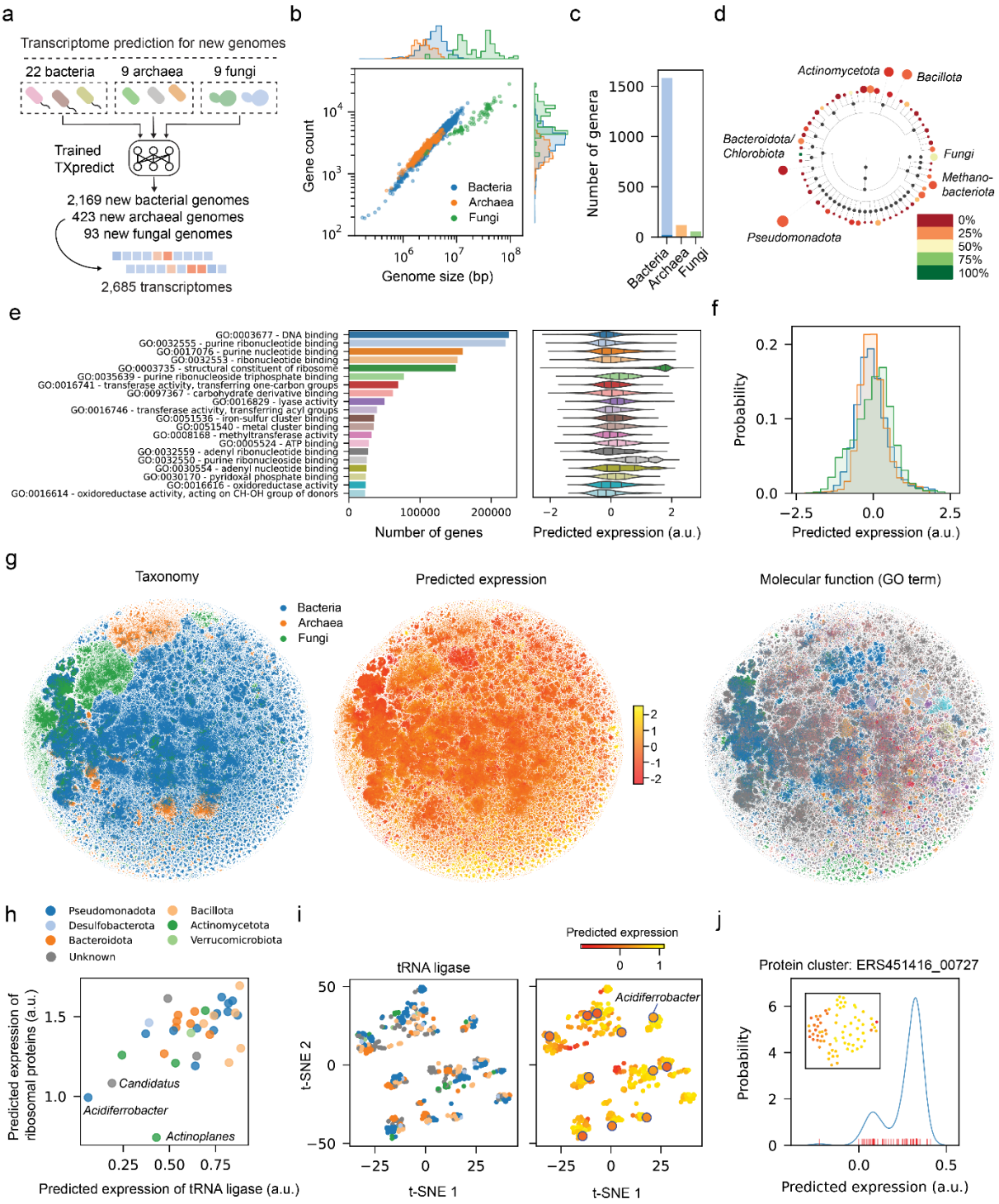


Figure 4. Application of TXpredict to new genomes. **a)** Transcriptome prediction for new genomes. **b)** Genome sizes and gene counts for the new genomes (n = 2,685). **c)** Number of genera in the new genomes (light colors) compared to training genomes (dark colors). **d)** Polygenetic tree of newly added genomes. Each dot represents a phylum. The dot size is proportional to the number of genera it contains and color indicates the proportion of genera with existing RNA-seq data based on the NCBI GEO database. **e)** Top 20 GO terms of

the genes in the new genomes calculated by the DeepFRI model39 (left) and their predicted gene expression levels (right). **f)** Distributions of predicted gene expression levels across new bacterial (blue), archaeal (orange) and fungal (green) genomes. **g)** t-SNE plots of protein embeddings generated by the ESM2 model24 (n = 9.8M). Each dot represents a gene in the new genomes, colored by taxa (left), predicted gene expression (middle), and GO terms (right). GO terms are colored as in panel (**e**). **h)** Predicted expression of tRNA ligases and ribosomal proteins across understudied genera (n = 40), colored by phylum. **i)** t-SNE plots of tRNA ligase embeddings across understudied genera, colored by phylum (left) and predicted expression level (right). **j)** Histogram of predicted expression levels for proteins in a high-similarity cluster. Inset shows the t-SNE plot of embeddings for these proteins, colored by predicted expression.