iScience, Volume *25*

Supplemental information

Metabolic signatures of regulation by phosphorylation and acetylation

Kirk Smith, Fangzhou Shen, Ho Joon Lee, and Sriram Chandrasekaran

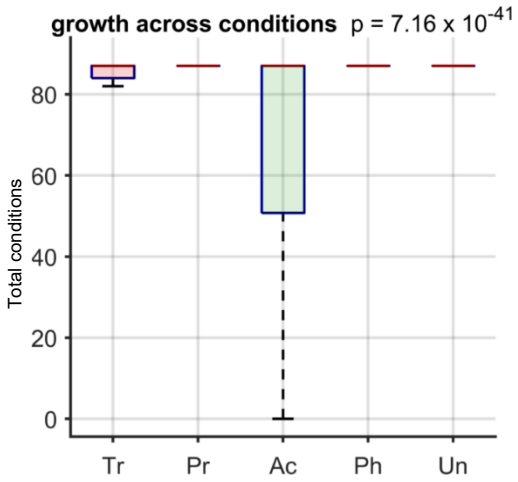


Figure S1. Distribution of regulation based on gene essentiality across 87 different conditions, Related to Figure 1. These conditions comprise 56 different carbon sources including glucose, and 31 different nitrogen sources including ammonium ions. The total number of conditions in which each gene deletion was viable was calculated. This total number was then compared between targets of each regulatory mechanism. The box plots show that acetylation preferentially regulates the genes that impact growth across the 87 conditions. The box plot whiskers extend to the 99.3rd percentile of each distribution. The ANOVA p-value comparing the means is 7.1 x 10^{-41}.

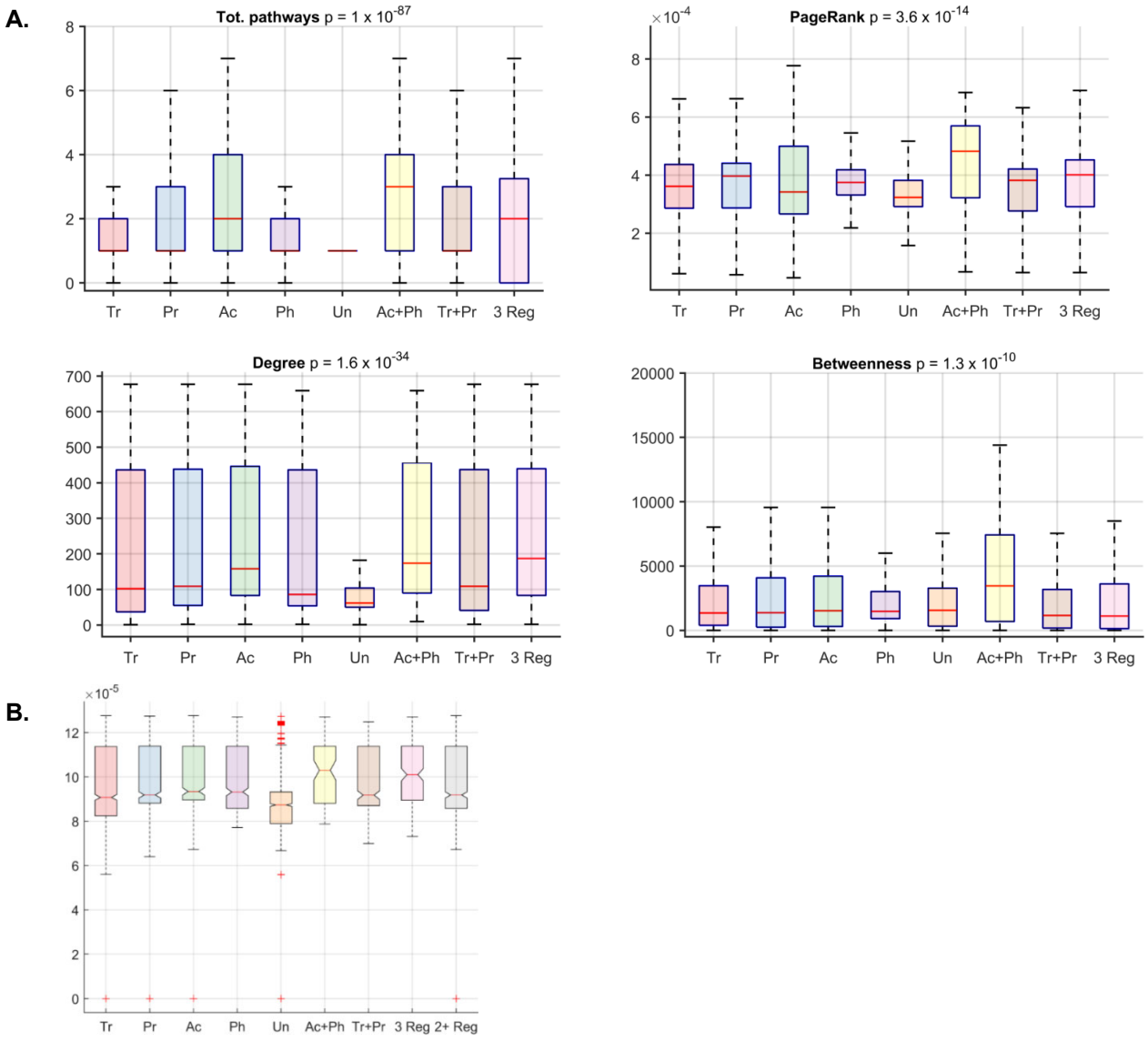


Figure S2. Distribution of regulation based on topological properties of each reaction, Related to Figure 1. **A.** Four different topological properties are shown in the box plots - the total number of annotated pathways each reaction participates (Tot. pathways), the number of times each reaction is traversed during a random walk between reactions in the network (Pagerank), the total number of connected reactions (Degree) and the number of times each reaction appears on a shortest path between two reactions (Betweenness). These show that reactions that are regulated by any mechanism have a higher connectivity compared to those that are unregulated or regulated by unknown mechanisms. Furthermore, reactions regulated by both acetylation and phosphorylation had the highest connectivity across all metrics. The ANOVA p-value comparing the means is provided in the title. (Abbreviations: regulation by both transcription and post-transcription (Tr + Pr), both acetylation and phosphorylation (Ac + Ph), at least 3 regulators (3 Reg), and Unknown regulation (Un)). **B.** Demonstration of robustness of topological analysis. Highly connected metabolites (ATP ADP AMP NADH NAD) were removed from the yeast model prior to the calculation of topological parameters. The box plots compare the properties of enzymes regulated by transcription (Tr), post-transcription (Pr), acetylation (Ac), phosphorylation (Ph), both transcription and post-transcription (Tr + Pr), both acetylation and phosphorylation (Ac + Ph), or at least 3 regulators (3 Reg). Reactions regulated by both acetylation and phosphorylation had the highest connectivity as measured by the Closeness. The ANOVA p-value comparing the means is 3e-46 for closeness, 2e-29 for degree (not shown) and 5e-15 for pagerank (not shown).

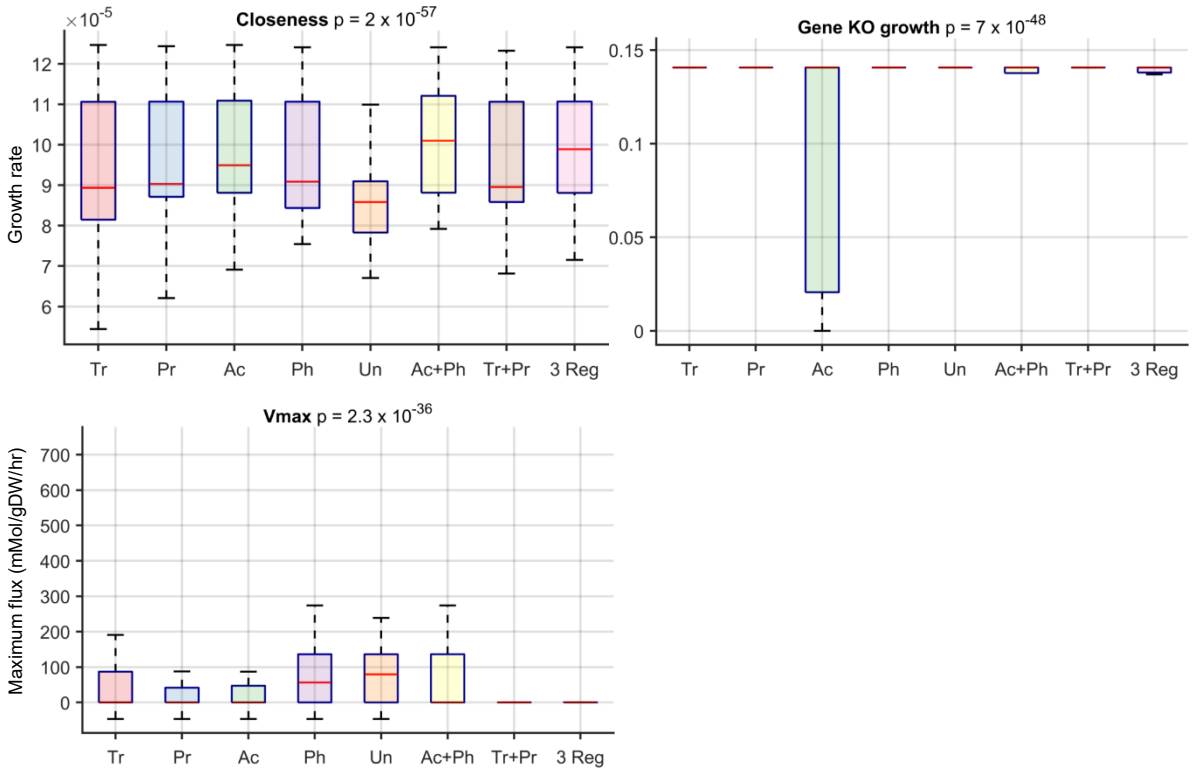


Figure S3. Properties of reactions regulated by multiple mechanisms, Related to Figure 1. The box plots compare the properties of enzymes regulated by transcription, post-transcription, acetylation, phosphorylation with those regulated by both transcription and post-transcription (Tr + Pr), both acetylation and phosphorylation (Ac + Ph), or at least 3 regulators (3 Reg). This set of combinations among regulators was chosen as both acetylation and phosphorylation are PTMs, and the transcriptome and proteome of yeast cells show significant correlation. Reactions regulated by both acetylation and phosphorylation had the highest connectivity as measured by the inverse sum of the distance from a reaction to all other reactions in the network (Closeness). Apart from connectivity, reactions regulated by two different mechanisms did not share properties of reactions regulated by each individual mechanism. For example, reactions regulated by acetylation and phosphorylation were not likely to be essential or have high maximum flux. The ANOVA p-value comparing the means is provided in the title.

