

RESEARCH ARTICLE

# Predicting bacterial growth conditions from mRNA and protein abundances

M. Umut Caglar, Adam J. Hockenberry , Claus O. Wilke\*

Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, United States of America

\* [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 thus 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 measurements from stationary phase cells typically provide less useful information for discriminating between conditions as compared to exponentially growing populations. 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.

## OPEN ACCESS

**Citation:** Caglar MU, Hockenberry AJ, Wilke CO (2018) Predicting bacterial growth conditions from mRNA and protein abundances. PLoS ONE 13(11): e0206634. <https://doi.org/10.1371/journal.pone.0206634>

**Editor:** Sam Paul Brown, Georgia Institute of Technology, UNITED STATES

**Received:** June 23, 2018

**Accepted:** October 16, 2018

**Published:** November 2, 2018

**Copyright:** © 2018 Caglar et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All data and code required to replicate the analyses are freely available at [https://github.com/umutcaglar/ecoli\\_learning\\_bacterial\\_response\\_optimization](https://github.com/umutcaglar/ecoli_learning_bacterial_response_optimization).

**Funding:** This study was funded by Army Research Office (<http://www.arl.army.mil/>; W911NF-12-1-0390), the National Institutes of Health (R01 GM088344) and the National Science Foundation (Cooperative agreement no. DBI-0939454, BEACON Center). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Introduction

Environmental conditions across the planet vary in terms of their capacity to support microbial life. Individual environments can also change rapidly over time, and these changes are likely to impact the composition of microbial communities and ecosystem functions in unpredictable ways [1,2]. To measure various properties of the environment, microbial cells can be engineered to act as biosensors via rational design of synthetic genetic circuits [3]. In contrast to gold standard approaches that are comparatively labor intensive and expensive, microbial cells can be engineered, for instance, to rapidly screen for the presence of heavy metals in aquatic environments [4]. Such applications can provide a useful, low-cost diagnostic for monitoring environmental changes and detecting pollutants and/or toxins [5], but individual synthetic biology applications take time and resources to develop. Additionally, there is an ever-present concern about potential dangers associated with releasing genetically engineered species into natural environments.

By contrast, prior work has shown that the natural species composition of an environment may be sufficient to serve as a rapid and low-cost biosensor to indicate the presence of various

**Competing interests:** The authors have declared that no competing interests exist.

contaminants according to the species abundances identified via meta-genomic sequencing [6–9]. However, many bacterial species within a community are generalists that are capable of thriving in diverse environments and must therefore sense and respond to various environmental signals [10]. For instance, *Escherichia coli* grows inside the comparatively warm, nutrient rich digestive tract of host organisms [11] but spends another portion of its life-cycle exposed to harsh environmental conditions upon being excreted and before finding another host. The mere presence of generalist species in an environment may provide little value for understanding past or current environmental conditions because their gene and expression diversity permits growth across variable environments [12]. The extent to which gene expression patterns of individual generalist species can be used to discriminate between environmental conditions—or to supplement species composition-based methods—remains unknown.

Gene expression profiles for individual cells or populations contain a wealth of information about their current physiological state, but measurements for thousands of genes across numerous conditions are challenging to integrate under traditional statistical methods. Further, combining different ‘omics’-scale technologies has been shown to provide more valuable information compared to monitoring only mRNA abundances alone, but integrating datasets is challenging due to the biases of individual methods [13] and the inevitability of batch-level effects that occur when datasets are generated across multiple labs and platforms [14,15]. Machine learning methods, by contrast, are frequently applied to such data-rich applications, for example to differentiate between cancerous and normal cells/tissues [16–20] using a variety of different machine learning models [21,22].

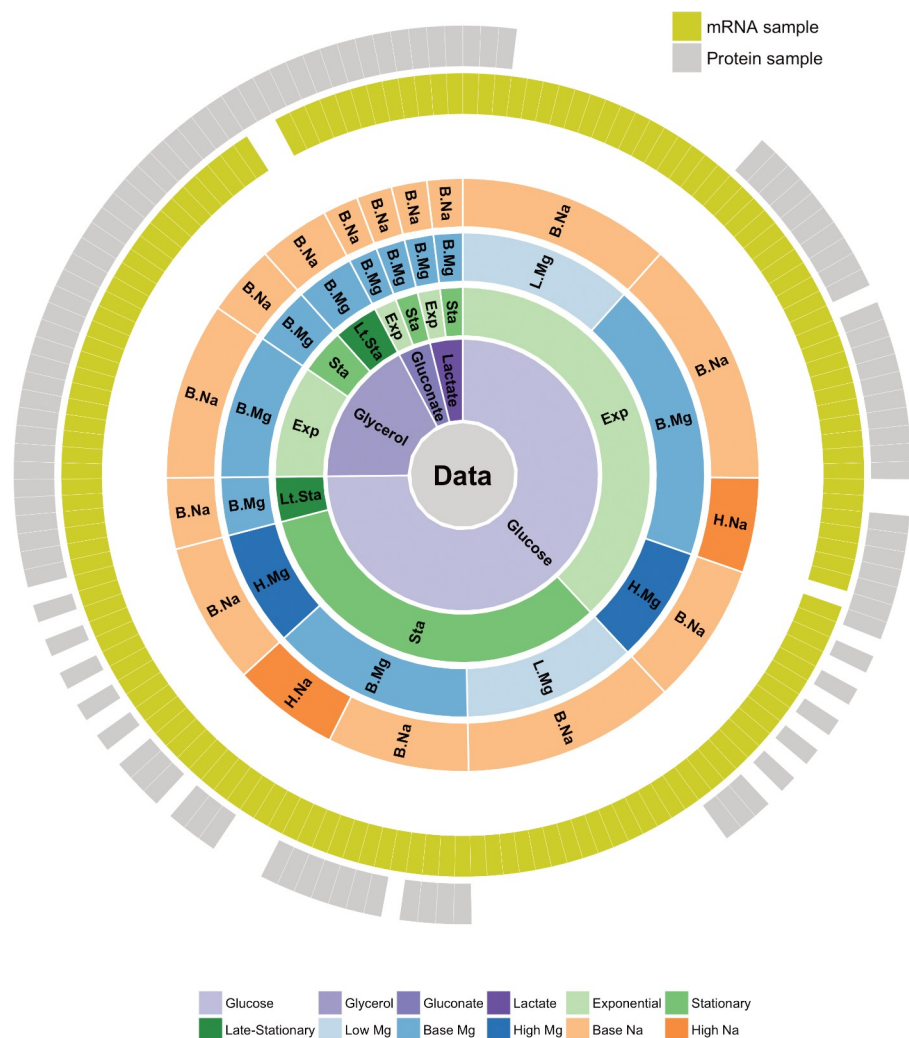
In microbiology applications, machine learning has been frequently applied to infer regulatory networks and molecular pathways from gene expression data [23–25], and from this knowledge to predict the growth capabilities of cells in different environments [26–28]. However, the primary focus in many of these studies has been to understand aspects of the cellular physiology. In this framework, environmental change serves as a perturbation that can be used to provide insight into *internal* cellular mechanisms/pathways [29]. While explicitly representing a cell’s internal state may help to predict cellular phenotypes such as growth capabilities across environments [30–32], it is unclear whether explicit representation of cellular metabolic pathways, for instance, are necessary to distinguish between cells growing in different environmental conditions [33,34]. Few studies have focused on using the abundance of cellular macromolecules to predict external environmental features across a range of partially-overlapping conditions and cellular growth states.

Here, we are interested in determining whether gene expression patterns can be leveraged to discriminate between environmental conditions in the absence of prior knowledge about the role and function of individual genes or explicit representation of cellular metabolism. Our study leverages a large dataset of transcriptomic and proteomic measurements of *E.coli* growth under multiple distinct but closely-related conditions [35]. We use mRNA and protein composition data to train several distinct machine learning models and find that highly similar environmental conditions can be discriminated with a high degree of accuracy. We also investigate which conditions are more- and less-challenging to discriminate and find that prediction accuracies decrease for stationary phase cells, indicating the importance of cellular growth for discriminating between conditions. Finally, we caution that the overall accuracy of our models may be limited by training set size; we found that the most difficult conditions to predict are the conditions for which we have the smallest amount of training data. This suggests that our findings may represent 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* (strain REL606) mRNA and protein abundances, measured under 34 different conditions [35,36]. This dataset consists of a total of 155 samples, for which mRNA abundances are available for 152 and protein abundances for 105 (Fig 1). For 102 samples, both mRNA and protein abundances are available. The 34 different experimental conditions were generated by systematically varying four parameters: carbon source, growth phase, Na<sup>+</sup> concentration, and Mg<sup>2+</sup> concentration. Here we further simplified the experimental conditions into a total of 16, by grouping similar conditions together (e.g., 100, 200, and 300mm Na<sup>+</sup> were all labelled as “high Na<sup>+</sup>”). For the remainder of this work (unless otherwise noted) we use the term “growth condition” to refer to the



**Fig 1. Overview of available gene expression data.** Our study uses a previously published dataset consisting of 155 samples [13, 14]. 152 samples have whole-transcriptome RNA-seq reads and 105 have mass-spec proteomics reads. 102 of the 155 samples have both mRNA and protein reads. Bacteria were grown on four different carbon sources (glucose, glycerol, gluconate, and lactate), two sodium concentrations (base and high), and three magnesium concentrations (low, base, and high). Samples were taken at multiple time points during a two-week interval, and they can be broadly subdivided into exponential phase, stationary phase, and late stationary phase samples.

