# Predicting bacterial growth conditions from mRNA and protein abundances

Mehmet U. Caglar1, Adam J. Hockenberry1, Claus O. Wilke1\*

1Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, USA

\*Corresponding author: [wilke@austin.utexas.edu](mailto:wilke@austin.utexas.edu)

**Abstract**

Cells respond to changing nutrient availability and external stresses by altering the expression of individual genes. Condition-specific gene expression patterns may provide a promising and low-cost route to quantifying the presence of various small molecules, toxins, or species-interactions in natural environments. However, whether gene expression signatures alone can predict individual environmental growth conditions remains an open question. Here, we used machine learning to predict 16 closely-related growth conditions using 155 datasets of *E. coli* transcript and protein abundances. We show that models are able to discriminate between different environmental features with a relatively high degree of accuracy. We observed a small but significant increase in model accuracy by combining transcriptome and proteome-level data, and we show that stationary phase conditions are typically more difficult to distinguish from one another than conditions under exponential growth. Nevertheless, with sufficient training data, gene expression measurements from a single species are capable of distinguishing between environmental conditions that are separated by a single environmental variable.

**Introduction**

Generalist bacterial species must be able to sense outside conditions and alter gene expression patterns according to nutrient availabilities and various environmental features [1]. For instance, *Escherichia coli* grows inside the comparatively warm, nutrient rich digestive tract of host [2] organisms but spends another portion of its life-cycle exposed to harsh environmental conditions upon being excreted and before finding another host. Maintenance and reproduction in these vastly different conditions is accomplished by expressing a controlled set of environmentally specific gene products in addition to more ubiquitous “housekeeping” genes [3].

On top of their native responses, microbial cells can be further engineered to act as sensors for a variety of environmental conditions via rational design of synthetic genetic circuits that may, for instance, cause the cells to fluoresce upon sensing of a particular small molecule [4]. However, individual synthetic biology applications take time to develop and there is still a concern about releasing genetically engineered species into natural environments where they may act as low-cost sensors for pollutants or various environmental phenomena of interest [5].

To partially alleviate this concern, previous work has shown that the species composition of an environment can serve as a rapid and low-cost biosensor to indicate the presence of various contaminants according to the gene repertoire identified via meta-genomics [6]. However, looking at the species composition alone fails to account for the fact that gene *expression* patterns of individual species may be able to provide even higher resolution into the past and current chemical composition of environments. What molecular level data is important, and at what resolution individual environmental differences can be resolved from the gene expression patterns of individual generalist species remains unknown.

Combining different ‘omics’-scale technologies is likely to provide better discriminatory capability compared with only monitoring mRNA abundances, for instance, but integrating datasets is challenging based on the biases of individual methods [7] and the inevitability of batch-level effects that occur when datasets are generated across multiple labs and platforms [8,9] . These problems are further exacerbated when considering the ultimate goal of detecting different environmental conditions *in situ*.

Previous work has looked into the question of predicting external conditions by using the cells’ internal variables [10,11]. Other studies have interrogated multi-omic datasets from different growth conditions to understand the function of regulatory networks, individual gene functions, and resource allocation strategies [3,12]. However, the main focus of many of these studies has been to understand differences in gene expression patterns across environmental conditions so as to provide insight into *internal* cellular mechanisms and pathways or to predict cellular level phenotypes such as specific growth rates. By contrast, few studies have focused on using the internal state of cells to predict external environmental conditions across a range of partially-overlapping conditions and cellular growth rates.

Here, we are interested in determining whether gene expression patterns can discriminate between environmental conditions in the absence of prior knowledge about the role and function of individual genes. Our study leverages a large dataset of transcriptomic and proteomic measurements of *E.coli* growth under multiple distinct but closely-related conditions [13]. We use mRNA and protein composition data to train machine learning models and find that highly similar environmental conditions can be discriminated with a relatively high degree of accuracy. We also investigate which conditions are more- and less-challenging to discriminate and find that prediction accuracies decrease substantially for stationary phase cells, indicating the importance of cellular growth for discriminating between conditions. Finally, we note that our accuracy remains limited by training set size such that our findings present a lower bound on the predictive power that is achievable given a greater availability of training data.