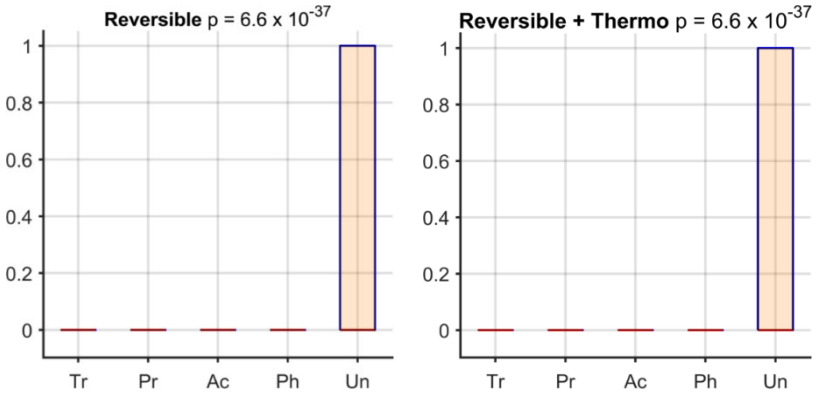
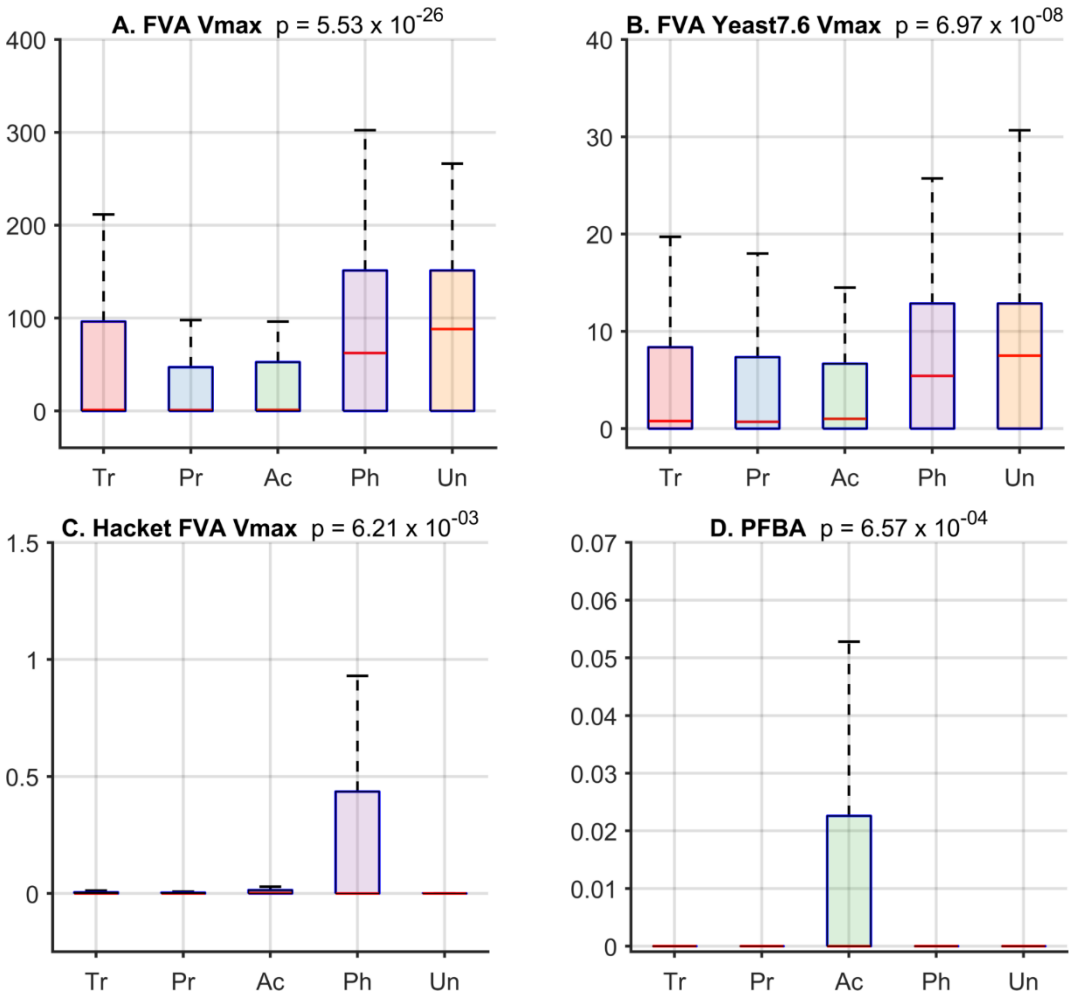


Figure S4. Distribution of regulation based on reaction reversibility, Related to Figure 1. Reversible reactions were highly likely to be not regulated by any of the four mechanisms. The left panel compares the distribution of regulation of reversible reactions based on the annotation from the Yeast 7 model (reversible reactions are set to 1 and irreversible reactions are set to 0). The panel on the right uses an updated list based on thermodynamic analysis of the Yeast metabolic model by Martinez *et al* [49].



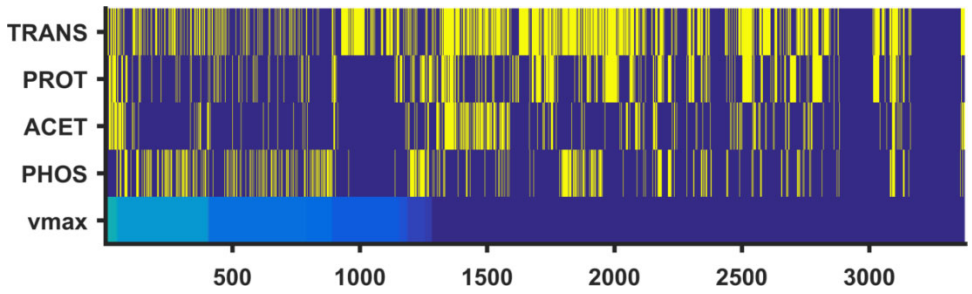


Figure S5. Distribution of regulation based on magnitude of maximum possible flux (mmol/gDW/hr) through each reaction, Related to Figure 1. The plots compare the distribution of regulation using flux calculated using various methods and models. The ANOVA p-value comparing the means is provided in the panel title of each plot. These results show that phosphorylated reactions are highly enriched among those reactions with high maximum flux. **A.** Maximum flux through each reaction was calculated using FVA using the Yeast 7 model without assuming that cells maximize their biomass (the default objective in FVA and FBA). The box plots compare the maximum flux value of reactions regulated by each mechanism. **B.** Maximum flux through each reaction was calculated using FVA without assuming that cells maximize their biomass using the Yeast 7.6 model (Yeast 7 model was used for all analyses). **C.** The flux through the model was first fit to the experimentally inferred flux data from Hackett *et al*[21]. The maximum flux through all reactions was then determined using FVA. **D.** The flux through each reaction was inferred from Parsimonious FBA (PFBA). Note that PFBA does not provide the maximum flux but the

flux value that minimizes the sum of flux through all reactions while maximizing the biomass objective. Hence it does not reveal any futile cycles or redundancy in the network. **E.** The heatmap shows the distribution of regulation based on magnitude of maximum possible flux (Vmax) through of each reaction. Reactions are sorted based on Vmax inferred from FVA. The columns correspond to each reaction-gene pair. Those that are regulated by each mechanism are shown in yellow, while those that are not regulated by a specific mechanism are in blue.