<https://doi.org/10.1371/journal.pone.0206634.g001>

four-dimensional vector of categorical variables defining: i) growth phase (exponential, stationary, late stationary), ii) carbon source (glucose, glycerol, gluconate, lactate), iii)  $Mg^{2+}$  concentration (low, base, high), and iv)  $Na^+$  concentration (base, high). While we note that growth phase is not strictly an environmental feature, we suspected that this indicator of cellular state would be an important feature to consider since prior research has shown that the macromolecular composition of cells varies substantially between exponentially growing and stationary phase cells [35,36]. With these data and features, the question we set out to answer is: to what extent are machine learning models capable of discriminating between the known growth parameters given only knowledge of gene expression levels?

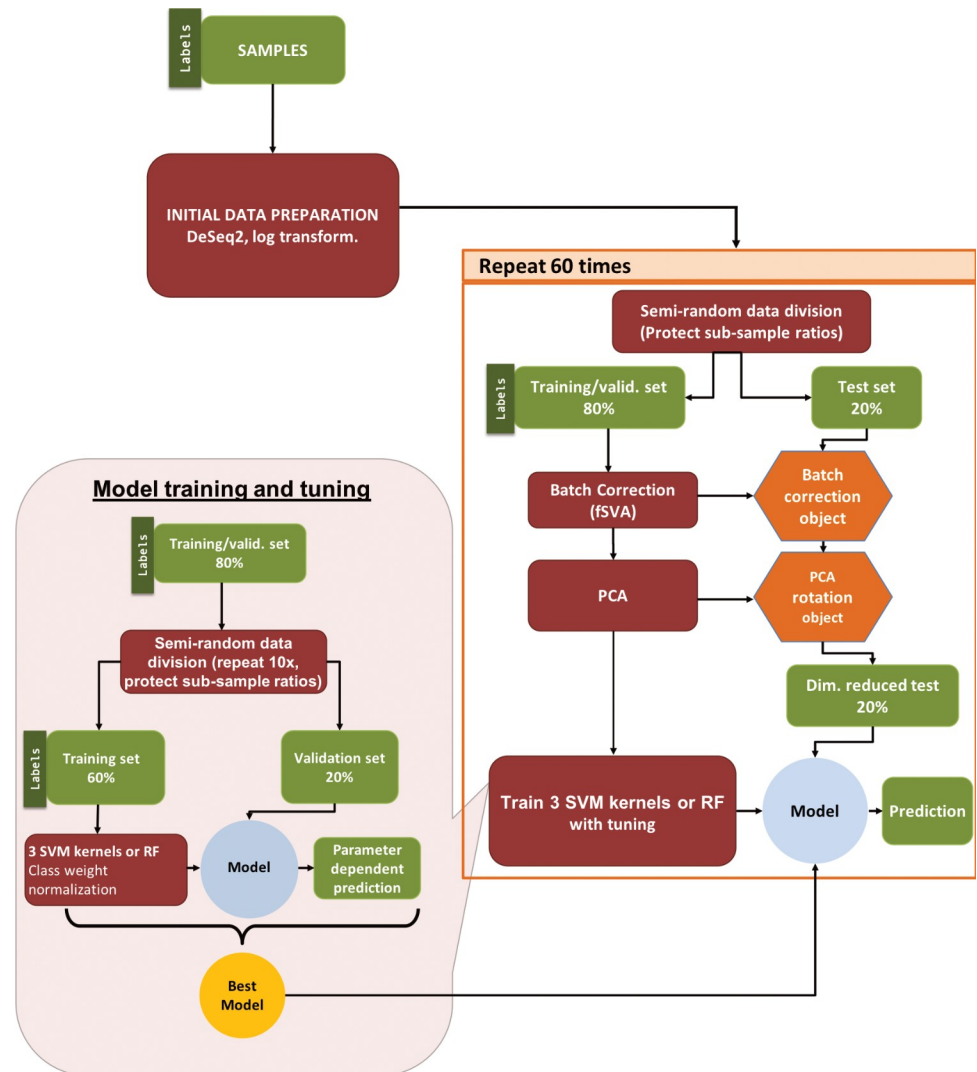
We first split samples into training/validation and test datasets using a semi-random approach that randomly splits data while preserving class balances. We performed several data processing steps, including batch correction and Principal Component Analysis (PCA), to reduce the dimensionality of the data (see [Materials and Methods](#) for details). We analyzed the top 10 genes contributing to the dominant principal components (PC1 and PC2, in both mRNA and protein datasets) and found that they all have orthologs in both B and K strains suggesting that data collection/extrapolation across different strains may not be particularly problematic for future studies ([S1 Table](#)). Additionally, PC1 was enriched for highly expressed genes in both mRNA and protein datasets (elongation factors, RNA polymerase subunits, outer membrane proteins, *etc.*), with the protein datasets also consisting of important chaperones (*dnaK* and *groEL*).

During the model tuning phase, we optimized hyperparameters in the machine learning pipeline by further splitting the training/validation data into training and validation sets, fitting models to the labeled training set, and optimizing for model accuracy on the validation set. We performed cross-validation by making 10 unique splits of the training/validation samples—with 75% of samples in training and 25% in validation sets—and searched across a parameter grid to select the hyperparameters that gave the highest  $F_1$  score on the validation set. Finally, we tested the accuracy of model predictions on the test dataset using the optimized hyperparameters from the tuning phase. To assess the overall robustness of our findings, we used repeated testing to replicate our entire pipeline 60 times and report the mean and range of variation in our final test set accuracies. Our pipeline is illustrated in [Fig 2](#) and described in greater 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. Since our overall goal was to demonstrate the feasibility and limitations of using machine learning on gene expression data to predict environmental features, we wanted to: i) ensure that our choice of machine learning algorithm did not substantially affect our results/conclusions and ii) determine the best method for this particular application since prior work has shown that the choice of machine learning model can substantially affect the accuracy of best fitting models [21,22]. We tested four different machine learning models: three based on Support Vector Machines (SVMs) with different kernels (radial, sigmoidal, and linear) and a fourth using random forest classification. We trained our models to predict [12,37] the entire four-dimensional condition vector at once for a given sample, and used the multi-class macro- $F_1$  score [38] to quantify prediction accuracy.

We note that various metrics can be applied to quantify model accuracy during classification tasks—each with particular strengths and limitations. The multi-class macro- $F_1$  score is the harmonic mean of precision (of all the positive predictions made by a model, “what



**Fig 2. Machine learning pipeline.** Our pipeline can be separated into three parts: (i) initial data preparation, (ii) training and prediction, and (iii) model tuning. After (i) initial data preparation, the samples are (ii) semi-randomly (preserving sub-sample ratios) separated into 2 parts, the training/validation set and the test set. After applying fSVA and PCA to the training/validation data, we train supervised SVM or random forest models on the training/validation set. After obtaining the tuned model we make predictions on the test data that has been batch corrected (via fSVA) and rotated (via PCA). This whole process is repeated 60 times to collect statistics on model performance. For model tuning (iii), the training/validation data set is similarly divided semi-randomly into training and validation datasets to optimize hyperparameters using a grid search approach. The tuning procedure is repeated 10 times and the parameter set that performs best—on average—during the 10 repeats is considered the winning model and is used for prediction on the test set data.

<https://doi.org/10.1371/journal.pone.0206634.g002>

fraction are correct?") and recall (of all the possible positive predictions, "what fraction does the model return?"). This quantity approaches zero if either quantity approaches zero, and it approaches one if both quantities approach one (representing perfect prediction accuracy). We further emphasize that our scoring scheme will classify a prediction as incorrect if even a single variable is incorrectly predicted, even if the predictions for the remaining three variables of interest are correct. We made this choice, rather than binary classification of individual variables, so that our findings would be conservative and represent a lower bound on the prediction accuracy for this task.

