# Predicting bacterial growth conditions from mRNA and protein abundances

**Supplementary file**

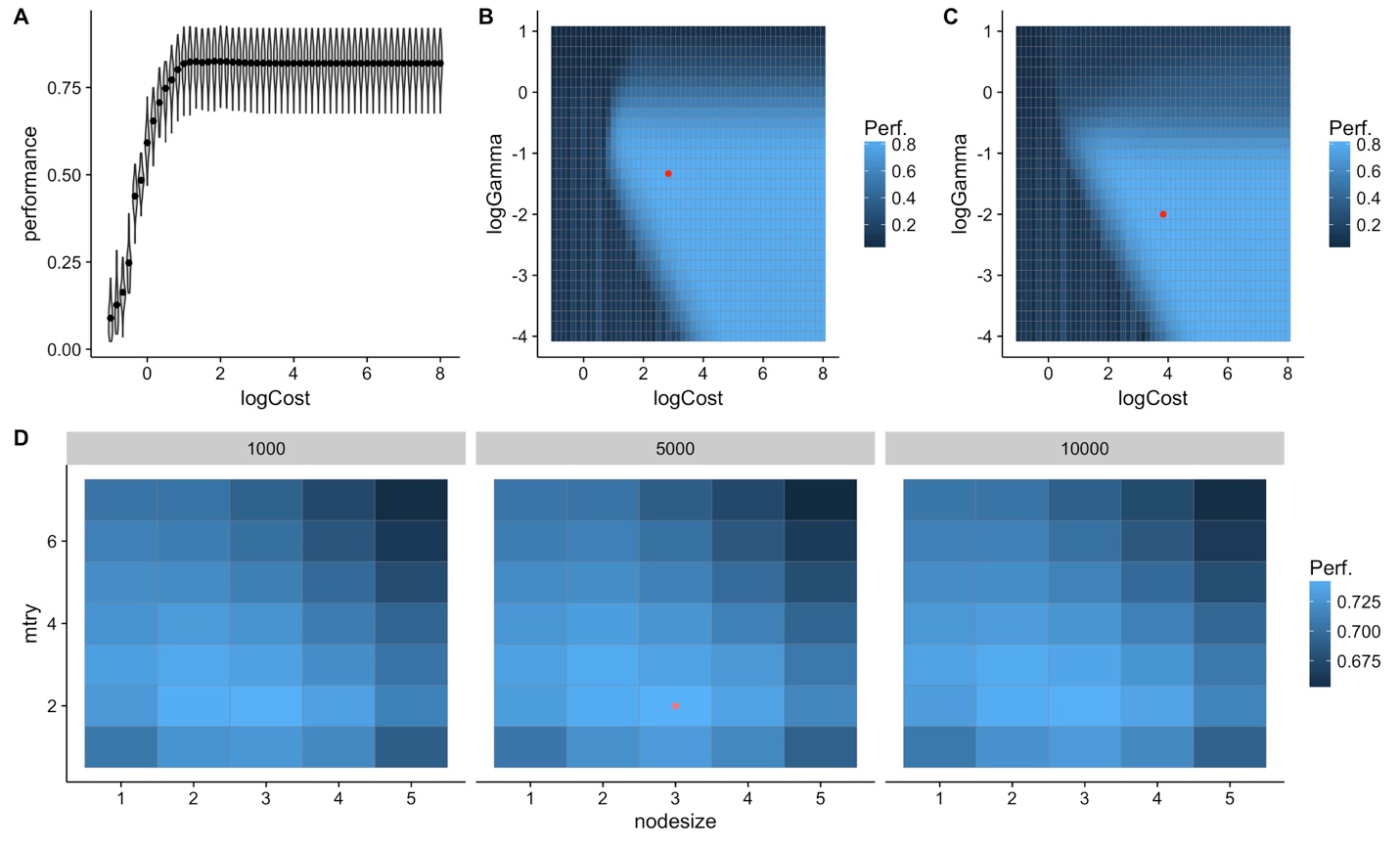
Mehmet U. Caglar1, Adam J. Hockenberry1, Claus O. Wilke1\*