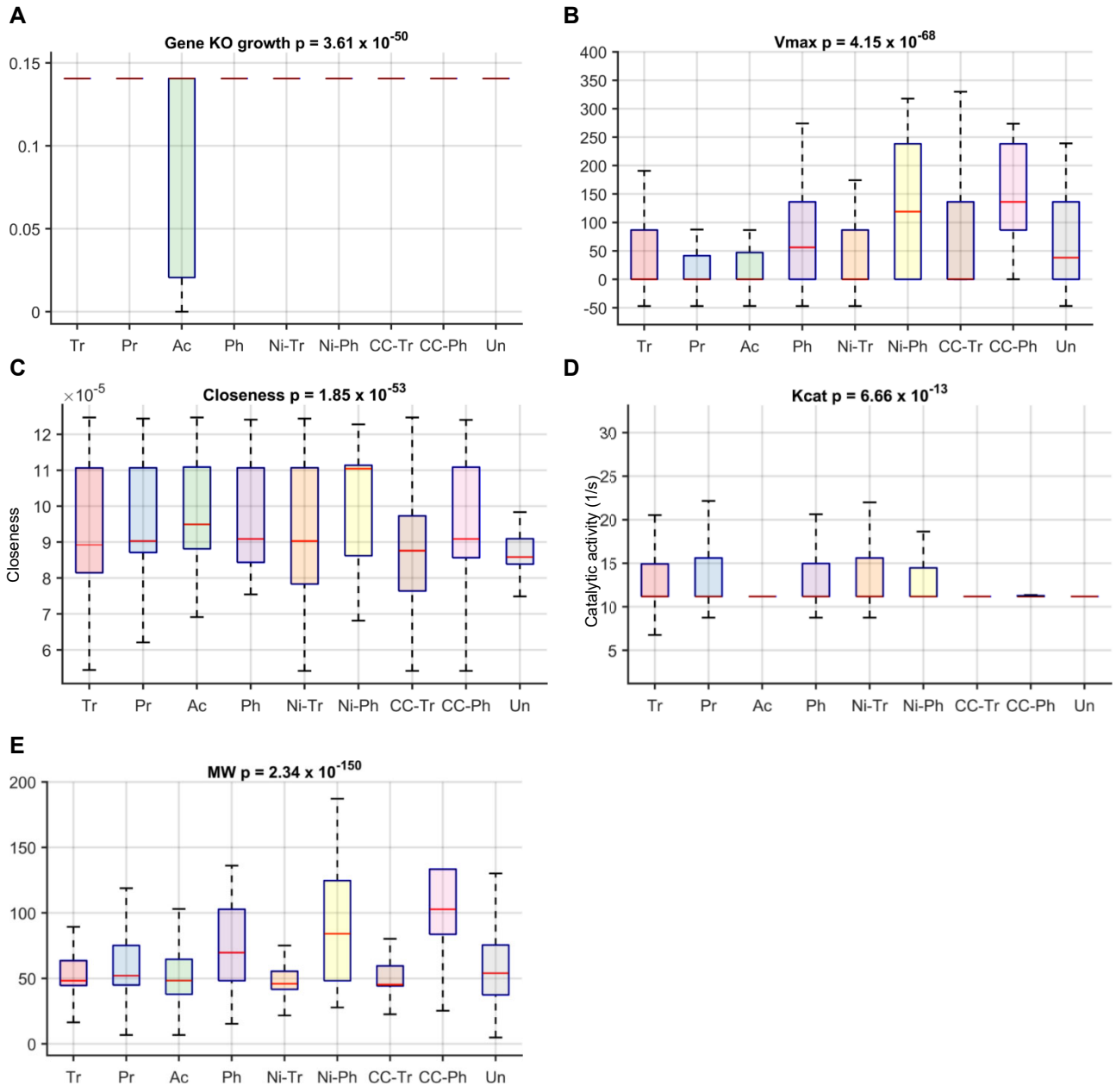


Figure S6. Comparison of the properties of enzymes in yeast regulated by each mechanism during the cell cycle (CC-Tr, CC-Ph) and nitrogen starvation (Ni-Tr, Ni-Ph), Related to Figure 1. Data from stationary phase conditions (transcription (Tr), post-transcription (Pr), acetylation (Ac), phosphorylation (Ph) or Unknown regulation (Un)) are shown for comparison. Similar to stationary phase, enzymes that impact growth when knocked out are likely to be acetylated **(A),** enzymes that catalyze reactions with high flux are likely to be regulated through phosphorylation in all three conditions **(B),** enzymes that are highly connected are likely to be regulated by one of the four mechanisms **(C).** No consistent difference across datasets was observed in regulation based on the enzyme catalytic activity (kcat) of the target enzyme **(D)** and enzymes regulated by phosphorylation on average tend to have high molecular weight **(E)**. The Anova p-value comparing the differences in means is shown in the title.

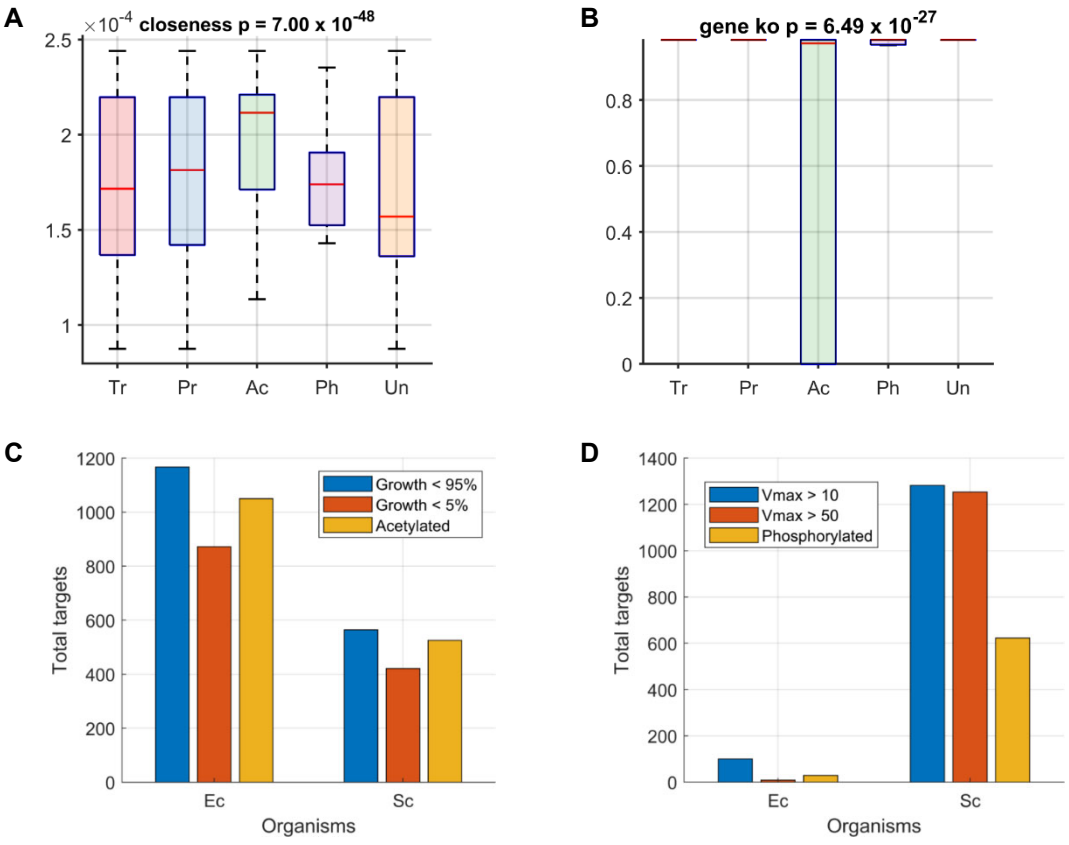


Figure S7. Comparison of properties of enzymes in E. coli regulated by each mechanism, Related to Figure 2. **A.** Similar to yeast, enzymes that are highly connected (i.e. high closeness) are likely to be regulated. **B.** Similar to our analysis in Figure 2A, which showed using the entire set of acetylated proteins the association between acetylation regulation and growth impacting enzymes, this figure shows that the subset of acetylated proteins regulated by the deacetylase cobB also show the same trend with reactions that impact growth when knocked out are highly likely to be acetylated and regulated by cobB. The Anova p-value comparing the differences in means is shown in the title. **C, D.** Comparison of total number of targets between species. Total number of regulation targets (i.e. gene-reactions) of PTMs in *E. coli* (Ec) and yeast (Sc) are compared with those that have high Vmax and are growth limiting in those species in the stationary phase condition.

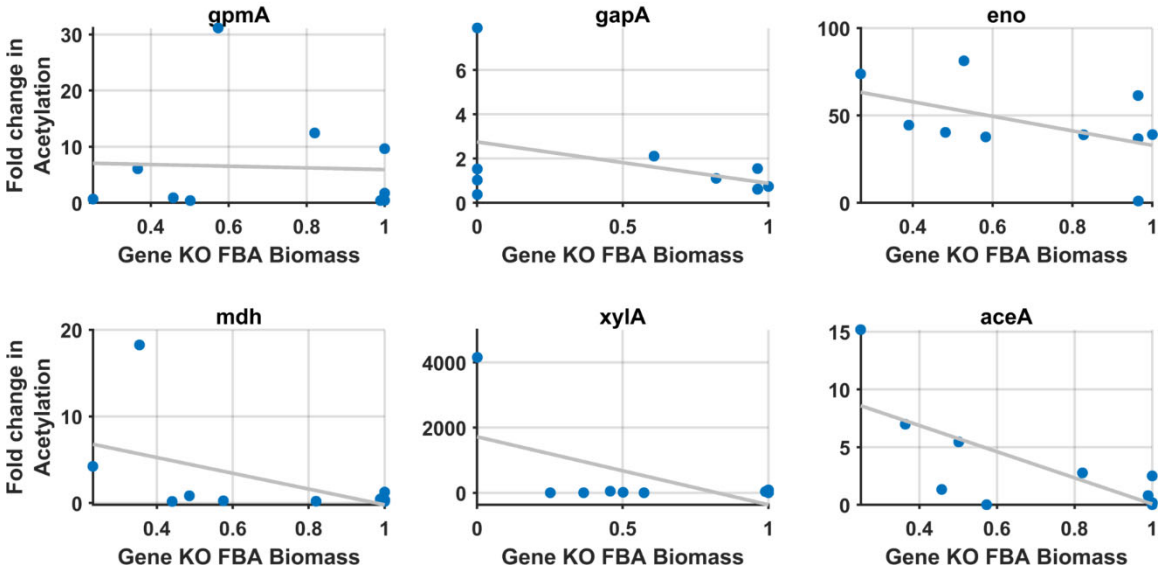


Figure S8. Condition-specific essentiality is correlated with acetylation, Related to Figure 2. The scatter plots show the association between the impact of a gene knockout on biomass from FBA with the acetylation levels of the corresponding protein in a given condition. On average, increased essentiality is associated with an increase in acetylation. All proteins with at least 2 fold change in acetylation between conditions and are part of the metabolic model are shown. The change in biomass relative to glucose is show in the x-axis. The correlations were observed even when the total absolute acetylation levels were considered instead of relative levels to proteins.

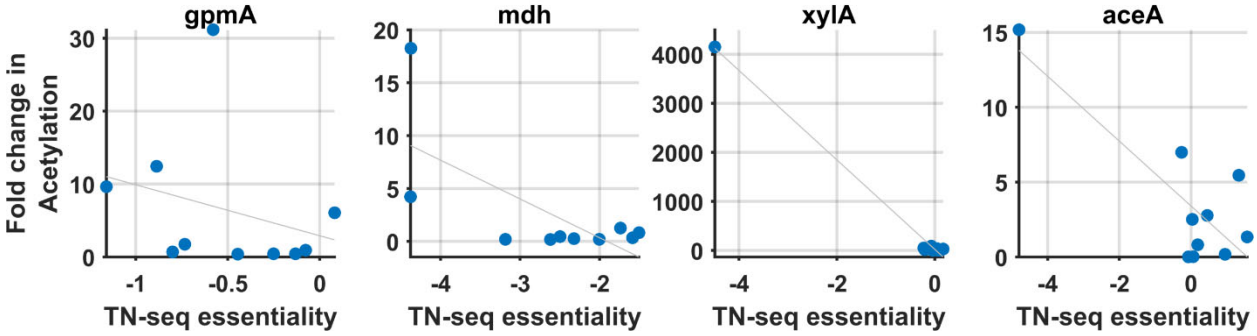


Figure S9. Condition-specific essentiality from TN-seq is correlated with acetylation, Related to Figure 2. The scatter plots show the association between the impact of a gene knockout on viability from Transposon mutagenesis screens with the acetylation levels of the corresponding protein in a given condition. All proteins in the metabolic model with available TN-seq data and acetylation data across conditions from Schmidt et al study are shown. Although FBA made false positive growth predictions for some enzymes such as XylA (Figure S8), our results were observed even with experimentally derived knockout screens, suggesting that this link between essentiality and acetylation is robust.