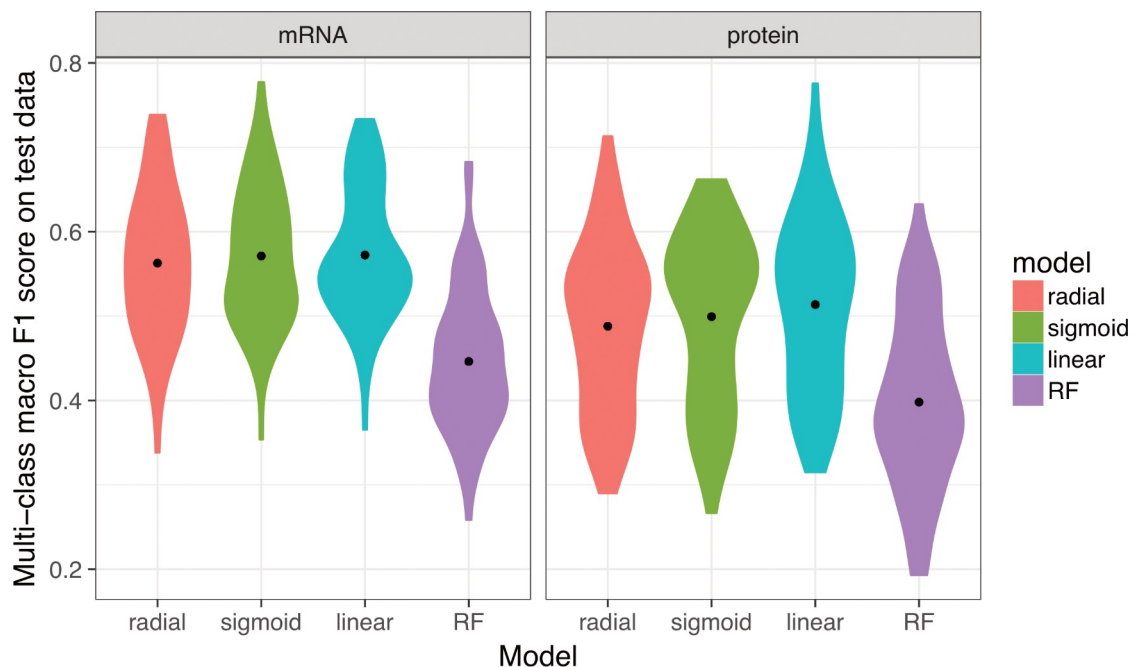
**Table 1. Winning-model distributions at the tuning stage.** Numbers show the number of times out of 60 independent runs that each given model had the highest  $F_1$  score in the tuning process. Results are shown separately for predictions on the mRNA and the protein data. The ties are counted for all the “winner” models as a result the sums are bigger than 60.

Model	mRNA	Protein
SVM, radial kernel	53	8
SVM, sigmoidal kernel	6	41
SVM, linear kernel	0	3
Random Forest	1	13

<https://doi.org/10.1371/journal.pone.0206634.t001>

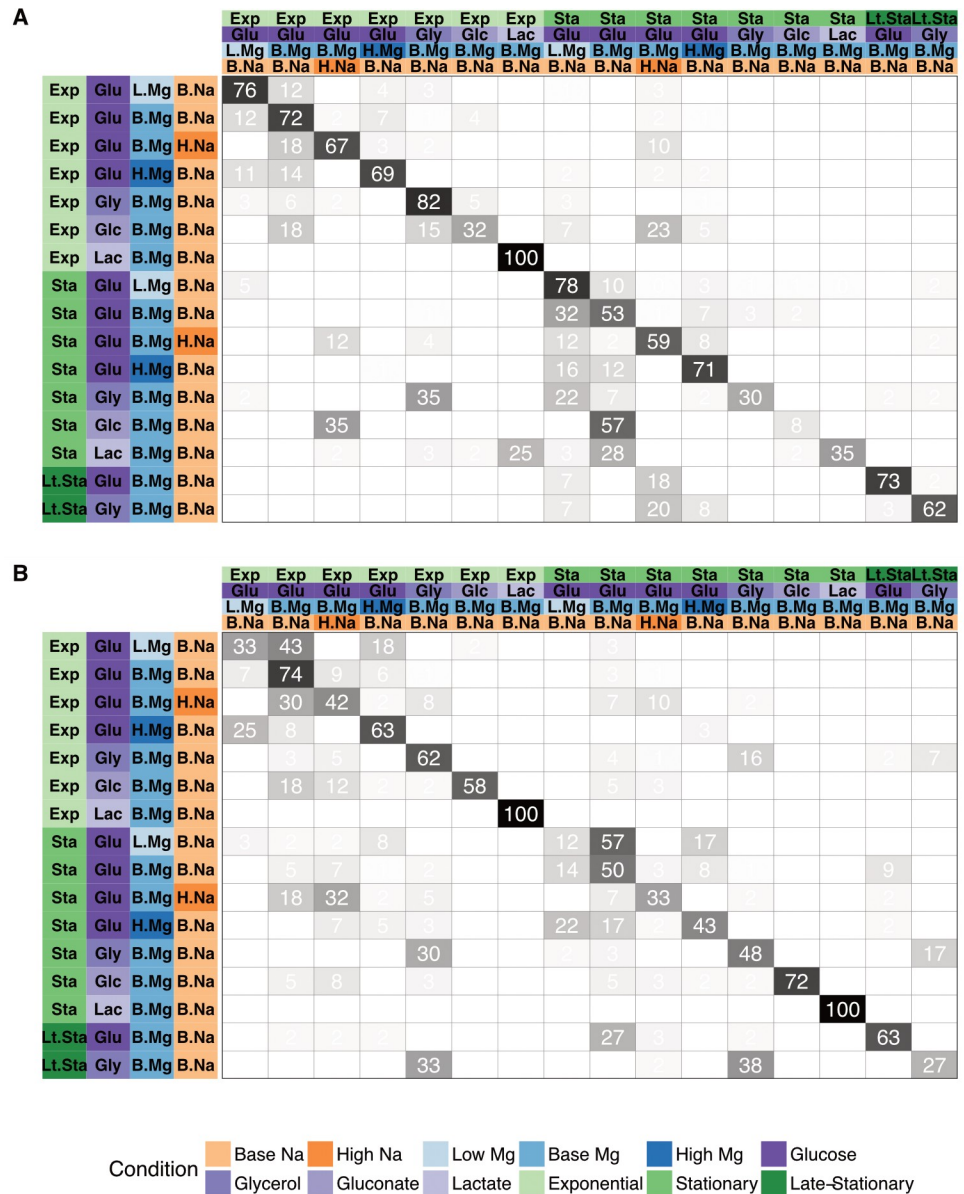
We assessed model performance during the tuning stage of our pipeline by recording which model and hyper-parameter set had the best macro- $F_1$  score for the validation set (S1 and S2 Figs). During this tuning stage, we found that the SVM model with a radial kernel clearly outperformed the other models when fit to mRNA data, and the random forest model outperformed 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 repeated testing of 60 independent replications were centered around a value of ~0.55 (Fig 3). The  $F_1$  score distributions were virtually identical for the three SVM models and was lower for the random forest model. Model performance on test data using only protein abundance measurements was slightly worse than what was achieved with mRNA abundance data. However, it is important to note that the protein abundance data contains fewer samples overall, which may partially explain



**Fig 3. Performance of multi-class predictions.** Distributions of multi-class macro  $F_1$  score for prediction of growth conditions from mRNA or protein abundances, using four different machine-learning algorithms (SVM with radial, sigmoidal, or linear kernel, and random forest [RF] models). For each model type, 60 independent models were trained on 60 independent subdivisions of the data into training/validation and test sets. We found that random forest models consistently performed worse than SVM models, and predictions based on mRNA data were slightly better than predictions based on protein data. The black dots represent the mean  $F_1$  scores.

<https://doi.org/10.1371/journal.pone.0206634.g003>



**Fig 4. Test set prediction accuracy for specific growth conditions.** In each matrix, rows represent true conditions and columns represent predicted conditions. The numbers in the cells and the shading of the cells represent the percentage (out of 60 independent replicates) with which a given true condition is predicted as a certain predicted condition. (A) Predictions based on mRNA abundances. Results are shown for the SVM with radial kernel, which was the best performing model in the tuning process on mRNA data, where it won 55 of 60 independent runs. In this panel, the average of the diagonal line is 60.5% and corresponding multi-class macro  $F_1$  score is 0.61. (B) Predictions based on protein abundances. Results are shown for the SVM with sigmoidal kernel, which was the best performing model in the tuning process on protein data, where it won 41 of 60 independent runs. In this panel, the average of the diagonal line is 55.1% and corresponding multi-class macro  $F_1$  score is 0.56.

<https://doi.org/10.1371/journal.pone.0206634.g004>

the decreased predictive accuracy of the protein-only model—a point to which we return to later.

In addition to assessing the overall accuracy of our predictive models using  $F_1$  scores, we also recorded the percentage of times specific growth conditions were accurately or erroneously predicted. We report these results in the form of a confusion matrix (Fig 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 (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 elements represent incorrect predictions. 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 during the training stage (S3 Fig).

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 features is thus treated the same as one that only differs from the true set of features by a single incorrect factor. In practice, however, we observed that the majority of incorrect predictions differed from their true condition vector by only a single value (S4 Fig).