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

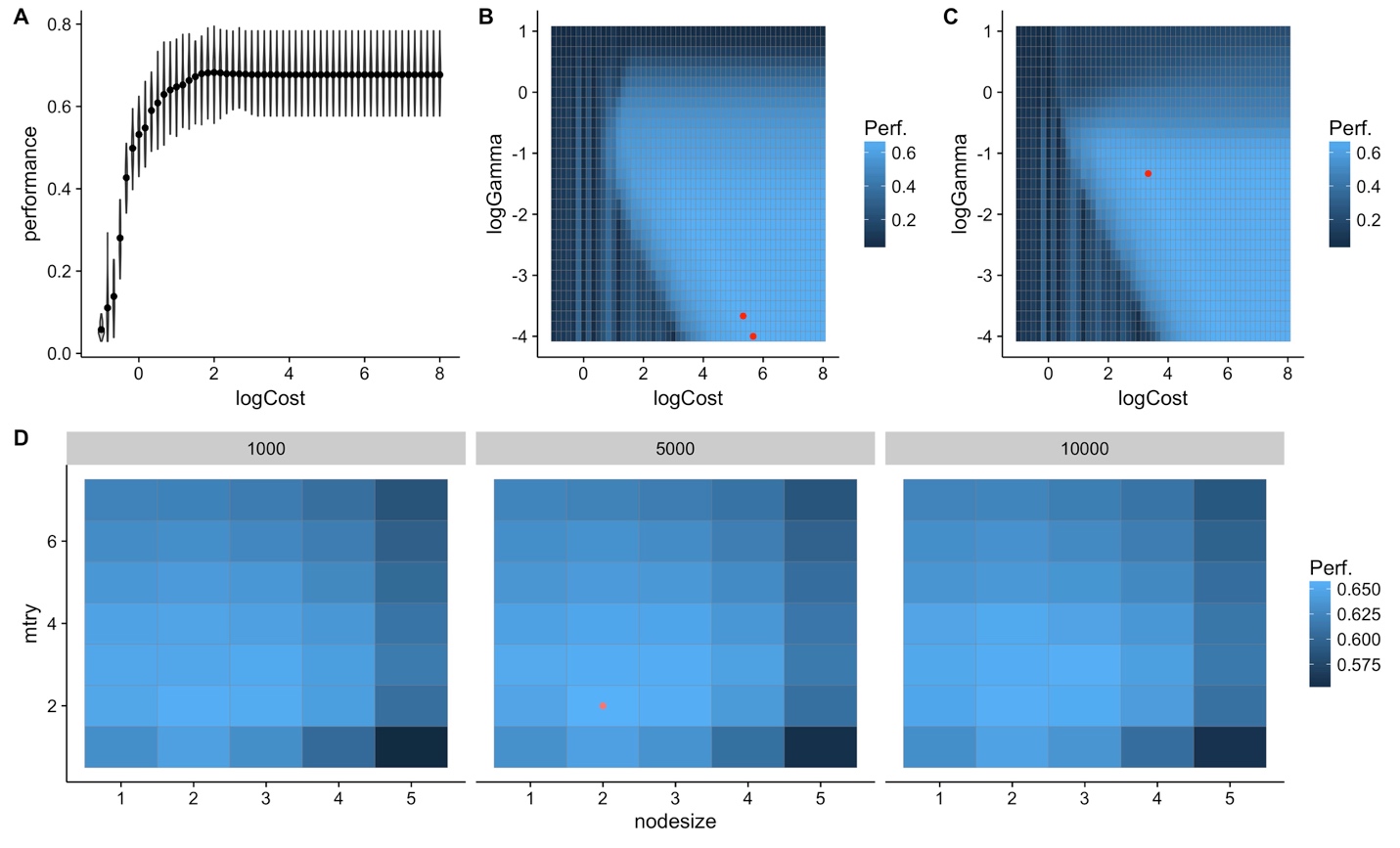
\*Corresponding author: [wilke@austin.utexas.edu](mailto:wilke@austin.utexas.edu)