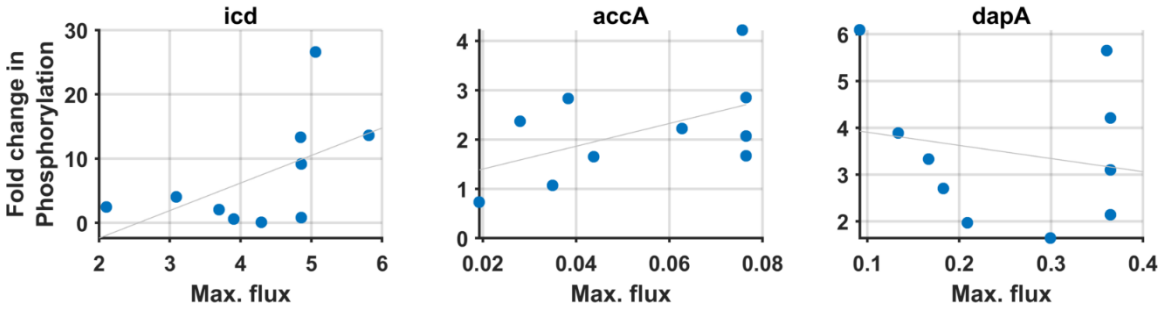


Figure S10. Correlation between maximum flux and phosphorylation levels (normalized to glucose), Related to Figure 2. All proteins that showed at least 2-fold change in phosphorylation levels between conditions are shown. This trend was observed with both the total phosphorylation levels and relative levels normalized to proteins. While in most cases a change in maximal flux or essentiality resulted in a change in regulation by PTMs (Figure 2F), there were exceptions. For example, dapA did not show this trend suggesting that other factors likely influence regulation by PTMs in a combinatorial fashion.

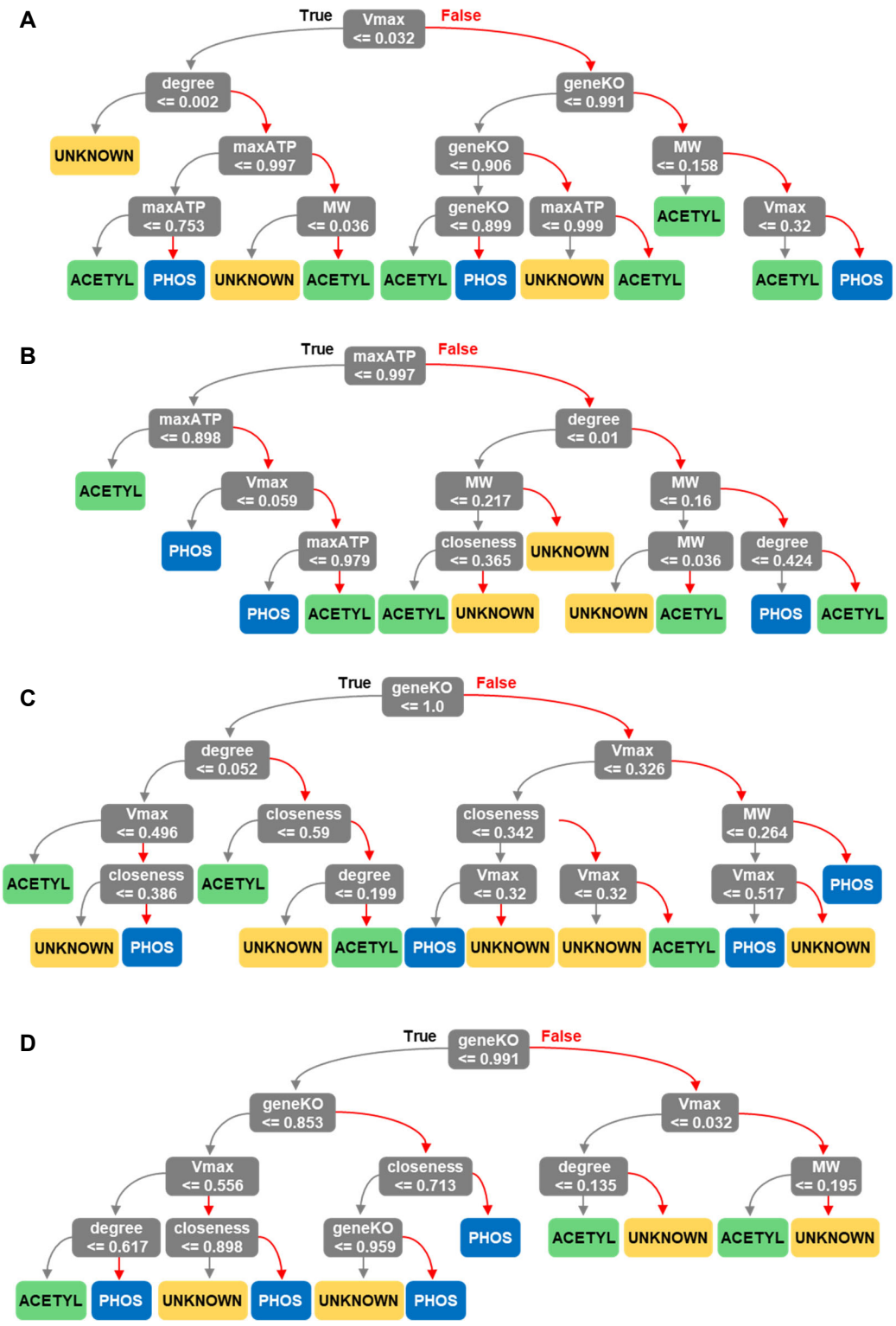


Figure S11. Representative decision trees with maximum depth of 4, Related to Figures 3-5. Single decision tree models were trained for the multi-organism **(A)**, *E. coli* **(B)**, yeast **(C)**, and mammalian **(D)** datasets. Only the top 50% most important features, as identified in the Shapley analysis, were used to train the trees.

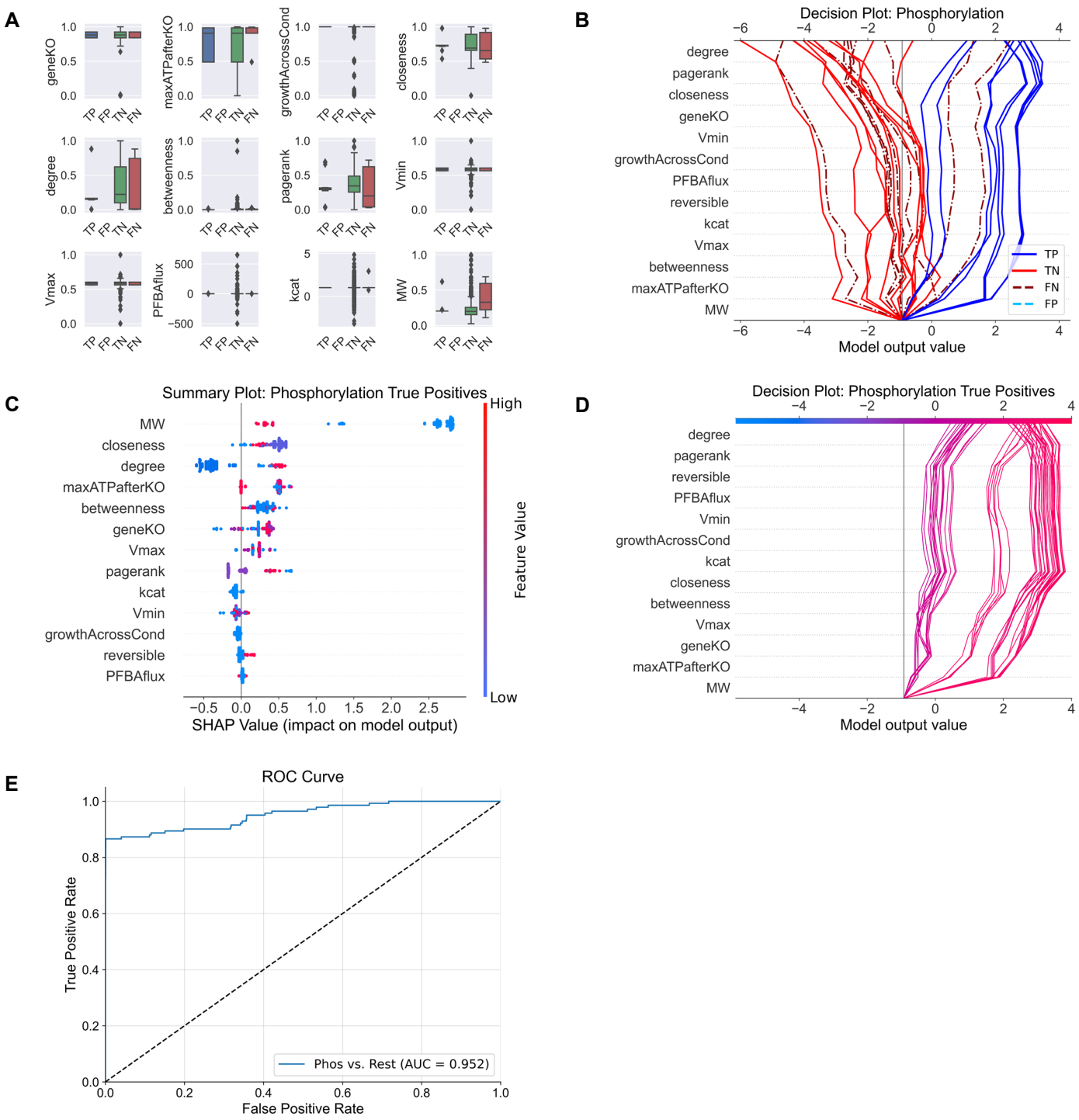


Figure S12. Analysis of model predictions on the cell-cycle phosphorylation data, Related to Figures 3-5. **A.** Feature distributions for phosphorylated gene-reaction pairs are compared between true positive (TP), true negative (TN) and false negative (FN) observations using boxplots. There were no false positives from this validation test. **B** SHAP decision plot was created for 50 random observations to compare trends between the classification groups. Values on the x-axis represent log odds of belonging to the phosphorylation class. **C** and **D.** The phosphorylated gene-reaction pairs that were correctly classified (true positives) are displayed in a SHAP summary plot (C) and decision plot (D). **E**. ROC curve for the model's phosphorylation predictions on the cell-cycle data.