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

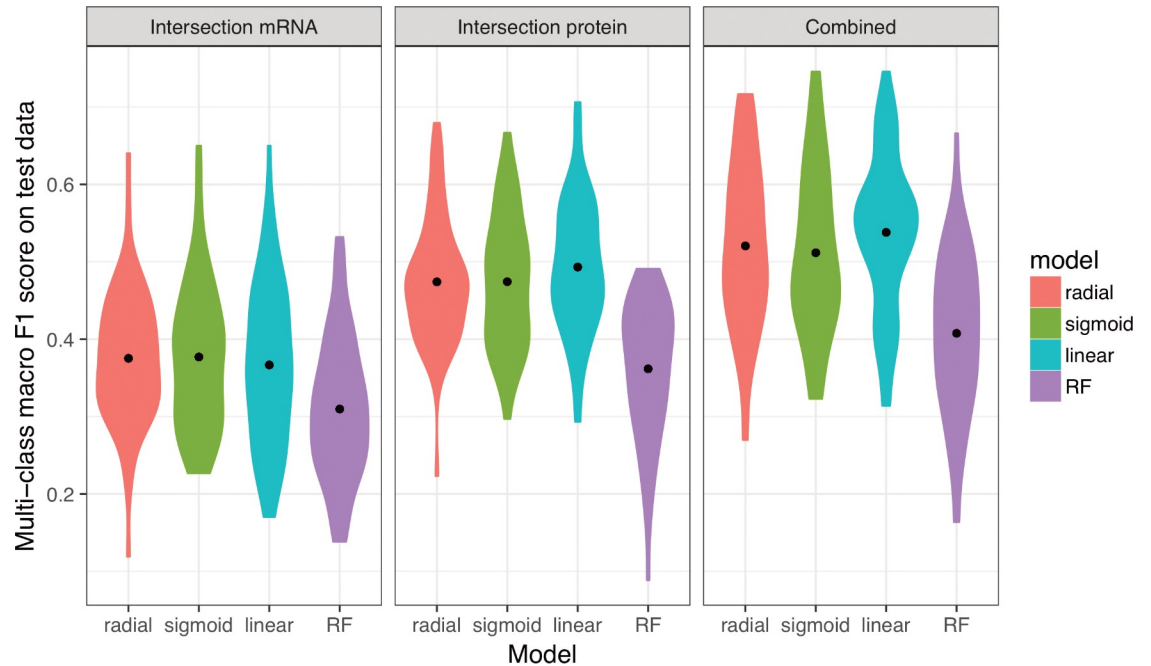
We next asked whether predictions could be improved by simultaneously considering both mRNA and protein abundances. To address this question, we limited our analysis to the subset of 102 samples for which both mRNA and protein abundances were available and ran our analysis pipeline for mRNA abundances 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 (Fig 5, Table 2). This is in contrast to Fig 3, where we saw increased accuracy using mRNA abundance data. However, as previously noted, our dataset contains more mRNA abundance samples, which results in a larger amount of training data for the results presented in Fig 3. When compared on the same exact conditions—as depicted in Fig 5—protein abundance data appears 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 (Fig 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 (S5 and S6 Figs). For example, when using mRNA abundances, many conditions were erroneously predicted as being exponential phase, glycerol, base  $Mg^{2+}$ , base  $Na^+$ ; or as stationary phase, glucose, base  $Mg^{2+}$ , high  $Na^+$ ; these particular erroneous 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  $Mg^{2+}$ , 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 suppressed, and only those predictions that arose for *both* mRNA and protein abundances individually remained present in the combined dataset (S7 Fig).

### Prediction accuracy differs between environmental features

We next assessed the sources of inaccuracy in our models. As previously noted, the majority of incorrect predictions differed by only a single factor (S4 Fig). The environmental features that





**Fig 5. Models trained on both mRNA and protein data perform better than models trained on only one data type.** The 102 samples for which we have both protein and mRNA abundances were used to compare the performance of machine learning models based on only mRNA, only protein, and mRNA and protein data combined (left to right, respectively). Regardless of the machine learning model used, prediction performance was higher for models that use protein data compared to mRNA data. Further, using both mRNA and protein data resulted in higher predictive power compared to either alone. Statistical significance of these differences is reported in Table 2.

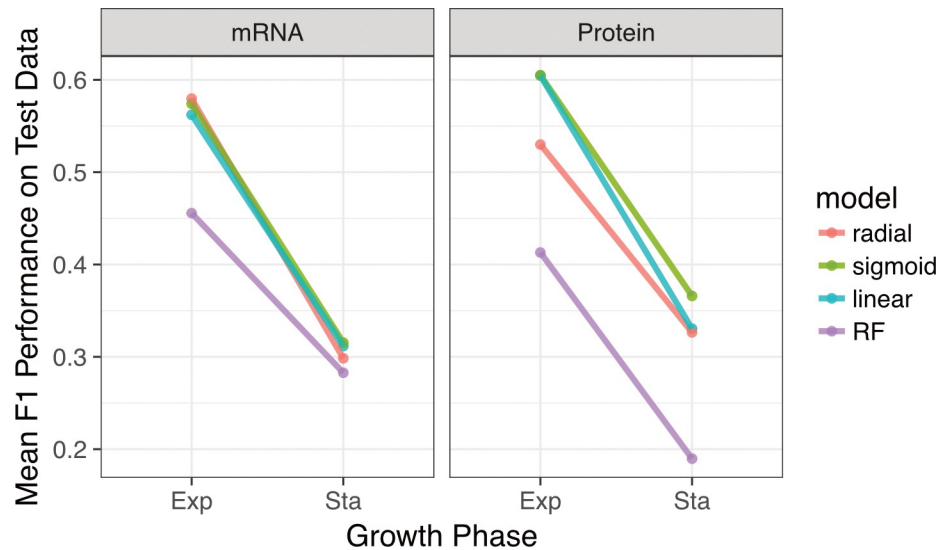
<https://doi.org/10.1371/journal.pone.0206634.g005>

accounted for most of these single incorrect predictions were  $Mg^{2+}$  concentration for the protein-only data and carbon source for mRNA-only data. Despite the importance of growth phase to macromolecular abundances, we reasoned that growth (e.g. exponential, stationary, late-stationary) is not an environmental variable and using this as a feature may partially skew our results if the goal is to predict *strictly external* conditions.

**Table 2. Statistical significance of comparisons shown in Fig 5.** Distributions of multi-class macro  $F_1$  scores were compared using t-tests. The adjusted  $P$  value reports the false discovery rate (FDR). All comparisons are statistically significant after correction for multiple testing via FDR.

Model	Comparison	$P$ value	Adjusted $P$ value
SVM, radial kernel	mRNA vs protein	1.943E-09	4.663E-09
SVM, radial kernel	mRNA + protein vs mRNA	3.908E-13	2.345E-12
SVM, radial kernel	mRNA + protein vs protein	8.425E-03	1.087E-02
SVM, sigmoidal kernel	mRNA vs protein	3.327E-08	6.654E-08
SVM, sigmoidal kernel	mRNA + protein vs mRNA	3.088E-11	1.235E-10
SVM, sigmoidal kernel	mRNA + protein vs protein	3.517E-02	3.517E-02
SVM, linear kernel	mRNA vs protein	4.728E-11	1.418E-10
SVM, linear kernel	mRNA + protein vs mRNA	1.595E-15	1.914E-14
SVM, linear kernel	mRNA + protein vs protein	9.441E-03	1.087E-02
Random forest	mRNA vs protein	1.818E-03	2.727E-03
Random forest	mRNA + protein vs mRNA	1.928E-07	3.306E-07
Random forest	mRNA + protein vs protein	9.968E-03	1.087E-02