**Results**

**Data structure and pipeline design**

We used a previously generated dataset of whole-genome *E. coli* mRNA and protein abundances, measured under 34 different conditions [13,14]. This dataset consists of a total of 155 samples, for which mRNA abundances are available for 152 and protein abundances for 105 (Figure 1). For 102 samples, both mRNA and protein abundances are available. The 34 different experimental conditions were generated by systematically varying four parameters. Here we further simplified the experimental conditions into a total of 16, by grouping similar conditions together (e.g., 100, 200, and 300mm Na+ were all labelled as “high Na”). For the remainder of this manuscript, we use the term “growth condition” to specifically mean the four-dimensional vector of categorical variables defining growth phase (exponential, stationary, late stationary), ii) carbon source (glucose, glycerol, gluconate, lactate), Mg2+ concentration (low, base, high), and Na+ concentration (base, high). The question we set out to answer is to what extent machine learning models are capable of discriminating between these growth parameters given only knowledge of gene expression levels, provided as mRNA abundances, protein abundances, or both.

We applied a general cross-validation strategy and first split samples into training and test datasets. We next used the training data to fit supervised models to the gene expression data to maximize correct predictions of the labeled environmental conditions. At the training stage, we employed parameter tuning, which required a further subdivision of the training data to identify the optimal tuning parameters. Finally, we use the trained and tuned models to predict test set data and report prediction accuracy. To assess robustness of our results to the choice of training and test data, we repeated this procedure 60 times. Our pipeline is illustrated in Figure 2 and described in more detail in Materials and Methods.

**Growth conditions can be predicted accurately from both mRNA and protein abundances**

After constructing our analysis pipeline, we first asked whether there were major differences in the performance of different machine learning approaches. We tested four different machine learning models, three based on Support Vector Machines (SVMs) with different kernels (radial, sigmoidal, and linear) and the fourth using random forest classification. We trained models to predict [3,15] the entire four-dimensional condition vector at once for a given sample, and we used the multi-conditional *F*1 score [16] to quantify prediction accuracy. The *F*1 score is the harmonic mean of precision and recall. It approaches zero if either quantity approaches zero, and it approaches one if both quantities approach one (representing perfect prediction accuracy). We note that this score is highly conservative as it will classify a prediction as incorrect if a single variable is wrongfully predicted, even if the predictions for the remaining three variables are correct. We assessed model performance during the tuning stage of our pipeline by recording which model had the best *F*1 score for each tuning run (Supplementary Figures 1 and 2). At the tuning stage, we found that the SVM model with a radial kernel clearly outcompeted the other models when fit to mRNA data, and the random forest model outcompeted the other models when fit to protein data (Table 1).

We next compared the *F*1 scores for model predictions applied to the test set. When using mRNA abundance data alone, the distribution of *F*1 scores from our 60 independent replications werecentered around a value of 0.7 (Figure 3). The *F*1 score distributions were virtually identical for the three SVM models and were somewhat lower for the random forest model. Model performance on test data using only protein abundance measurements was slightly worse than those achieved with mRNA abundance data. However, it is important to note that the protein abundance data contains fewer conditions overall, which may partially explain the decreased predictive accuracy of the protein-only model—a point to which we return to later.

In addition to assessing the overall predictive power using *F*1 scores, we also recorded the percentage of times specific growth conditions were accurately or erroneously predicted, and we report these results in the form of a confusion matrix (Figure 4). Here, the column headings at the top show the predicted condition from the model on the test set and the rows show the true experimental condition. The numbers and shading in the interior of the matrix represent the percentage of cases that a given experimental condition was predicted to be a certain growth condition. The numbers within each row add up to 100. The large numbers/dark colorings along the diagonal highlight the high percentage of true positive predictions whereas any off-diagonal element represents the percentage of incorrect predictions for a given condition. We found that the erroneous off-diagonal predictions are partially driven by the uneven sampling of different conditions in the original dataset. Even though we used sample-number-adjusted class weights in all fitted models, we observed a trend of increasing fractions of correct predictions with increasing number of samples available under training (Supplementary Figure 3).