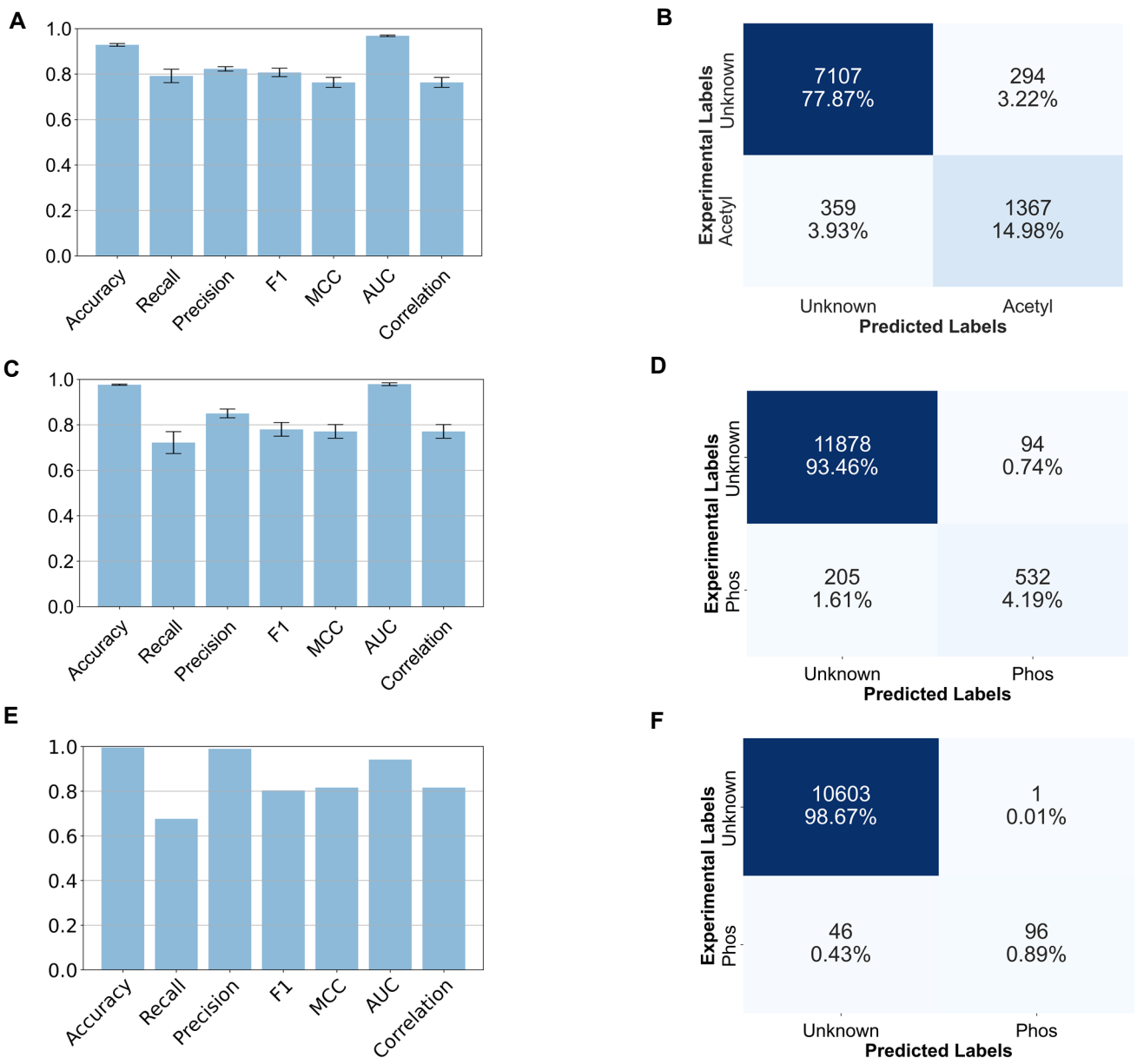


Figure S13. Binary classification models for predicting acetylation and phosphorylation separately, Related to Figures 3-5. The pipeline for training the models was identical to process used for the multi-class model. **A, B** The 5-fold cross-validation results for the acetylation model. **C, D.** The 5-fold cross-validation results for the phosphorylation model. **E, F.** The phosphorylation model was used to predict the cell-cycle validation dataset, which includes the G1, S and G2 phases. Overall, these results show that the ternary classification model outperforms the binary classification models.

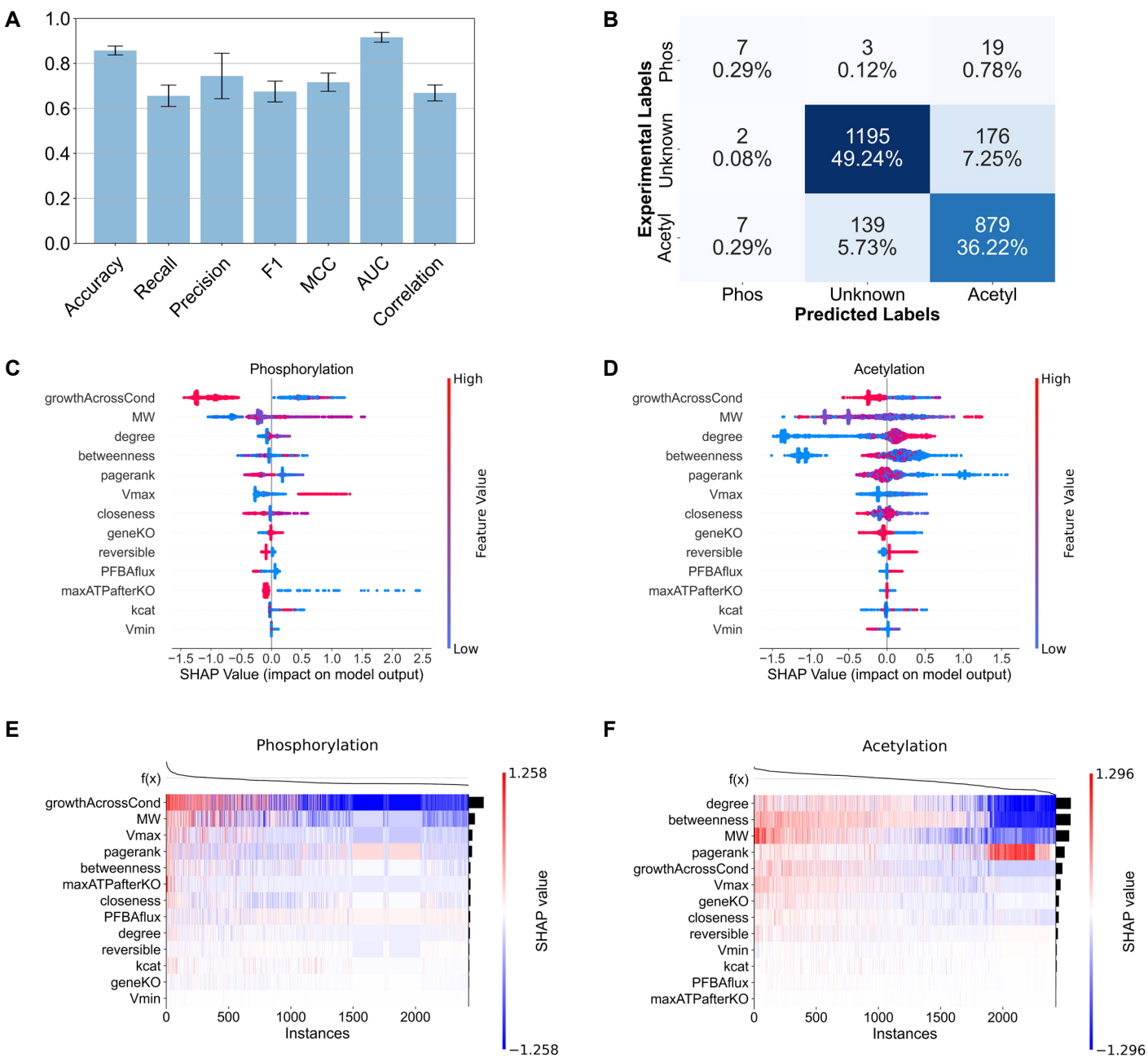


Figure S14. Organism-specific ML models – E. coli, Related to Figures 3-5. XGBoost model trained on the *E. coli* dataset. **A, B.** 5-fold cross-validation results. Bar graph shows the mean scores across the 5 folds with a 95% confidence interval. **C, D.** SHAP value summary plots for the phosphorylation and acetylation classes. **E, F.** SHAP value heatmaps for the phosphorylation and acetylation classes. Observations are clustered by the model output, f(x).