<https://doi.org/10.1371/journal.pone.0206634.t002>

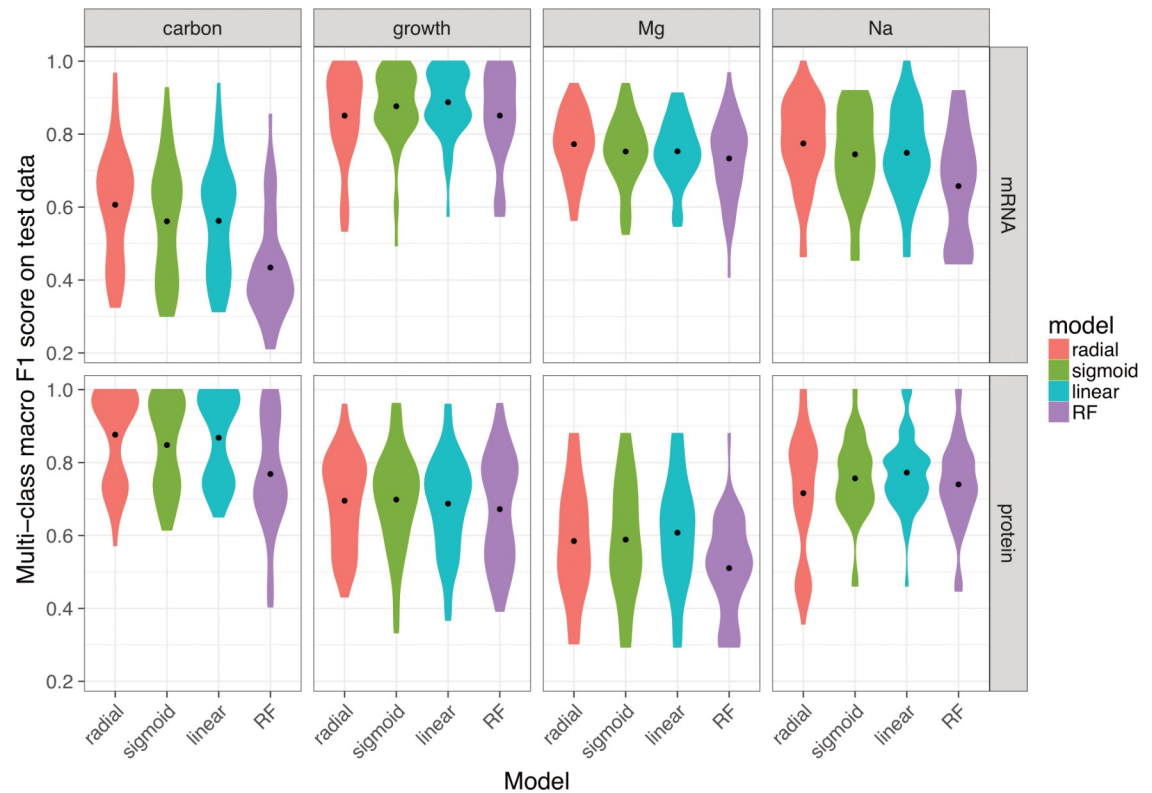


**Fig 6. Prediction accuracy systematically declines from exponential to stationary phase growth.** We separated data by growth phase and then trained separate models to predict carbon source, magnesium level, and sodium level within each growth phase. Regardless of the data source, prediction accuracy was substantially lower for stationary-phase samples than for exponential-phase samples. For each model and growth phase, dots show the mean  $F_1$  score over 60 replicates and lines connect mean  $F_1$  scores calculated for the same model.

<https://doi.org/10.1371/journal.pone.0206634.g006>

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,  $[Mg^{2+}]$ , 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) (Fig 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. Despite the lower accuracies, however, predictive accuracy from models trained solely on stationary phase cells was still much higher than random expectation, highlighting the fact 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. This is an easier task when compared to predicting all 4 dimensions simultaneously, and this ease is reflected in the relatively accurate confusion matrices that we observed (S8 Fig). For predictions based on mRNA abundances only, models were most accurate in predicting growth phase and least accurate for carbon source, with  $Mg^{2+}$  and  $Na^+$  concentration falling between these two extremes. By contrast, for predictions based on protein abundances, the most predictable feature was carbon source, the least predictable was  $Mg^{2+}$  concentration with  $Na^+$  concentration and growth phase fell in-between these two extremes (Fig 7, S8 Fig). Finally, for the combined mRNA and protein abundance dataset, we found that accuracy for carbon source and  $Mg^{2+}$  concentration fell between the accuracies observed using mRNA and protein abundances individually. By contrast, accuracies for the  $Na^+$  concentration and growth phase were as good as—or better than—the prediction accuracies of the individual datasets (S9 Fig). Together, these findings highlight that mRNA and protein abundances differ in their ability to discriminate between particular environmental conditions.



**Fig 7. Model performance on univariate predictions.** The multi-class macro- $F_1$  score of tuned models over test data for four individual conditions: carbon source, growth phase,  $Mg^{2+}$  levels, and  $Na^+$  levels. To keep mRNA-based and protein-based predictions comparable, we used the 102 samples with both mRNA and protein abundances for this analysis. To facilitate comparison with our previous results, we used the multi-class macro- $F_1$  score even for univariate predictions by averaging the component  $F_1$  scores for the individual outcomes (such as the different carbon sources).

<https://doi.org/10.1371/journal.pone.0206634.g007>

### 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 [12]. However, the largest external comparison dataset that we could find consisted of measurements for only ~2,000 proteins, which is substantially less the 4196 proteins that we measured and constructed our models on. Further, the particular bacterial strain (BW25113, a “K” strain) used in this external dataset was distinct from ours (REL606, a “B” strain), so not all of the proteins from our model have direct orthologs in this external dataset. Based on our analysis of the dominant genes contributing to the principal components (S1 Table), however, this strain level-variation may be less important than the missing data values. 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 (Table 3). In the second approach, we restricted our training dataset to only include proteins that appeared in the external validation data set (Table 4). These two approaches lead to comparable results.

**Table 3. Performance of the protein model on external data.**

Sample	Na <sup>+</sup> level	Mg <sup>2+</sup> level	Carbon source	Growth phase
A (Base)	base	high <sup>‡</sup>	Glucose	Exponential
B (Glycerol)	base	high <sup>‡</sup>	Glucose <sup>‡</sup>	Exponential
C (High Na <sup>+</sup> )	base <sup>†</sup>	high <sup>‡</sup>	Glucose	Exponential
D (Stationary)	base	base	Glucose	Stationary
E (Late stationary)	base	base	Glucose	Stationary <sup>†</sup>

For each of the five external samples we matched to conditions in our dataset, we show the predicted sodium level, magnesium level, carbon source, and growth phase. Regular text indicates a correct prediction for the sample in the given column.

The ‡ symbol indicates an incorrect prediction, and the † symbol indicates a prediction where the external data falls between two categories in our data (see [Methods](#) for details).

Predictions here are based on a model trained using our complete dataset, and any missing protein abundances in the external test data were replaced by the median values from the training dataset.

<https://doi.org/10.1371/journal.pone.0206634.t003>

Notably, our model made mostly correct predictions on this entirely independent dataset. The model was most accurate at distinguishing between different growth phase data, and moderately accurate at distinguishing Na<sup>+</sup> concentration and carbon source. The external data did not consist of samples with variable Mg<sup>2+</sup> concentrations, however, and we note that our model incorrectly predicted several samples to have high Mg<sup>2+</sup>.

## Discussion

Our central goal here was to determine whether gene expression measurements from a single species of bacterium are sufficient to predict environmental features. We analyzed a rich dataset of 152 samples for mRNA data and 105 samples for protein data across 16 distinctly classified laboratory conditions as a proof-of-concept. We showed that *E. coli* gene expression is responsive to external conditions in a measurable and consistent way that permits identification of environmental features from gene signatures alone via supervised machine learning techniques.

While *E. coli* is a well-characterized species, our analysis relies on none of this *a priori* knowledge. Previous approaches have focused on modeling cellular biology and metabolism in order to predict the growth capabilities of individual species in various environments [27–29]. Rather than using varied environmental conditions to interrogate cellular regulation [23,25], we instead determined that the abundances of cellular macromolecules themselves are sufficient to provide accurate information about environmental conditions.

**Table 4. Performance of the protein model on external data with different missing value assumptions.**

Sample	Na <sup>+</sup> level	Mg <sup>2+</sup> level	Carbon source	Growth phase
A (Base)	base	base	Gluconate <sup>‡</sup>	Exponential
B (Glycerol)	base	base	Gluconate <sup>‡</sup>	Exponential
C (High Na <sup>+</sup> )	high	base	Glucose	Exponential
D (Stationary)	base	base	Glucose	Stationary
E (Late stationary)	base	base	Glucose	Stationary <sup>†</sup>

Similar to [Table 3](#), here we show the accuracy of predictions based on a model that was trained only on the subset of proteins from our dataset that were present in the external test data.

Regular text indicates a correct prediction for the sample in the given column, the ‡ symbol indicates an incorrect prediction, and the † symbol indicates a prediction where the external data falls between two categories in our data.

<https://doi.org/10.1371/journal.pone.0206634.t004>

Interestingly, we found that consideration of mRNA and protein datasets alone is 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 [13,37,39,40]. Such regulation may result from post-translational modifications [41], stress coping mechanisms [42], differential translation of mRNAs, or protein-specific degradation patterns.

Our results show 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 may be 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 differences between cells growing on different external carbon sources [43,44]. 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 [45]. Negative correlations between population average gene expression and noise have been shown for *E. coli* and *Saccharomyces cerevisiae*, lending support for this theory [46,47]. Finally, we note that stationary phase cells are likely to have depleted the externally supplied carbon sources after several days of growth. The similarity of stationary phase cells to other stationary phase cells may be a consequence of them actually inhabiting more similar chemical environments to one another compared to during exponential growth where nutrient concentrations are more varied across conditions. Despite these caveats with regard to cellular growth phase, 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.

Another 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 may be the differences in life cycles between mRNAs and proteins [36,48]. 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 [49–51], which may be a possible reason for this finding (Fig 7, S9 Fig).

