Response to editor:

Editors' comments to the author:

I enjoyed this MS, and encourage minor revisions in light of the comments by the expert reviewer (below). I would also encourage the authors to introduce / motivate the alternative SVM models (radial / sigmoidal / linear) in the main text, as they occupy plenty of real estate in figs 3,5,7 but are unexplained in the text. Fig 6 is interesting but suffers from the caveat stated around line 222 that the sample count is unbalanced. could this be addressed via sub-sampling?

We thank you and the reviewer for your time and the thoughtful feedback. We have made numerous changes as a result of yours and reviewer comments that has strengthened the message of our manuscript. We address reviewer comments in the next section. With regard to specific editorial comments:

1. We have included motivation and interpretation about the need to analyze different models:

From the Results:

> "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-F1 score [38] to quantify prediction accuracy."

From the Discussion:

> "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 with regard to 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."

2. In Fig. 6, we observed a decrease in predictability when using stationary phase cells but the editor points out that unbalanced sample number may partially be the cause of this decrease. Our previously written caveat about fewer stationary phase samples was, however, erroneous and we thank the editor for pointing this out. Figure 1 (which remains entirely unchanged in this submission) highlights that the exponential and stationary phase conditions are roughly evenly divided across our samples. The much larger sources of class imbalance owes to carbon source (glucose dominates) and salt concentrations (baseline sodium, for instance, dominates). We agree with the editor that sub-sampling to maintain equal sample sizes would the be best way to test this phenomenon if these samples were particularly imbalanced. However, at this stage we have corrected the text to remove the previously written caveat given the balanced sample number across this particular variable.

Response to reviewer 1:

Comments to the Author:

Reviewer #1: Caglar et al show a novel perspective to use bacteria as biosensors using expression data coupled with machine learning. The authors use previously published expression data (mRNA and protein) to infer features of the E. coli environment. This idea is interesting, as this question often centers on understanding the physiology of the bacteria; it is exciting to flip the question and inquire about the environment instead. The proposed strategy is also very simple, which can allow for a simple extension in other lab or field studies. Overall the manuscript is well written, although there are parts where seemingly disconnected paragraphs hinder the reader's ability to follow along.

The introduction is too short, and the history of approaches to this problem is not adequately cited. The manuscript proposes the use of expression data as input for classification of environmental features. Their proposed strategy Support Vector Machines (SVM) have been extensively explored to interrogate physiological cues in bacteria using expression data. Previous studies have implemented SVM to classify different physiological states with a great deal of success. Applying this strategy to infer features of the environment can be phrased as an extension of these ideas. It would be useful for the reader to provide a short description of the differences of this approach with previously implemented strategies. Then in the discussion, it would be helpful to discuss how the problem evaluated is different.

We thank the reviewer for their time, kind words, and thoughtful comments. As a result of this feedback, we have revised our manuscript substantially. We have taken care to clarify the writing throughout, expanded and revised the introduction to further discuss previous applications of machine learning in the context of gene expression data, and elaborated our discussion section to particularly highlight the novelty of the problem we are studying as well as the uniqueness of our question/approach. Responses to individual comments follow below in black text.

Major comments

1. The strains used are not described. During the training step, the axis selected for PCA may correspond to a small number of genes in E. coli. If these genes are strain specific (which is unlikely) then the predictions of this model are restricted to a lab setting, if not a point could be made about the usability of the pipeline. The section “Model validation on external data” mentions the external dataset does not contain values for all of E. coli proteins. Is this due to a lack in the measurement tool, or are the strains missing specific orthologous groups? In the former, the correction applied by Caglar et al is an educated guess; in the latter it adds information that does not represent the system.

We thank the reviewer for pointing out our previous failure to discuss the strains or strain-level variation. In this revision, we explicitly state which strain we used (REL606, a “B strain”), and draw attention to the fact that the external dataset uses a different strain (BW25113, a “K strain”). Despite the strain differences, missing values predominantly stem from size differences between the measured proteome in the external dataset versus ours (~2,050 proteins compared to ~4,000 proteins, respectively). For all individual proteins contained in our dataset, we extracted the corresponding ortholog by gene name identification in the external dataset. “Missing values” in the external dataset therefore are nevertheless partially caused by both a failure to measure these values as well as a lack of orthologs in the corresponding strains, but we did not specifically investigate the source of the missing data. Additionally, our response to the next comment discusses the ubiquity of the genes that drive the dominant principal components. To clarify these points we have added/edited several sentences throughout the manuscript.

Notably, from the Results:

> “We used a previously generated dataset of whole-genome E. coli (strain REL606) mRNA and protein abundances, measured under 34 different conditions [35,36].”