As we previously noted, the *F*1 score quantifies accuracy by only considering perfect predictions (i.e. when all 4 features are correctly predicted). A sample that is incorrectly classified for all four factors is thus treated the same as one that only differs from the true set of features by a single incorrect factor. In practice, we observed that the majority of incorrect predictions differed from their true condition vector by only a single value (Supplementary Figure 4).

**Joint consideration of mRNA and protein abundances improves model accuracy**

We next asked whether predictions could be improved by simultaneously considering mRNA and protein abundances. To address this question, we limited our analysis to the subset of 102 samples for which both mRNA and protein abundances were available, and ran our analysis pipeline for mRNA abundances only, protein abundances only, and for the combined dataset containing both mRNA and protein abundances. For all four machine-learning algorithms, protein abundances yielded significantly better predictions than mRNA abundances (Figure 5, Table 2). This is in contrast to Figure 3, where we saw increased accuracy using mRNA abundance data. However, as previously noted, our dataset contains a larger number mRNA abundance samples, which results in a larger amount of training data. When compared on the same exact conditions—as depicted in Figure 5—protein abundance data appears to be more valuable for discriminating between different growth conditions. Notably, the combined dataset consisting of both mRNA and protein abundance measurements yielded the best overall predictive accuracy, irrespective of machine-learning algorithm used (Figure 5, Table 2).

When considering the confusion matrices for the three scenarios (mRNA abundance, protein abundance, and combined), we found that many of the erroneous predictions arising from mRNA abundances alone were not that common when using protein abundances and vice versa (Supplementary Figures 5, 6). For example, when using mRNA abundances, many conditions were erroneously predicted as being exponential phase, glycerol, base Mg2+, base Na+, or as stationary phase, glucose, base Mg2+, high Na+; these particular predictions were rare or absent when using protein abundances. By contrast, when using protein abundances, several conditions were erroneously predicted as being stationary phase, glycerol, base Mg2+, base Na+, and these predictions were virtually absent when using mRNA abundance data. For predictions made from the combined dataset, erroneous predictions unique to either mRNA or protein abundances were generally suppressed, and only those predictions that arose for both mRNA and protein abundances individually remained present in the combined dataset (Supplementary Figure 7).

**Prediction accuracy differs between environmental features**

We also assessed the sources of inaccuracy in our models. As previously noted, the majority of incorrect predictions differed by only a single factor. The environmental features that accounted for most of these single incorrect predictions were Mg2+ concentration for the protein-only data and carbon sources for mRNA-only data. Moreover, growth phase (e.g. exponential, stationary, late-stationary) is not strictly an environmental variable and using this as a feature may partially skew our results if the goal is to predict *strictly* *external* conditions.

We thus trained and tested separate models using only exponential or only stationary phase datasets and asked to what extent these models could predict the remaining 3 environmental features (carbon source, [Mg2+], and [Na+]). We found that prediction accuracy was consistently better for models trained on exponential-phase samples compared to models trained on stationary-phase samples, irrespective of the machine-learning algorithm used or the data source (mRNA, protein abundances, or both) (Figure 6). This observation implies that *E. coli* gene expression patterns during stationary phase are less indicative of the external environment compared to cells experiencing exponential growth. A notable caveat is that we have fewer stationary phase samples and this decrease in accuracy may partially be due to the size of the training dataset. Even despite the lower accuracies, however, predictive accuracy from models trained solely on stationary phase cells was still much higher than random expectation, illustrating that quiescent cells retain a unique signature of the external environment for the conditions studied.