We investigated over 150 samples spanning 16 unique conditions, but 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 [12,52–54]). The comparison between all available data with the more limited set that includes only the samples for which we have both mRNA and protein abundances indicates that prediction accuracy decreases as the size of our training sets gets smaller (152 vs 102 mRNA samples, Fig 3 compared to Fig 5), strongly implying that training set sizes limit overall model accuracy for at least a portion of our results. A second but related possible issue with our study is associated with sample number bias [55–57]. We made corrections with weight factors [58,59] and used the multi-class macro- $F_1$  score [60] to account for

the fact that some conditions contained more samples than others, but the predictability of *individual* conditions nevertheless increased with the number of training samples for that particular condition (S3 Fig). Accuracy limitations could be more thoroughly evaluated through the use of learning curves to determine whether test set accuracies plateau with increasing training set size, but the class imbalance problem and fairly low number of overall samples per condition in our data make it difficult to evaluate accuracies across a broad range of training set sizes. Future work with larger sample numbers will be useful to interrogate whether accuracies are ultimately limited by training set sizes or by some other features inherent to the data and/or methods.

Another caveat of our study is our choice of score that we used to both optimize hyperparameters during the training phase and report for our test set accuracies. The most comprehensive and intuitive evaluation of our results is contained within confusion matrices (Fig 4); collapsing these data-rich matrices into a single number is convenient but can also be problematic. Quantifying the accuracy of multi-class classifiers (simultaneously predicting 4 separate vectors) is challenging and standards are generally lacking but the multi-class macro- $F_1$  score provides an intuitive scale (ranging from 0 to 1, with 1 representing perfect accuracy) and should account for all possible errors by averaging across predictions for each class. We recognize that the use of other scoring schemes, such as multi-class AUROC [61,62], could alter the model fits during the training phase and the final reported accuracies but the magnitude of these differences should be minor.

We also chose to evaluate different machine learning models throughout this manuscript to ensure the robustness of results and to determine if model choice had a substantial impact on classification accuracy. Overall, we found that the three SVM models performed equivalently to one-another and outperformed random forest models on most tasks. While machine learning models can be difficult to interrogate owing to data transformations, linear kernel SVM models return interpretable output that can be used to determine the most important features and therefore would be preferred for future work in this space given the seeming equivalence between linear, sigmoidal, and radial kernel models. The differences between all models were minor, however, and this finding shows that the accuracy of our classification task is robust to different assumptions.

Our study is a proof-of-principle, demonstrating that gene expression patterns of natural species may provide useful information for assessing various aspects of the environment. Other research has shown that the microbial species composition, derived from meta-genomic sequencing, may be useful for determining the presence of particular contaminants [6]. Our results 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 low-cost environmental biosensors, we show that—with enough training data—the macromolecular 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 [35,36]. 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 [63], and applied a variance stabilizing transformation [64,65] (VST).

For each separate analysis, we divided the data into two subsets, (i) the training/validation set and (ii) the test set, using an 80:20 split (Fig 2). This division was done semi-randomly, such that our algorithm preserved the ratios of different conditions between the training/validation and the test subsets. We retained the condition labels in the training/validation data (thus our learning was supervised) but we discarded the sample labels for the test set. We then applied frozen Surrogate Variable Analysis [66] (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 [67] (PCA) to define the principal axes of the training/validation 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/validation dataset for learning and prediction. Finally, we trained and tuned our candidate machine learning algorithms with the dimension reduced training/validation dataset and 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 (Fig 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 [68] for implementing SVM models and the R package randomForest [69] for implementing random forest models. SVMs with radial and sigmoidal kernels were set to use the *c*-classification [70] algorithm.

### Model scoring

Our goal throughout this work was to predict multiple parameters (i.e., growth phase, carbon source, Mg<sup>2+</sup> concentration, or Na<sup>+</sup> concentration) of each growth condition at once. Therefore, we could not measure model performance via ROC or precision–recall curves, which assume a simple binary (true/false) prediction. Instead, we assessed prediction accuracy via  $F_1$  scores, which 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 [38,60,71] 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  $i$  [60]:

$$F_{1, \text{macro}} = \langle F_{1,i} \rangle$$

where  $\langle \dots \rangle$  indicates the average and the individual  $F_1$  scores are defined as:

$$F_{1,i} = 2 * \text{Precision}_i * \text{Recall}_i / (\text{Precision}_i + \text{Recall}_i).$$

Alternatively, we can average precision and recall and then combine those averages into an  $F_1$  score [38]:

$$F_{1, \text{macro}} = 2 \langle \text{Precision}_i \rangle \langle \text{Recall}_i \rangle / (\langle \text{Precision}_i \rangle + \langle \text{Recall}_i \rangle).$$

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/validation data further into separate training and validation datasets, using a 75:25 split (Fig 2). As before, for the subdivision between training/validation and test data, we did this semi-randomly while trying to preserve the ratios of individual conditions. We repeated this procedure 10 times to generate 10 independent pairs of training and validation 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 (S1 Fig). For the random forest algorithm, we optimized three parameters; "mtry", "ntrees", and "nodesize".

We trained each of the four machine learning models on all 10 training datasets and made predictions on the 10 validation 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 validation dataset and then averaged the scores over all validation 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 (Fig 2).

## Model validation on external data

We validated our predictions against independently published external data [12]. This external dataset consisted of 22 conditions, of which we could match five to our conditions. For all five samples,  $Mg^{2+}$  levels were held constant in the external dataset at a level that *approximately* matched our base  $Mg^{2+}$  concentrations. The first sample used glucose as carbon source, did not experience any osmotic stress (no elevated sodium), and was collected during 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 [35], 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 [35]. 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](https://github.com/umutcaglar/ecoli_multiple_growth_conditions) (permanent archived version available via zenodo: 10.5281/zenodo.1294110). mRNA and protein abundances have been previously published [35,36]. Raw Illumina read data and processed files of read counts per gene are available from the NCBI GEO database [72] (accession numbers GSE67402 and GSE94117). Mass spectrometry proteomics data are available via PRIDE [73] (accession numbers PXD002140 and PXD005721).

## Supporting information

**S1 Table. Feature importance in principal component analysis.** Listed are the top 10 genes that contribute the most to the indicated dataset and principal component. (XLSX)



**S1 Fig. Tuning results for predictions based on mRNA data, generated from one of 60 independent runs and chosen for demonstration purposes.** Model performance is measured as the mean  $F_1$  score over 10 independent tuning runs. Higher numbers indicate better performance. (A) Tuning results for SVMs with linear kernel. Only the cost parameter was tuned. (B) Tuning results for SVMs with radial kernel. The cost and gamma parameters were tuned. The red dot indicates the winning parameter combination. (C) Tuning results for SVMs with sigmoidal kernel. The cost and gamma parameters were tuned. The red dot indicates the winning parameter combination. (D) Tuning results for random forest models. The mtry, node-size, and ntrees parameters were tuned. We used three values for ntrees, 1000, 5000, and 10000, shown as three separate panels. The red dot indicates the winning parameter combination.

(TIFF)

**S2 Fig. Tuning results for predictions based on protein data, generated from one of 60 independent runs and chosen for demonstration purposes.** (A) Tuning results for SVMs with linear kernel. Only the cost parameter was tuned. (B) Tuning results for SVMs with radial kernel. The cost and gamma parameters were tuned. The red dots indicate the winning parameter combinations. (C) Tuning results for SVMs with sigmoidal kernel. The cost and gamma parameters were tuned. The red dot indicates the winning parameter combination. (D) Tuning results for random forest models. The mtry, nodesize, and ntrees parameters were tuned. We used three values for ntrees, 1000, 5000, and 10000, shown as three separate panels. The red dot indicates the winning parameter combination.

(TIFF)

**S3 Fig. Percentage of correct predictions as a function of the number of samples during training.** (A) Predictions based on mRNA abundances. (B) Predictions based on protein abundances.

(TIFF)

**S4 Fig.** The error count distribution for mRNA (A) and protein (B) confusion matrices. The number of mis-predicted labels (x-axis) indicates how many of the 4 possible condition variables that an individual prediction got wrong. 0 mis-predicted labels (the majority in both cases) means that model predictions were 100% accurate. In both cases (mRNA and protein), when an incorrect prediction was made, it was most frequently due to a single variable being incorrectly predicted (number of mis-predicted labels with a value of 1) as compared to errors predicting more than one variable for a given condition (2 and 3 mis-predicted labels).

(TIFF)

**S5 Fig. Prediction accuracy for specific growth conditions for intersection mRNA data.** Rows represent true conditions and columns represent predicted conditions. The numbers in the cells and the shading of the cells represent the percentage (out of 60 independent replicates) with which a given true condition is predicted as a certain predicted condition. Predictions based on mRNA abundances, generated by using subset of mRNA samples which has matching protein pairs. Results are shown for the SVM with radial kernel, which was the best performing model in the tuning process on mRNA data, where it won 48 of 60 independent runs. The average of the diagonal line is 44.1% and multi class macro  $F_1$  score is 0.43.