Also within the Results:

> “However, the largest external comparison dataset that we could find consisted of measurements for ~2,000 proteins, which is substantially less the 4196 proteins that we measured and constructed our models on. Further, the particular 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...”

Transforming the data using PCA is a sensible choice given the limitations of SVM to make predictions when many features are present. However, an exploration of the resulting dimensions could add to the discussion and overall impact of the manuscript. If the genes

driving the variance of PCA are represented in multiple strains it extends usefulness of the method beyond a laboratory setting.

We thank the reviewer for the suggestion and have included a supplementary table (S1 Table) in this revision which lists the top 10 genes that contribute to PC1 and PC2 for both mRNA and protein datasets. In this revision, we also discuss these findings in the text—particularly within the context of strain-level variation in *E. coli*:

> "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*)."

2. The explanation of the method seems lacking. In line 112 is mentioned use of a cross-validation strategy that separates test and training set. However, in the methods section is clear that the "training set" is further subdivided to tune the model. This tuning as described should be described as cross-validated (the optimization procedure should also be described). And the multiple separation should be referred as repeated testing rather than cross-validation. The standard terminology should be 60 times replicated 10fold cross-validation with balanced sampling.

We agree that the terminology here can be potentially confusing and have altered our wording to better align with the standard terminologies when possible. We now note two initial "sets": i) training/validation set and ii) test set. We refer to this new nomenclature throughout this revision and additionally use the terminology of cross-validation to describe the optimization of hyper-parameters and repeated testing when it is clear that we are repeating the entire pipeline with different initial splits of the data into the various test sets to estimate the error in our overall pipeline. Below, we highlight the most relevant paragraph from our Results section setting up this nomenclature (this paragraph also highlights that our optimization procedure was an exhaustive grid-search):

> "We first split samples into training/validation and test datasets using a semi-random approach that randomly splits data while preserving class balances. During the model tuning phase, we optimized hyperparameters in the machine learning pipe-line 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 F1 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 the Materials and Methods."

Using Fscore is useful in evaluating the ability of the model to retrieve positive examples. However, because this metric is invariant to true negative classification, it makes it more difficult to interrogate potential weaknesses of the model that could lead to misclassification (like the imbalance of the data described in the main text) as well as ways to address them. Thoroughly exploring other metrics such as receiving operator characteristic (ROC) can expand on the weaknesses of the model. In addition, other metrics may offer a more intuitive interpretation to an audience familiar with machine learning procedures. SVM models can bet set up to return class membership probabilities. These would be the usual value for which a threshold would be varied to produce a ROC, the area under the curve can then summarize the information in a simple way to be compared among tests.

We thank the reviewer for this suggestion. Ultimately, after many discussions with several colleagues, we found that the most common usage and interpretation of ROC curves is for binary classification and extensions to multi-class prediction are relatively novel (including volume under ROC surfaces: doi:10.1109/TMI.2007.908687, doi:10.1007/978-3-540-39857-8_12), not entirely standardized (doi:10.1016/j.patrec.2007.05.001, doi:10.1109/IJCNN.2008.4633979, doi:10.1023/A:1010920819831), and can therefore become confusing to describe and interpret. On top of this, we note that some researchers question the usage of ROC curves in general, preferring Precision-Recall curves in cases of class imbalance such as we have here (doi:10.1371/journal.pone.0118432). With regard to the true negatives, the reviewer is correct that these are ignored by the F1 score. However, when averaging across classes these prediction errors are nevertheless incorporated since one class's false negative prediction will by definition be another class's false positive prediction. Based on the reviewers comment, and to make our decision more explicit, we have included the following discussion of scoring schemes within the text. We additionally focus the readers' attention on the confusion matrix results, which is our preferred way to visualize and interpret the data and which largely avoids these issues regarding how to summarize multi-dimensional data with a single or small set of numbers.

From the Results:

> "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-F1 score is the harmonic mean of precision (of all the positive predictions made by a model, "what 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."

From the Discussion:

> "Another caveat of our study is our choice of score that we used to both optimize hyper-parameters 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-F1 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 training phase and the final reported accuracies but the magnitude of these differences should be minor."

As part of the conclusions of the manuscript is hinted that increasing the amount of training data will lead to an increase in accuracy, this could be shown by constructing learning curves of the models. Additionally this can provide information on the internal confidence of the model after training compared to test. This is useful to determine where is the model having faults (e.g. the imbalance of the dataset).