**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.

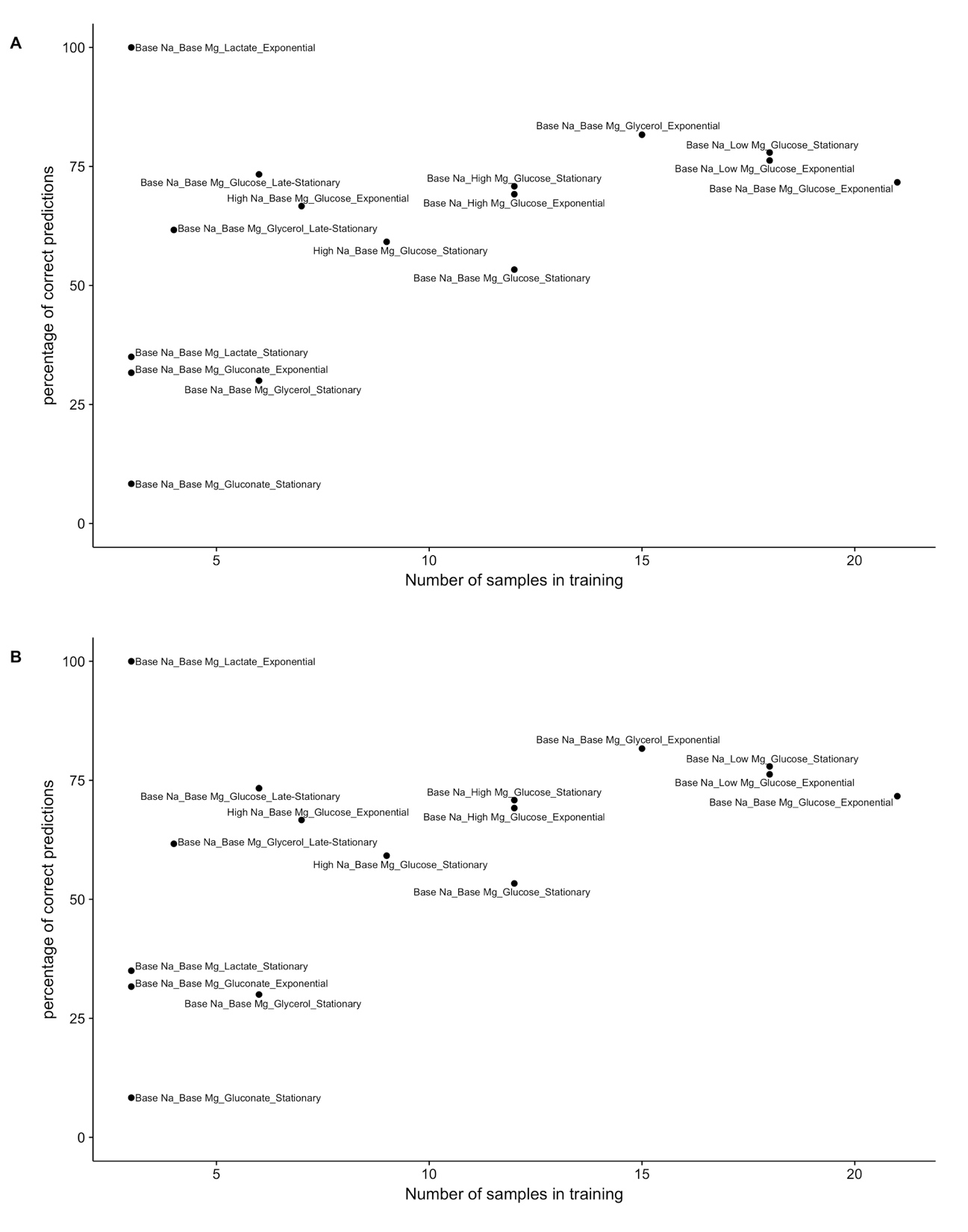
|  |  |  |
| --- | --- | --- |
|  | PC1 | PC2 |
| mRNA | *ompF* | *glpD* |
|  | *lpp* | *glpX* |
|  | *yfiA* | *garP* |
|  | *ompA* | *glpT* |
|  | *lpxC* | *glpF* |
|  | *cspA* | *glpA* |
|  | *ychH* | *menD* |
|  | *fadE* | *ygfA* |
|  | *icdA* | *mdlA* |
|  | *tufA* | *nanA* |
|  |  |  |
| Protein | *groEL* | *mglA* |
|  | *tufA* | *glpD* |
|  | *gapA* | *ydcS* |
|  | *fusA* | *fadA* |
|  | *dnaK* | *glpQ* |
|  | *metE* | *fadE* |
|  | *rpoC* | *fadB* |
|  | *rpoB* | *cspA* |
|  | *cysK* | *araF* |
|  | *eno* | *acs* |
|  |  |  |



**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, 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.



**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.

**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.



**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).



**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. In this figure average of the diagonal line is 44.1% and multi class macro F1 score is 0.43.