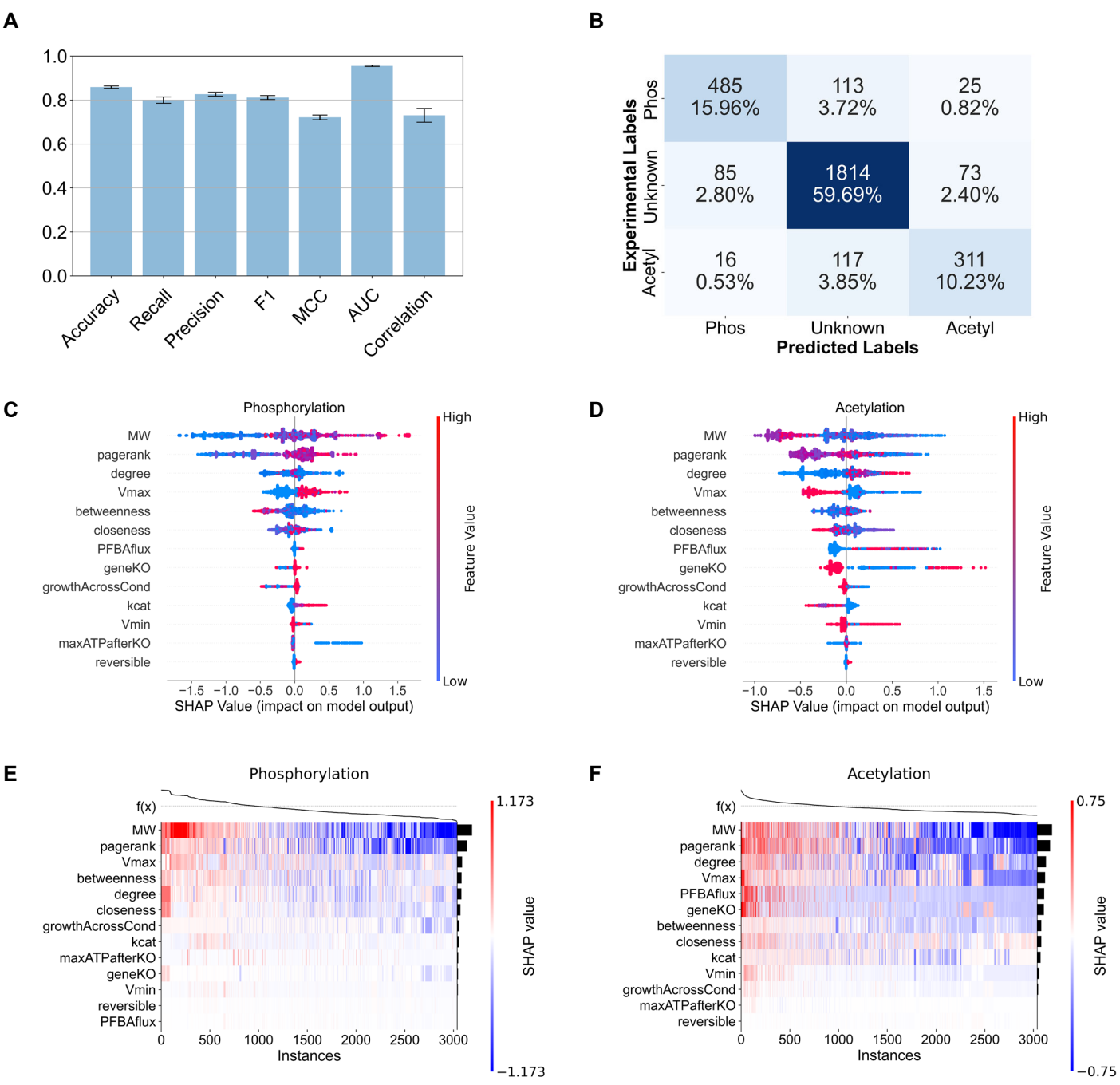


Figure S15. Organism-specific ML models – S. cerevisiae, Related to Figures 3-5. XGBoost model trained on the yeast dataset. **A, B.** 5-fold cross-validation results. Bar graph shows the mean scores across the 5 folds with a 95% confidence interval. **C, D.** SHAP value summary plots for the phosphorylation and acetylation classes. **E, F.** SHAP value heatmaps for the phosphorylation and acetylation classes.

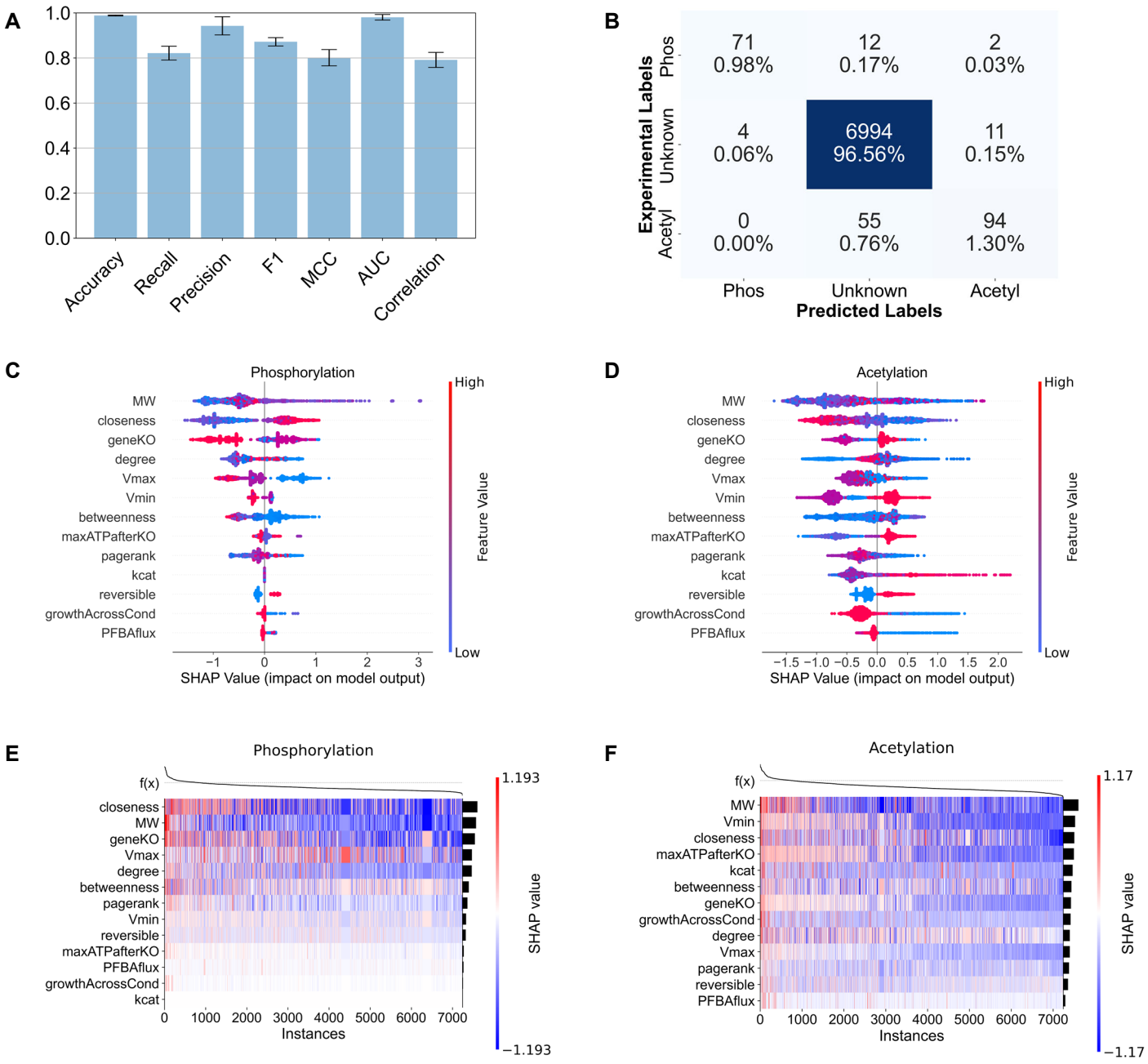


Figure S16. Organism-specific ML models – mammalian cells, Related to Figures 3-5. XGBoost model trained on the mammalian dataset. **A, B.** 5-fold cross-validation results. Bar graph shows the mean scores across the 5 folds with a 95% confidence interval. **C, D.** SHAP value summary plots for the phosphorylation and acetylation classes. **E, F.** SHAP value heatmaps for the phosphorylation and acetylation classes.

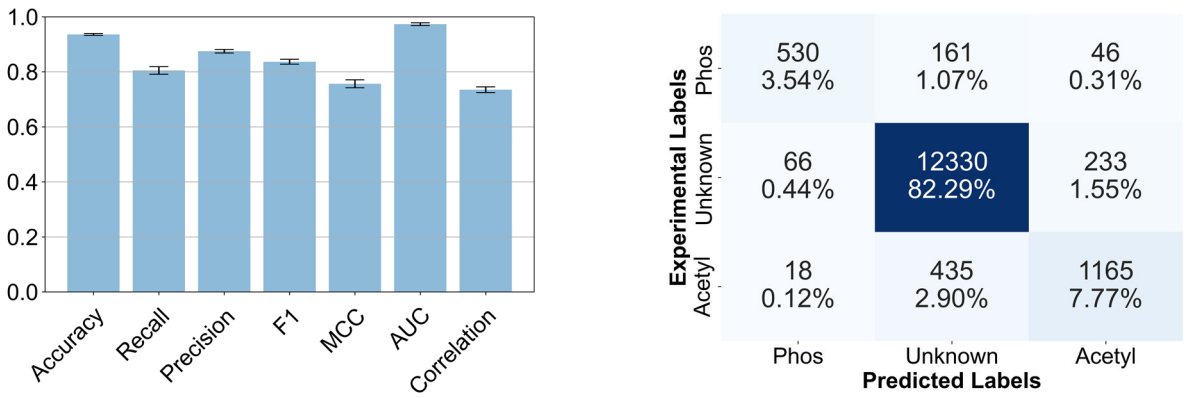


Figure S17. Impact of including organism type in the ML model, Related to Figures 3-5. 5-fold cross-validation results for XGBoost model with organism-type included in the training data. Bar graph shows the mean scores across the 5 folds with a 95% confidence interval. The organism type was added as a categorical array where a 1 designated *E. coli*, 2 for yeast and 3 for human. The cross-validation results were extremely consistent with those from the primary model, suggesting that the model's decision-making is not influenced by organism type