(TIFF)

**S6 Fig. Prediction accuracy for specific growth conditions for intersection protein data.** Rows represent true conditions and columns represent predicted conditions. The numbers in the cells and the shading of the cells represent the percentage (out of 60 independent

replicates) with which a given true condition is predicted as a certain predicted condition. Predictions based on protein abundances, generated by using subset of protein samples which has matching mRNA pairs. Results are shown for the SVM with sigmoid kernel, which was the best performing model in the tuning process on mRNA data, where it won 47 of 60 independent runs. The average of the diagonal line is 52.3% and corresponding multi class macro F1 score is 0.53.

(TIFF)

**S7 Fig. Prediction accuracy for specific growth conditions for intersection mRNA & protein data.** Rows represent true conditions and columns represent predicted conditions. The numbers in the cells and the shading of the cells represent the percentage (out of 60 independent replicates) with which a given true condition is predicted as a certain predicted condition. Predictions based on protein abundances, generated by using subset of mRNA & protein samples which has matching pairs. Results are shown for the SVM with sigmoid kernel, which was the best performing model in the tuning process on combined intersection data, where it won 27 of 60 independent runs. The average of the diagonal line is 56.1% and corresponding multi class macro F1 score is 0.57.

(TIFF)

**S8 Fig. Prediction accuracy for univariate predictions using intersection mRNA and intersection protein data, as in the main text Fig 7.** (A) Prediction of carbon source from mRNA abundances. (B) Prediction of carbon source from protein abundances. (C) Prediction of growth phase from mRNA abundances. (D) Prediction of growth phase from protein abundances. (E) Prediction of  $Mg^{2+}$  levels from mRNA abundances. (F) Prediction of  $Mg^{2+}$  levels from protein abundances. (G) Prediction of  $Na^+$  levels from mRNA abundances. (H) Prediction of  $Na^+$  levels from protein abundances.

(TIFF)

**S9 Fig. Prediction accuracy for univariate predictions based on intersection mRNA abundances, intersection protein abundances, or the combined dataset including both mRNA and protein abundances.** Protein abundances are more predictive for carbon source and  $Mg^{2+}$  levels, and mRNA abundances are more predictive for  $Na^+$  levels and growth phase.

(TIFF)

## Acknowledgments

The authors acknowledge support from the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources.

## Author Contributions

**Conceptualization:** M. Umut Caglar, Claus O. Wilke.

**Data curation:** M. Umut Caglar.

**Formal analysis:** M. Umut Caglar, Adam J. Hockenberry.

**Funding acquisition:** Claus O. Wilke.

**Methodology:** Adam J. Hockenberry.

**Project administration:** Claus O. Wilke.

**Resources:** Claus O. Wilke.

**Software:** M. Umut Caglar.

**Supervision:** Claus O. Wilke.

**Writing – original draft:** M. Umut Caglar, Adam J. Hockenberry, Claus O. Wilke.

**Writing – review & editing:** M. Umut Caglar, Adam J. Hockenberry, Claus O. Wilke.

## References

1. Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, D'Agrosa C, et al. A global map of human impact on marine ecosystems. *Science*. 2008; 319: 948–952. <https://doi.org/10.1126/science.1149345> PMID: 18276889
2. Sahney S, Benton MJ, Ferry PA. Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land. *Biol Lett*. 2010; 6: 544–547. <https://doi.org/10.1098/rsbl.2009.1024> PMID: 20106856
3. Slomovic S, Pardee K, Collins JJ. Synthetic biology devices for *in vitro* and *in vivo* diagnostics. *Proc Natl Acad Sci*. 2015; 112: 14429–14435. <https://doi.org/10.1073/pnas.1508521112> PMID: 26598662
4. Bereza-Malcolm LT, Mann G, Franks AE. Environmental sensing of heavy metals through whole cell microbial biosensors: A synthetic biology approach. *ACS Synth Biol*. 2015; 4: 535–546. <https://doi.org/10.1021/sb500286r> PMID: 25299321
5. Roggo C, van der Meer JR. Miniaturized and integrated whole cell living bacterial sensors in field applicable autonomous devices. *Curr Opin Biotechnol*. 2017; 45: 24–33. <https://doi.org/10.1016/j.copbio.2016.11.023> PMID: 28088093
6. He Z, Zhang P, Wu L, Rocha AM, Tu Q, Shi Z, et al. Microbial functional gene diversity predicts groundwater contamination and ecosystem functioning. *mBio*. 2018; 9: e02435–17. <https://doi.org/10.1128/mBio.02435-17> PMID: 29463661
7. Poisot T, Kéfi S, Morand S, Stanko M, Marquet PA, Hochberg ME. A continuum of specialists and generalists in empirical communities. *PloS One*. 2015; 10: e0114674. <https://doi.org/10.1371/journal.pone.0114674> PMID: 25992798
8. Flynn TM, Sanford RA, Ryu H, Bethke CM, Levine AD, Ashbolt NJ, et al. Functional microbial diversity explains groundwater chemistry in a pristine aquifer. *BMC Microbiol*. 2013; 13: 146. <https://doi.org/10.1186/1471-2180-13-146> PMID: 23800252
9. Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, Barua S, et al. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J*. 2010; 4: 660–672. <https://doi.org/10.1038/ismej.2009.154> PMID: 20182523
10. Sriswasdi S, Yang C, Iwasaki W. Generalist species drive microbial dispersion and evolution. *Nat Commun*. 2017; 8: 1162. <https://doi.org/10.1038/s41467-017-01265-1> PMID: 29079803
11. Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, Kupiec M, et al. Adaptive prediction of environmental changes by microorganisms. *Nature*. 2009; 460: 220–224. <https://doi.org/10.1038/nature08112> PMID: 19536156
12. Schmidt A, Kochanowski K, Vedelaar S, Ahmé E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent *Escherichia coli* proteome. *Nat Biotechnol*. 2016; 34: 104–110. <https://doi.org/10.1038/nbt.3418> PMID: 26641532
13. Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat Commun*. 2016; 7. <https://doi.org/10.1038/ncomms13090> PMID: 27713404
14. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010; 11. <https://doi.org/10.1038/nrg2825> PMID: 20838408
15. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA. A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics*. 2011; 12: 33–50. <https://doi.org/10.1093/biostatistics/kxq043> PMID: 20625178
16. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci*. 2001; 98: 15149–15154. <https://doi.org/10.1073/pnas.211566398> PMID: 11742071
17. Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*. 2002; 18: 1216–1226. <https://doi.org/10.1093/bioinformatics/18.9.1216> PMID: 12217913

18. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*. 2002; 18: 39–50. <https://doi.org/10.1093/bioinformatics/18.1.39> PMID: 11836210
19. Lee Y, Lee C-K. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*. 2003; 19: 1132–1139. <https://doi.org/10.1093/bioinformatics/btg102> PMID: 12801874
20. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. 2000; 16: 906–914. <https://doi.org/10.1093/bioinformatics/16.10.906> PMID: 11120680
21. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*. 2005; 21: 631–643. <https://doi.org/10.1093/bioinformatics/bti033> PMID: 15374862
22. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*. 2008; 9: 319. <https://doi.org/10.1186/1471-2105-9-319> PMID: 18647401
23. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol*. 2006; 7: R36. <https://doi.org/10.1186/gb-2006-7-5-r36> PMID: 16686963
24. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol*. 2007; 3. <https://doi.org/10.1038/msb4100120> PMID: 17299415
25. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*. 2007; 5: e8. <https://doi.org/10.1371/journal.pbio.0050008> PMID: 17214507
26. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, et al. A predictive model for transcriptional control of physiology in a free living cell. *Cell*. 2007; 131: 1354–1365. <https://doi.org/10.1016/j.cell.2007.10.053> PMID: 18160043
27. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc Natl Acad Sci*. 2010; 107: 17845–17850. <https://doi.org/10.1073/pnas.1005139107> PMID: 20876091
28. Carrera J, Estrela R, Luo J, Rai N, Tsoukalas A, Tagkopoulos I. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *Escherichia coli*. *Mol Syst Biol*. 2014; 10: 735–735. <https://doi.org/10.15252/msb.20145108> PMID: 24987114
29. Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol*. 2014; 10: e1003580. <https://doi.org/10.1371/journal.pcbi.1003580> PMID: 24762745
30. Brandes A, Lun DS, Ip K, Zucker J, Colijn C, Weiner B, et al. Inferring carbon sources from gene expression profiles using metabolic flux models. *PLoS One*. 2012; 7: e36947. <https://doi.org/10.1371/journal.pone.0036947> PMID: 22606312
31. Sridhara V, Meyer AG, Rai P, Barrick JE, Ravikumar P, Segrè D, et al. Predicting growth conditions from internal metabolic fluxes in an *in-silico* model of *E. coli*. *PLoS One*. 2014; 9: e114608. <https://doi.org/10.1371/journal.pone.0114608> PMID: 25502413
32. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Mol Syst Biol*. 2015; 11: 784. <https://doi.org/10.15252/msb.20145697> PMID: 25678603
33. Airoidi EM, Huttenhower C, Gresham D, Lu C, Caudy AA, Dunham MJ, et al. Predicting cellular growth from gene expression signatures. *PLoS Comput Biol*. 2009; 5: e1000257. <https://doi.org/10.1371/journal.pcbi.1000257> PMID: 19119411
34. Gutteridge A, Pir P, Castrillo JI, Charles PD, Lilley KS, Oliver SG. Nutrient control of eukaryote cell growth: a systems biology study in yeast. *BMC Biol*. 2010; 8: 68. <https://doi.org/10.1186/1741-7007-8-68> PMID: 20497545
35. Caglar MU, Houser JR, Barnhart CS, Boutz DR, Carroll SM, Dasgupta A, et al. The *E. coli* molecular phenotype under different growth conditions. *Sci Rep*. 2017; 7: 45303. <https://doi.org/10.1038/srep45303> PMID: 28417974
36. Houser JR, Barnhart C, Boutz DR, Carroll SM, Dasgupta A, Michener JK, et al. Controlled measurement and comparative analysis of cellular components in *E. coli* reveals broad regulatory changes in response to glucose starvation. *PLoS Comput Biol*. 2015; 11: e1004400. <https://doi.org/10.1371/journal.pcbi.1004400> PMID: 26275208
37. Wilmes A, Limonciel A, Aschauer L, Moenks K, Bielow C, Leonard MO, et al. Application of integrated transcriptomic, proteomic and metabolomic profiling for the delineation of mechanisms of drug induced