To better understand which conditions were the most problematic to predict, we constructed models to predict only *individual* features rather than the entire set of 4 features. When making predictions based on mRNA abundances only, models were most accurate in predicting growth phase and least accurate for carbon source, with Mg2+ and Na+ concentration falling between these two extremes. By contrast, when making predictions based on protein abundances, the most predictable feature was carbon source, the least predictable was Mg2+ concentration, and Na+ concentration and growth phase fell in-between these two extremes (Figure 7, Supplementary Figure 7). Finally, for the combined mRNA and protein abundance dataset, we found that accuracy for carbon source and Mg2+ concentration generally fell between the accuracies observed using mRNA and protein abundances individually. By contrast, accuracies for the Na+ concentration and growth phase were generally as good as—or better than—the prediction accuracies of the individual datasets (Supplementary Figure 9). Together, these findings highlight that mRNA and protein abundances differ in their ability to discriminate between particular environmental conditions.

**Model validation on external data.**

The samples that we studied throughout this manuscript are fairly heterogeneous and were collected by different individuals over a span of several months/years. However, different sample types were still analyzed within the same labs, by the same protocols, and thus may be more consistent than one might expect from data collected and analyzed independently by different labs—which would be an ultimate goal of future applications of this methodology. We thus applied our best-fitting protein abundance model to analyze protein data with *similar* conditions that was independently collected and analyzed [3]. Since this external dataset did not contain measurements for all of the 4196 proteins that we measured and constructed our model on, we tested two alternative approaches of applying our model to the external data. For the first approach, we filled the missing parts of the external data with the median values of our in-house data before making predictions. In the second approach, we restricted our training dataset to only include proteins that appeared in the external validation data set. These two approaches lead to comparable results (Table 3). Notably, our model made mostly correct predictions on this dataset. The model was most accurate at distinguishing between different growth phase data, and moderately accurate at distinguishing Na+ concentration and carbon source. The external data did not have variation in Mg2+ levels, however, and our model incorrectly predicted several samples to have high Mg2+.

**Discussion**

Our central goal in this manuscript was to determine whether gene expression measurements from a single species of bacterium are sufficient to predict environmental growth conditions. We analyzed a rich dataset of 152 samples for mRNA data and 105 samples for protein data across 16 distinct laboratory conditions as a proof-of-concept. We could show that *E. coli* gene expression is responsive to external conditions in a measurable and consistent way that permits identification of external conditions from gene signatures alone using supervised machine learning techniques. While *E. coli* is a well-characterized species, our analysis relies on none of this *a priori* knowledge. It is thus likely that increasing the number and diversity of training samples and conditions will produce further improvements in accuracy and discrimination between a wider array of conditions.

Interestingly, we found that consideration of mRNA and protein datasets alone are sufficient to produce accurate results, but that joint consideration of both datasets results in superior predictive accuracy. This finding implies that post-transcriptional regulation is at least partially controlled by external conditions, which has been observed by previous studies that have investigated multi-omics datasets [7,15,17,18]. Such regulation may result from post-translational modifications [19], stress coping mechanisms [20], differential translation of mRNAs, or protein-specific degradation patterns.

An important finding that we discovered was that cellular growth phase places limits on the predictability of external conditions, with stationary phase cells being particularly difficult to distinguish from one another irrespective of their external conditions. A possible explanation for this behavior might be associated with endogenous metabolism, whereby stationary phase cells start to metabolize surrounding dead cells instead of the provided carbon source. This new carbon source, which is independent of the externally provided carbon source, may suppress the differences between the cells in different external carbon source environments [21,22]. Another reason for this behavior might be related to strong coupling between gene expression noise and growth rate. Multiple studies have concluded that lower growth rates are associated with higher gene expression noise, which might be a survival strategy in harsh environments [23]. Negative correlations between population average gene expression and noise have been shown for *E. coli* and *Saccharomyces cerevisiae,* lending support for this theory [24,25]. Finally, we note that stationary phase cells have likely depleted the externally supplied carbon sources after several weeks of growth. The similarity of stationary phase cells to other stationary phase cells may be a consequence of them inhabiting more similar chemical environments to one another than during exponential growth. Nevertheless, discrimination of external environmental factors in stationary phase cells was still much better than random—indicating that these populations continue to retain information about the external environment despite their overall quiescence.