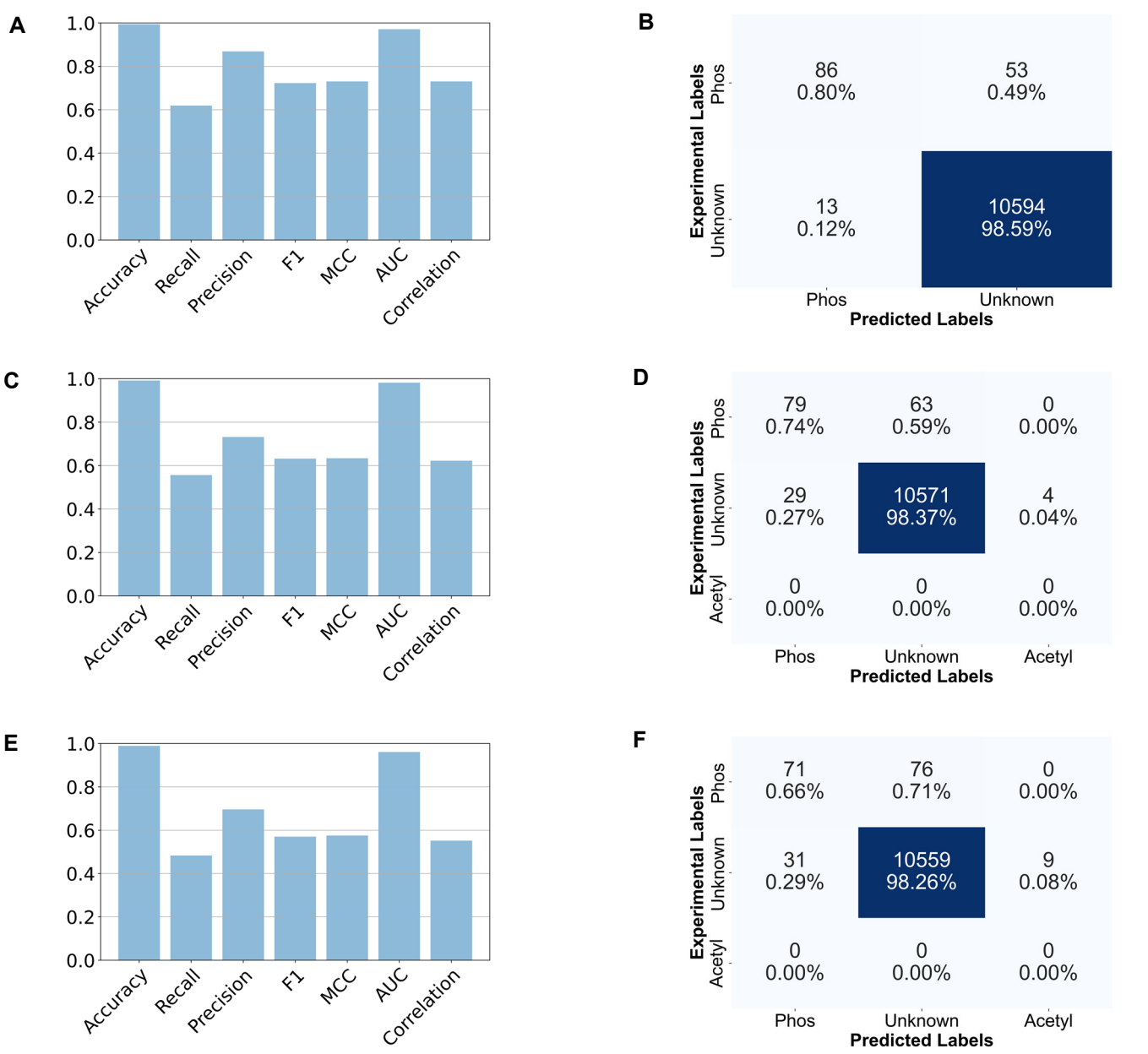


Figure S18. Impact of training on different phases of the cell cycle, Related to Figures 3-5. Models were trained by replacing the G0 cell-cycle data from the training set with the feature matrix from the remaining phases: G1, S, and G2. Each model was then used to predict the phosphorylated genes from the phases not featured in the training. These results are shown here for the G1-model **(A, B)**, S-model **(C, D)** and G2-model **(E, F)**. All three models, especially for S and G2, performed inferior to the primary CAROM-ML model in regard to this validation test. These results suggest that S and G2 conditions have a distinct phosphorylation pattern from the remaining conditions.

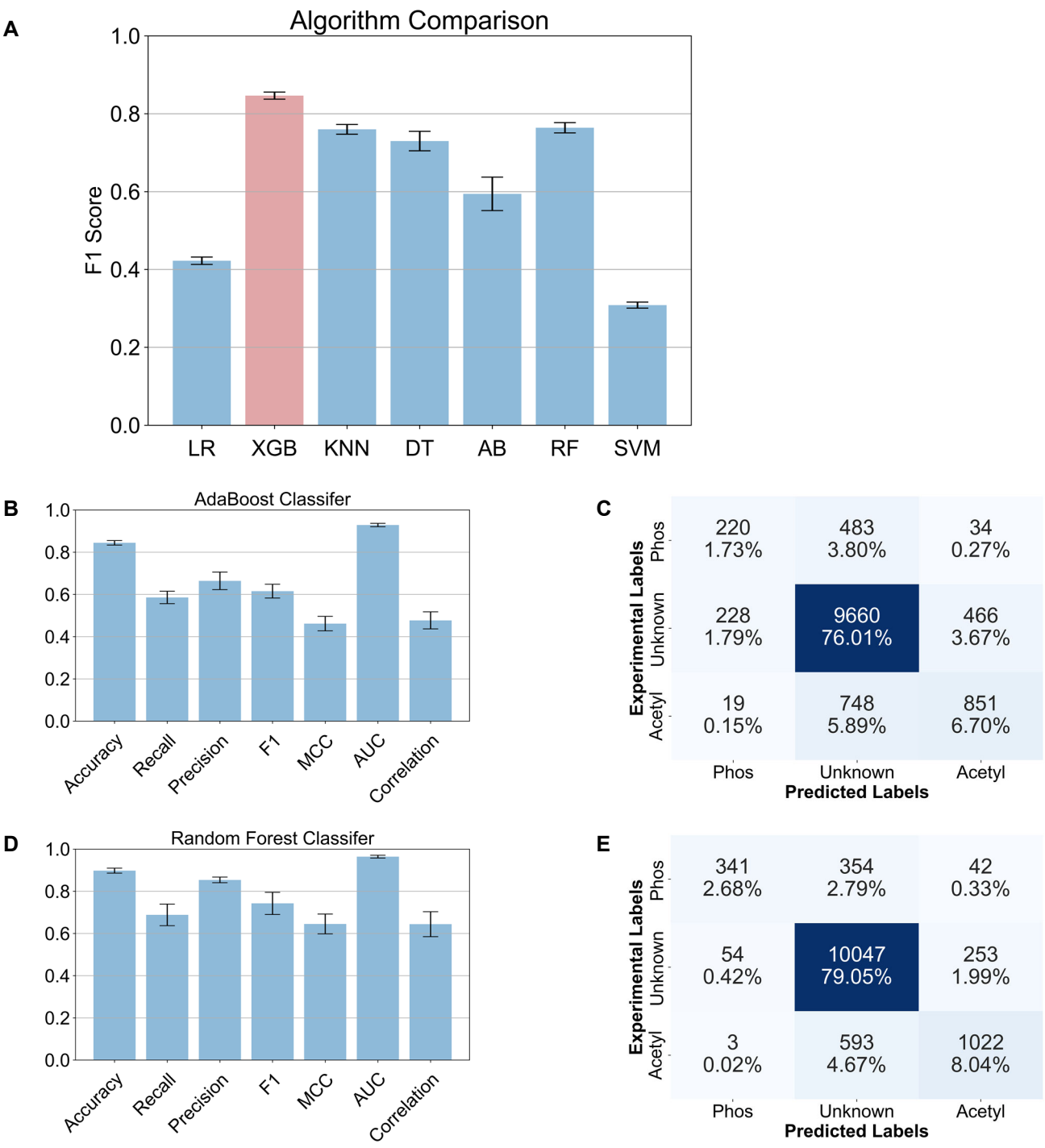


Figure S19. CAROM-ML model performance using various ML algorithms, Related to Figures 3-5. 5-fold cross-validation results were compared for various untuned algorithms, with F1 score used as the metric **(A)**. XGBoost, colored in red, had the best performance and was therefore used for the main CAROM-ML model. AdaBoost **(B, C)** and random forest **(D, E)** models were further tested by tuning their hyperparameters and performing 5-fold cross-validation. For all bar graphs, the mean scores across the 5 folds are shown with a 95% confidence interval.

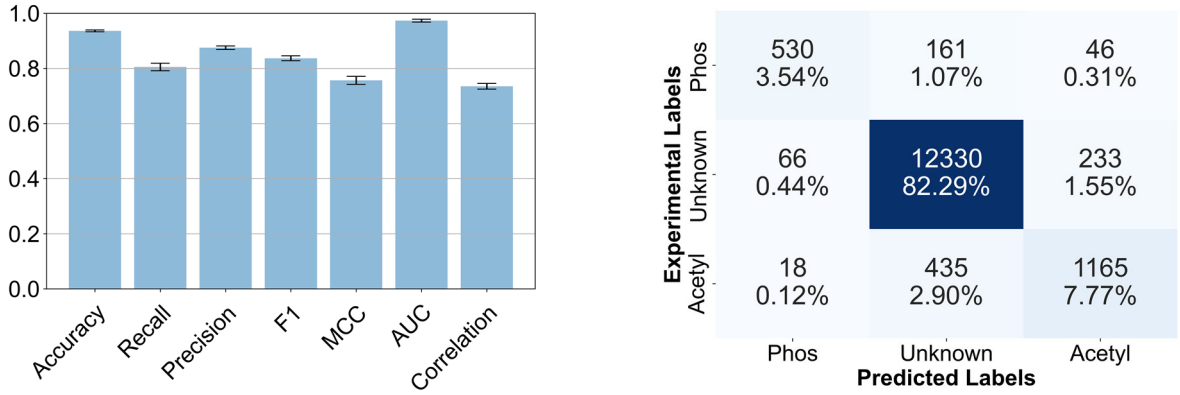


Figure S20. Impact of retaining genes that do not have evidence for phosphorylation or acetylation Related to Figures 3-5. 5-fold cross-validation results for model trained on full set of genes is shown. Bar graph shows the mean scores across the 5 folds with a 95% confidence interval. For the main CAROM-ML model, online databases were used to compile a list of enzymes that have been found to be phosphorylated or acetylated in published studies. Non-annotated enzymes were removed from the training data. Here we show the results for the model which had these non-annotated enzymes included in the training data did not differ from the model with these genes removed during the model construction.