- cell stress. *J Proteomics*. 2013; 79: 180–194. <https://doi.org/10.1016/j.jprot.2012.11.022> PMID: 23238060
38. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009; 45: 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
  39. Nie L, Wu G, Culley DE, Scholten JCM, Zhang W. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit Rev Biotechnol*. 2007; 27: 63–75. <https://doi.org/10.1080/07388550701334212> PMID: 17578703
  40. Zhang W, Li F, Nie L. Integrating multiple “omics” analysis for microbial biology: application and methodologies. *Microbiology*. 2010; 156: 287–301. <https://doi.org/10.1099/mic.0.034793-0> PMID: 19910409
  41. Oliveira AP, Sauer U. The importance of post-translational modifications in regulating *Saccharomyces cerevisiae* metabolism. *FEMS Yeast Res*. 2012; 12: 104–117. <https://doi.org/10.1111/j.1567-1364.2011.00765.x> PMID: 22128902
  42. de Nadal E, Ammerer G, Posas F. Controlling gene expression in response to stress. *Nat Rev Genet*. 2011; 12: 833–845. <https://doi.org/10.1038/nrg3055> PMID: 22048664
  43. Kolter R, Siegele DA, Tormo A. The stationary phase of the bacterial life cycle. *Annu Rev Microbiol*. 1993; 47: 855–874. <https://doi.org/10.1146/annurev.mi.47.100193.004231> PMID: 8257118
  44. Maier RM, Pepper IL. Chapter 3—Bacterial Growth. *Environmental Microbiology* (Third edition). San Diego: Academic Press; 2015. pp. 37–56. <https://doi.org/10.1016/B978-0-12-394626-3.00003-X>
  45. Keren L, van Dijk D, Weingarten-Gabbay S, Davidi D, Jona G, Weinberger A, et al. Noise in gene expression is coupled to growth rate. *Genome Res*. 2015; gr.191635.115. <https://doi.org/10.1101/gr.191635.115> PMID: 26355006
  46. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O’Shea E, Pilpel Y, et al. Noise in protein expression scales with natural protein abundance. *Nat Genet*. 2006; 38: 636–643. <https://doi.org/10.1038/ng1807> PMID: 16715097
  47. Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*. 2010; 329: 533–538. <https://doi.org/10.1126/science.1188308> PMID: 20671182
  48. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res*. 2010; 38:D750–D753. <https://doi.org/10.1093/nar/gkp889> PMID: 19854939
  49. Martínez-Gómez K, Flores N, Castañeda HM, Martínez-Batallar G, Hernández-Chávez G, Ramírez OT, et al. New insights into *Escherichia coli* metabolism: carbon scavenging, acetate metabolism and carbon recycling responses during growth on glycerol. *Microb Cell Factories*. 2012; 11: 46. <https://doi.org/10.1186/1475-2859-11-46> PMID: 22513097
  50. Perrenoud A, Sauer U. Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. *J Bacteriol*. 2005; 187: 3171–3179. <https://doi.org/10.1128/JB.187.9.3171-3179.2005> PMID: 15838044
  51. Kumar R, Shimizu K. Transcriptional regulation of main metabolic pathways of *cyoA*, *cydB*, *fnr*, and *fur* gene knockout *Escherichia coli* in C-limited and N-limited aerobic continuous cultures. *Microb Cell Factories*. 2011; 10: 3. <https://doi.org/10.1186/1475-2859-10-3> PMID: 21272324
  52. Soufi B, Krug K, Harst A, Macek B. Characterization of the *E. coli* proteome and its modifications during growth and ethanol stress. *Front Microbiol*. 2015; 6: 103. <https://doi.org/10.3389/fmicb.2015.00103> PMID: 25741329
  53. Lewis NE, Cho B-K, Knight EM, Pálsson BO. Gene expression profiling and the use of genome-scale *in silico* models of *Escherichia coli* for analysis: providing context for content. *J Bacteriol*. 2009; 191: 3437–3444. <https://doi.org/10.1128/JB.00034-09> PMID: 19363119
  54. Yoon SH, Han M-J, Jeong H, Lee CH, Xia X-X, Lee D-H, et al. Comparative multi-omics systems analysis of *Escherichia coli* strains B and K-12. *Genome Biol*. 2012; 13: R37. <https://doi.org/10.1186/gb-2012-13-5-r37> PMID: 22632713
  55. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor Newsl*. 2004; 6: 20–29. <https://doi.org/10.1145/1007730.1007735>
  56. Chawla NV. Data mining for imbalanced datasets: An overview. In: *Data Mining and Knowledge Discovery Handbook*. Springer US; 2005. pp. 853–867. [https://doi.org/10.1007/0-387-25465-X\\_40](https://doi.org/10.1007/0-387-25465-X_40)
  57. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009; 21: 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
  58. Huang Y-M, Du S-X. Weighted support vector machine for classification with uneven training class sizes. 2005 International Conference on Machine Learning and Cybernetics. 2005; 7:4365–4369 <https://doi.org/10.1109/ICMLC.2005.1527706>

59. Support Vector Machines [Internet]. [cited 24 Apr 2017]. Available: <http://www.di.fc.ul.pt/~jpn/r/svm/svm.html>
60. Yang Y. An evaluation of statistical approaches to text categorization. *Inf Retr.* 1999; 1: 69–90. <https://doi.org/10.1023/A:1009982220290>
61. Hand DJ, Till RJ. A simple generalisation of the Area Under the ROC Curve for multiple class classification problems. *Mach Learn.* 2001; 45: 171–186.
62. Landgrebe TCW, Duijn RPW. Approximating the multiclass ROC by pairwise analysis. *Pattern Recognit Lett.* 2007; 28: 1747–1758. <https://doi.org/10.1016/j.patrec.2007.05.001>
63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; 15: 550. <https://doi.org/10.1186/s13059-014-0550-8> PMID: [25516281](https://pubmed.ncbi.nlm.nih.gov/25516281/)
64. Differential analysis of count data—the DESeq2 package [Internet]. 27 Jun 2016 [cited 12 Apr 2016]. Available: <http://www.bioconductor.org/packages//2.13/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf>
65. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010; 11: R106. <https://doi.org/10.1186/gb-2010-11-10-r106> PMID: [20979621](https://pubmed.ncbi.nlm.nih.gov/20979621/)
66. Parker HS, Bravo HC, Leek JT. Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ.* 2014; 2: e561. <https://doi.org/10.7717/peerj.561> PMID: [25332844](https://pubmed.ncbi.nlm.nih.gov/25332844/)
67. Jolliffe I. *Principal Component Analysis*. Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd; 2014. <https://doi.org/10.1002/9781118445112.stat06472>
68. Meyer D, Wien TU. Support Vector Machines. The interface to libsvm in package e1071. Online-Documentation of the package e1071 for “R”. 2001.
69. Liaw A, Wiener M. Classification and regression by randomForest. *R News.* 2002; 2: 18–22.
70. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol.* 2011; 2: 27:1–27:27. <https://doi.org/10.1145/1961189.1961199>
71. Ghamrawi N, McCallum A. Collective multi-label classification. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management.* 2005;195–200. <https://doi.org/10.1145/1099554.1099591>
72. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013; 41: D991–D995. <https://doi.org/10.1093/nar/gks1193> PMID: [23193258](https://pubmed.ncbi.nlm.nih.gov/23193258/)
73. Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnol.* 2014; 32:223–226. <https://doi.org/10.1038/nbt.2839> PMID: [24727771](https://pubmed.ncbi.nlm.nih.gov/24727771/)