We thank the reviewer for this suggestion and agree that learning curves are a common and useful way to assess potential limitations of training data sizes. However, in our case we believe that the class imbalance problem would be largely prohibitive to extracting meaningful learning curves outside of a relatively narrow window of training set sizes due to the small size of our dataset. With our limited number of samples for certain labels, smaller training set sizes would

simply either fail to capture many of these labels at all if we trained without considering class imbalance, or our training set would be subject to pseudo-randomization problems and would always include roughly the same datasets regardless of overall training set size for certain classes (possibly resulting in erroneous interpretations). While the reviewer is correct that S3 Fig does not guarantee that the *overall* performance of our method is limited by training data, the presence of a correlation between accuracy and dataset size for particular labels is suggestive of this point without requiring extensive re-training/re-testing of the model. Also, we highlight that the comparison between mRNA results in Fig 3 (152 samples) vs Fig 5 (102 samples) further indicates some degree of training set size limitation. We have altered the text at several points to draw attention to this issue, most notably in the substantially revised Discussion:

> "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-F1 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."

3. SVM models and particularly SVM-RFE have been used to explore physiology of bacteria for a long time. Its important that the introduction and the discussion make clear that the problem at hand is not to use machine learning in a generic way to distinguish between physiological states but rather to obtain information about environmental cues. The use of growth phase data, which is inherently a physiological condition detracts (in my view) from the point. Splitting the data by this condition is justified by the results, as predictive ability is lower at different at different stages of growth. However, this is not the focus of the paper. If the focus is to predict features of the environment the discussion should center around the predictive capabilities in this regard. Scenarios where this method could be used to obtain valuable information about the environment are not discussed as well.

We thank the reviewer for this comment and have substantially revised our text to include a broader discussion of these models and their use in physiological state prediction as well as the novelty of our approach:

From the Introduction:

> "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."

Also from the Introduction:

> "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."

From the Results:

> "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]."

From the Discussion:

> "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."

Additionally it is unclear why the authors chose to focus on multiclass classification creating a four dimensional vector including all four classes, this detracts from their conclusions as the imbalances in test and fold selection are made more obvious with a limited number of samples. And the predictive power of the model is reduced. Supplementary figure S8 shows that binary classification has better predictive ability that their base model. If there is any additional merit in the selection of multiclass classification is not obvious and should be stated in the main text.

We thank the reviewer for this comment and have written additional clarification in the text about this decision. Briefly, our goal was simply to be as conservative as possible so as not to overstate our results. This is a proof-of-principle study, and as such we felt that we should make the challenge as difficult as possible to think of a "worst case scenario" given that even though our environments are varied they are nevertheless controlled laboratory conditions. Additionally, we note here that the challenge of predicting a multi-dimensional vector can partially mitigate batch effects. By requiring our models to simultaneously predict all 4 dimensions, the models should be prevented from simply over-fitting to individual batch effects that may be present for particular sets of conditions. From the Results section:

> "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."

Minor comments

-Figure 3. In line 143 is described the replications were centered around 0.7 with a reference to figure 3 however the figure shows the value to be lower than this.

We apologize for this error and have corrected the sentence to read:

> "When using mRNA abundance data alone, the distribution of F1 scores from repeated testing of 60 independent replications were centered around a value of ~0.55 (Fig 3)."

-Lack of italics through the reference section (e.g. E. coli, et al).

We thank the reviewer for pointing out this oversight and have thoroughly updated our reference section to correct numerous issues including italicization of latin terms and several other formatting errors.

-Figure 8. Could be summarized as a table. Asterisks or special characters can be used to highlight the miss classifications and intermediate states.'

We thank the reviewer for this suggestion and have subsequently converted our Figure 8 into two separate tables (Tables 3 and 4) using special characters to denote the various categories as the reviewer suggests.

-Supplementary figure 3. This data could be better summarized constructing training curves (see above), which depict the Fscore as the number of samples increase and how do training and test sets differ.

We greatly appreciate this suggestion and direct the reviewer to our response under point 2 of the major comments for a more elaborate discussion of this issue.

-By using feature selection they could shed light on what features matter from the ones that do not and made their predictions more applicable to different sets, as well as unravel reasons why some conditions are more difficult to predict than others (e.g Na vs Mg)

We thank the reviewer for this suggestion. Feature weights are readily accessible for Random Forest classifiers but less so for SVM. While interpretable weights can be extracted from SVM models with a linear kernel, to the best of our knowledge the corresponding coefficients for sigmoidal and radial kernels are not readily interpretable. Since our goal was to provide a proof-of-principle for the task of predicting environmental features, we felt that including a section that was limited in its scope to discussing the nuances of particular models would be potentially confusing. We have, however, included a brief discussion of this possibility in the Discussion section to highlight a future direction for improvement of these methods:

> "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."