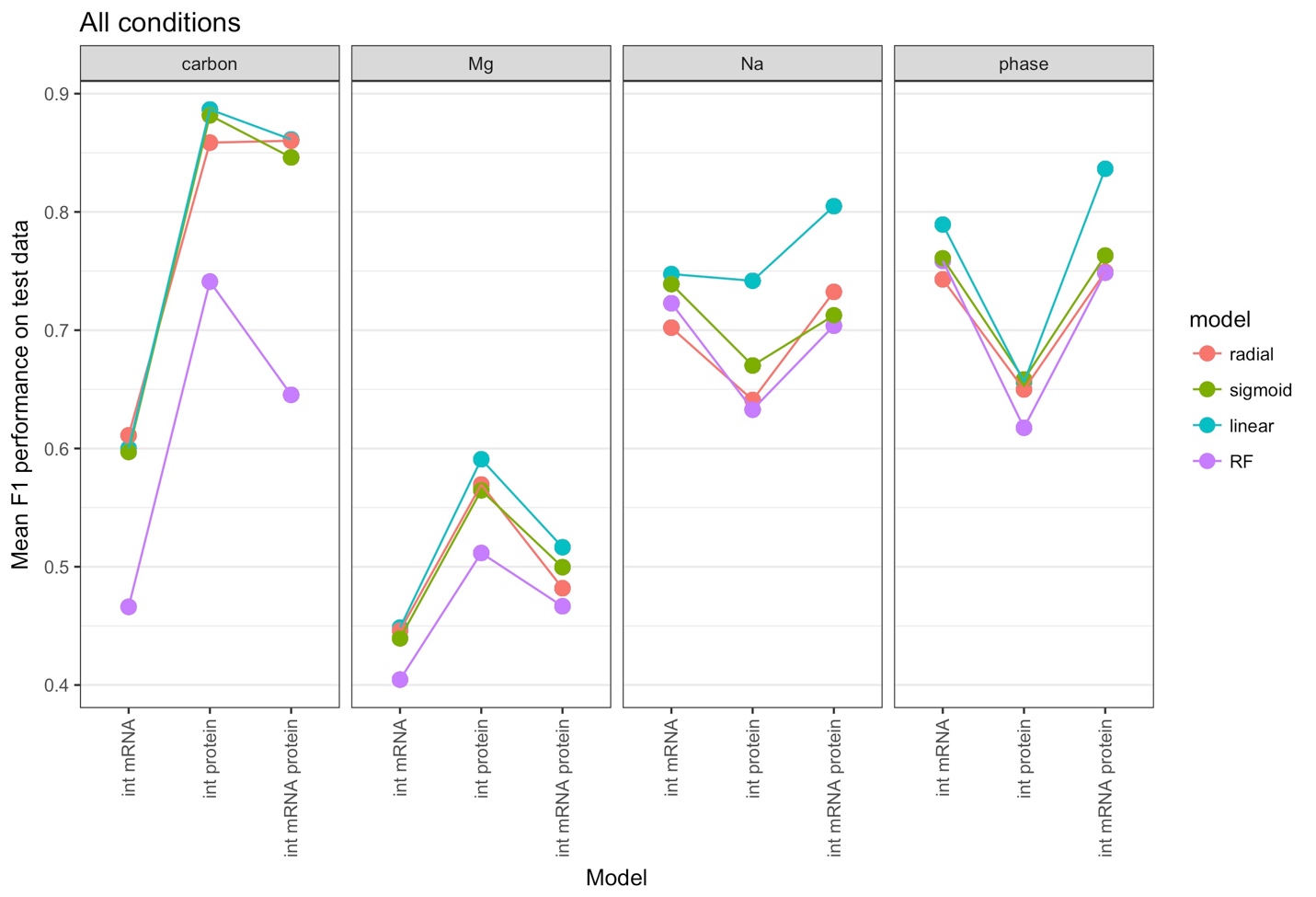
**S6 Fig.** Prediction accuracy for specific growth conditions for intersection protein data.Rows represent true conditions and columns represent predicted conditions. The numbers in the cells and the shading of the cells represent the percentage (out of 60 independent replicates) with which a given 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. In this figure average of the diagonal line is 52.3% and corresponding multi class macro F1 score is 0.53.

****

**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. In this figure average of the diagonal line is 56.1% and corresponding multi class macro F1 score is 0.57.



**S8 Fig.** Prediction accuracy for univariate predictions using intersection mRNA and intersection protein data, as in the main text Figure 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 Mg2+ levels from mRNA abundances. (F) Prediction of Mg2+ levels from protein abundances. (G) Prediction of Na+ levels from mRNA abundances. (H) Prediction of Na+ levels from protein abundances.



**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 Mg2+ levels, and mRNA abundances are more predictive for Na+ levels and growth phase.