A relevant finding to emerge from our study is that different features of the environment may be more or less easy to discriminate from one another and this discrimination may depend on which molecular species is being interrogated. Growth phase, for instance, can be reliably predicted from mRNA concentrations but similar predictions from protein concentrations were less accurate. A possible explanation for this observation is the fact that mRNAs and proteins have different life-cycles [14,26]. Given the comparably slow degradation rates of proteins, a large portion of the stationary phase proteome is likely to have been transcribed during exponential phase growth. As another example, carbon sources can be reliably predicted from protein concentrations, but the accuracy of carbon source predictions from models trained on mRNA concentrations was more limited. Carbon assimilation is known to be regulated by post-translational regulation [27–29], which may be a possible reason for this finding (Figure 7, Supplementary Figure 8).

Despite the fact that we investigated over 150 samples spanning 16 unique conditions, a limitation of our work and conclusions is nevertheless sample size (though our study is comparable to or larger than similar multi-conditional transcriptomic and/or proteomic studies [3,30–32]). The comparison between all of our data with the more limited set that includes only the intersection of samples for which we have both mRNA and protein abundance data (Figure 4 compared to Supplementary Figures 4 and 5) indicates that prediction accuracy decreases as the size of our training sets gets smaller. This trend indicates that our training set sizes are still ultimately limiting to model accuracy. A second possible issue with our study is associated with sample number bias [33–35]. We made corrections with weight factors [36,37] and displayed the multi-conditional F1 score [38] to account for the fact that some conditions contained more samples, but the predictability of *individual* conditions nevertheless increased with the number of training samples for that particular condition (Supplementary Figure 3). This finding again highlights that increasing training data will likely result in higher prediction accuracy.

Our study is a proof-of-principle towards the goal of using gene expression patterns of natural species as a rapid and low-cost method for assessing environmental conditions. Other research has shown that the gene repertoire, derived from meta-genomic sequencing, may be useful for determining the presence of particular contaminants [6]. Our findings suggest that further incorporation of species-specific gene expression patterns can likely improve the accuracy of such methods. While genetically engineered strains may play a similar role as environmental biosensors, our study highlights that—with enough training data—the molecular composition of natural populations may provide sufficient information to accurately resolve past and present environmental conditions.

## **Materials and Methods**

### **Data preparation and overall analysis strategy**

We used a set of 155 *E. coli* samples previously described [13,14]. Throughout this study, we used different subsets of these samples in different parts of the analysis. For mRNA only and protein only analyses we used all 152 samples with mRNA abundances and all 105 samples with protein abundances, respectively. For performance comparison of machine learning models between mRNA and protein abundances we used the subset of 102 samples that have both mRNA and protein abundance data. After selecting appropriate subsets of the data for a given analysis, we added abundances from technical replicates, normalized abundances by size factors calculated via DeSeq2 [39], and applied a variance stabilizing transformation [40,41] (VST).

For each separate analysis, we divided the data into two subsets, (i) the training & tune set and (ii) the test set, using an 80:20 split (Figure 2). This division was done semi-randomly, such that our algorithm preserved the ratios of different conditions between the training & tune and the test subsets. We retained the condition labels in the training & tune data (thus our learning was supervised) but we discarded the sample labels for the test set. We then applied frozen Surrogate Variable Analysis [42] (fSVA) to remove batch effects from the samples. This algorithm can correct for batch effects in both the training & tune and the test data, without knowing the labels of the test data. After fSVA, we used principal component analysis [43] (PCA) to define the principal axes of the training & tune set and then rotated the test data set with respect to these axes. We then picked the top 10 most significant axes in the training & tune dataset for learning and prediction. Finally, we trained and tuned our candidate machine learning algorithms with the dimension reduced training & tune datasetand then applied those trained and tuned algorithms on the dimension-reduced test dataset to make predictions. This entire procedure was repeated 60 times for each separate analysis (Figure 2).

We used four different machine learning algorithms: SVM models with (i) linear, (ii) radial, and (iii) sigmoidal kernels, and (iv) random forest models. We used the R package **e1071** [44] for implementing SVM models and the R package **randomForest** [45] for implementing random forest models. SVMs with radial and sigmoidal kernels were set to use the c-classification [46] algorithm.

### **Model scoring**

We performed multiple sets of predictions throughout this manuscript: We predicted either (i) an individual parameter such as growth phase, carbon source, Mg2+ concentration, or Na+ concentration for each sample, or (ii) all conditions at once for each sample. In all cases, we assessed prediction accuracy via *F*1 scores that jointly assess precision and recall. In particular, for predictions of multiple conditions at once, we scored prediction accuracy via the multi-class macro *F*1 score [16,38,47] that normalizes individual *F*1 scores over individual conditions, i.e., it gives each condition equal weight instead of each sample. There are two different macro *F*1 score calculation that have been proposed in the literature. First, we can average individual *F*1 scores over all conditions *i* [38]:

where indicates the average and the individual *F*1 scores are defined as:

Second, we can average precision and recall and then combine those averages into an *F*1 score [16]:

Between these two options, we implemented the first, because it is not clear that individually averaging precision and recall before combining them into *F*1 appropriately balances prediction accuracies from different conditions with very different prediction accuracies.

### **Model training and tuning**

For training, we first divided the training & tune data further into separate training and tuning datasets, using a 75:25 split (Figure 2). As before for the subdivision between training and test data, we did this again semi-randomly, trying to preserve the ratios of individual conditions. We repeated this procedure 10 times to generate 10 independent pairs of training and tuning datasets. Next, we generated a parameter grid for the tuning process. We optimized the "cost" parameter for all three SVM models and the "gamma" parameter for the SVM models with radial and sigmoidal kernels (Supplementary Table 1A). For the random forest algorithm, we optimized three parameters; "mtry", "ntrees", and "nodesize" (Supplementary Table 1B).

We trained each of the four machine learning models on all 10 training datasets and made predictions on the 10 tuning datasets. We applied a class weight normalization during training, where class weights are inversely proportional to the corresponding number of training samples and calculated independently for each training run. We calculated macro *F*1 scores for each model parameter setting for each tuning dataset and then averaged the scores over all tuning datasets to obtain an average performance score for each algorithm and for each parameter combination. The parameter combination with the highest average *F*1 score was considered the winning parameter combination and was subsequently used for prediction on the test dataset (Figure 2).

### **Calculation for the external data**

We validated our predictions against independently published external data [3]. This external dataset consisted of 22 conditions, of which we could match five to our conditions. For all five samples, Mg2+ levels were held constant and approximately matched our base Mg2+ levels. The first sample used glucose as carbon source, did not experience any osmotic stress (no elevated sodium), and was collected in the exponential growth phase. The second sample used glycerol as carbon source, did not experience any osmotic stress (no elevated sodium), and was collected in the exponential growth phase. The third sample included 50mM sodium, glucose as carbon source, and was collected in the exponential growth phase. Because our high-sodium samples all included 100mM of sodium or more [1], this third sample fell in-between what we consider base sodium and high sodium. Samples four and five used glucose as carbon source, did not experience osmotic stress, and were measured after 24 and 72 hours of growth, respectively. In our samples, we defined stationary phase as 24–48 hours and late stationary phase as 1 to 2 weeks [1]. Thus, sample four matched our stationary phase samples and sample five fell in-between our stationary and late-stationary phase samples.

### **Statistical analysis and data availability**

All statistical analyses were performed in R. All processed data and analysis scripts are available on GitHub: <https://github.com/umutcaglar/ecoli_multiple_growth_conditions>. mRNA and protein abundances have been previously published [13,14]. Raw Illumina read data and processed files of read counts per gene are available from the NCBI GEO database [48] (accession numbers GSE67402 and GSE94117). Mass spectrometry proteomics data are available via PRIDE [49] (accession numbers PXD002140 and PXD005721).

**References**

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.