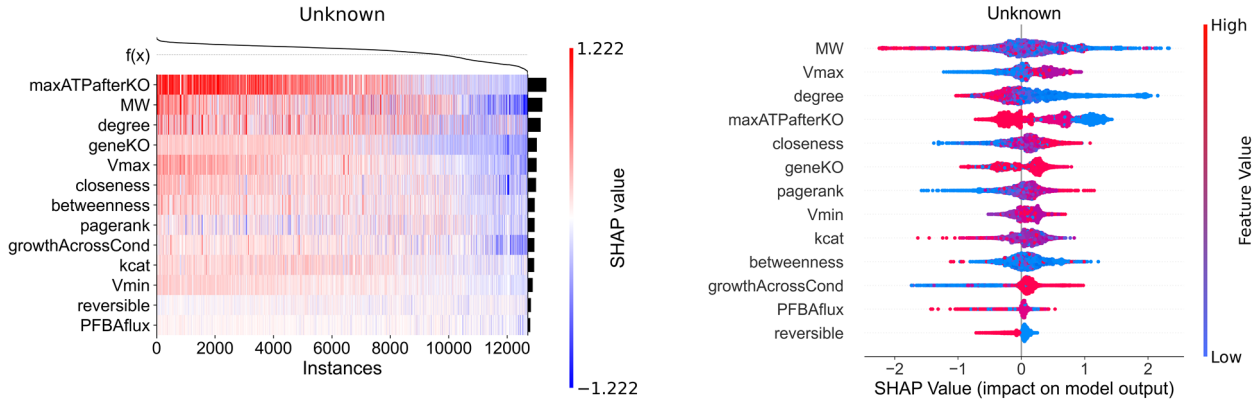


Figure S21. Interpretation of the CAROM-ML model using Shapley analysis: Unknown class, Related to Figures 3-5. Corresponding plots for the phosphorylation and acetylation classes are shown in Figure 4 in the main text.

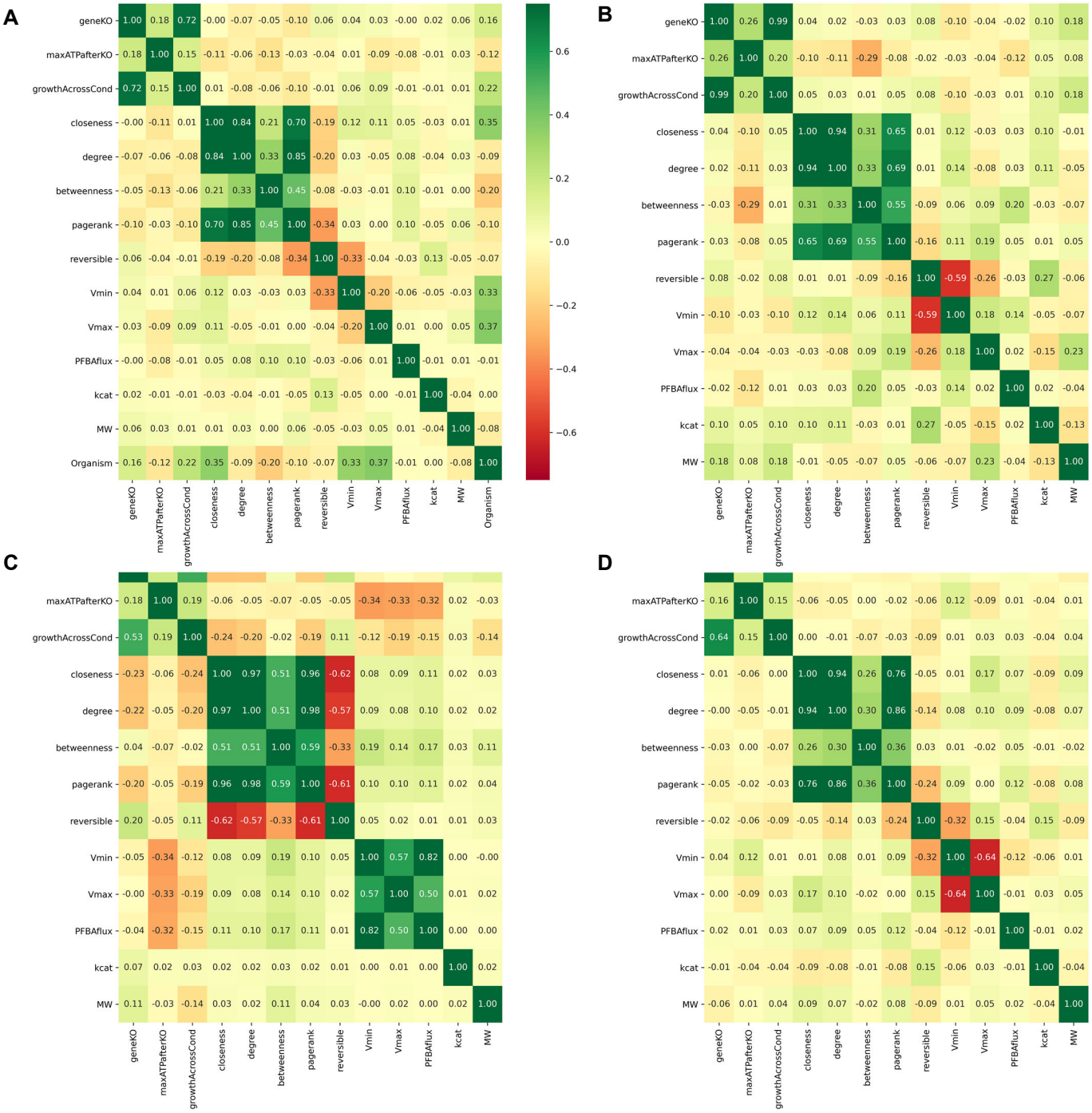


Figure S22. Correlation map of all model features, Related to Figures 3-5. Heatmap of Pearson's correlation between feature values for the following datasets: all organism types **(A)**, yeast **(B)**, *E. coli* **(C)**, and human **(D)**.

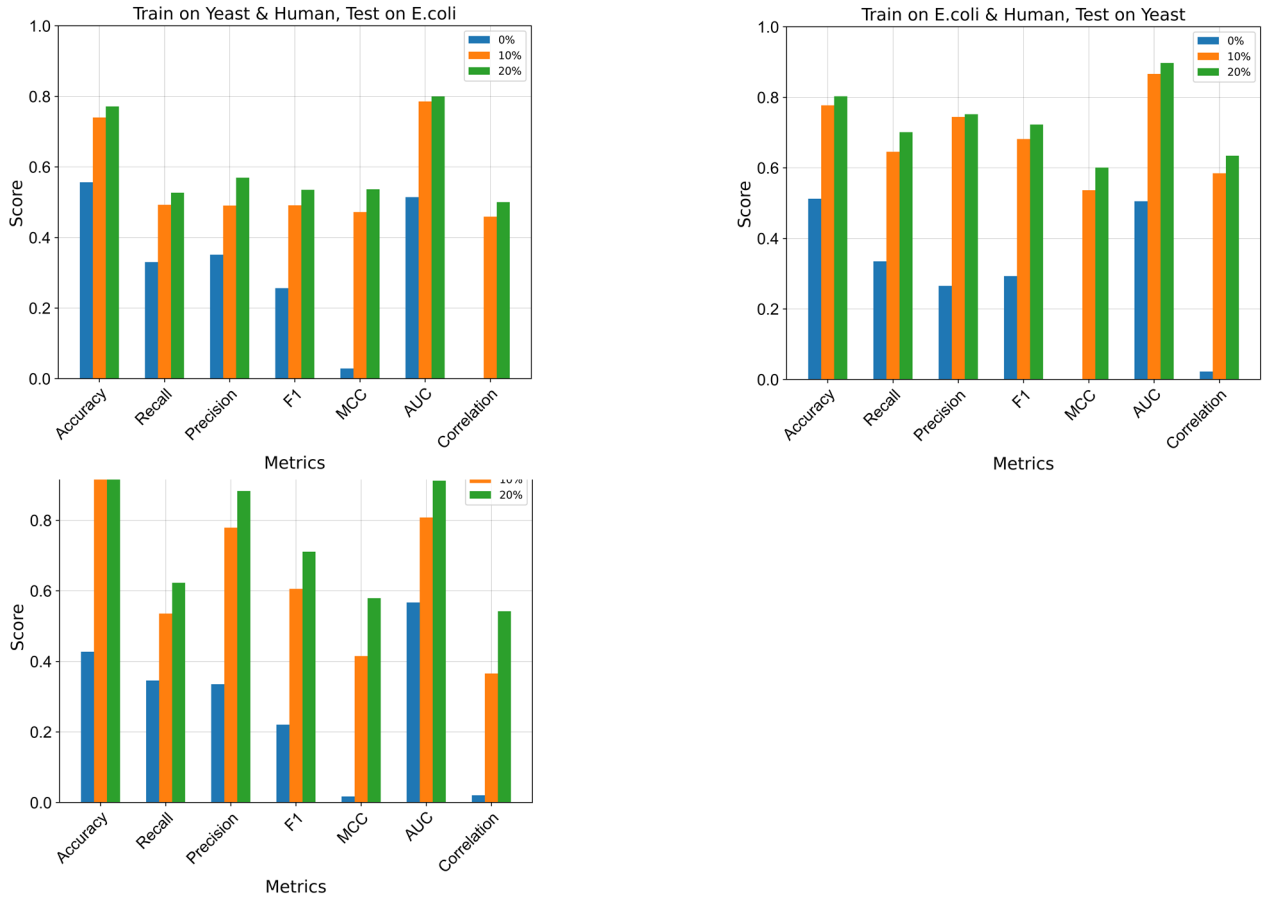


Figure S23. Predicting on unseen organisms, Related to Figures 3-5. XGBoost models were trained on the data from two organisms and used to make predictions on the third (e.g. train on E. coli and yeast, test on mammalian). Data from the test organism was moved to the training data in increments of 0%, 10% and 20%. Model performance improved significantly after including a small number of samples from the test organism in the training dataset.

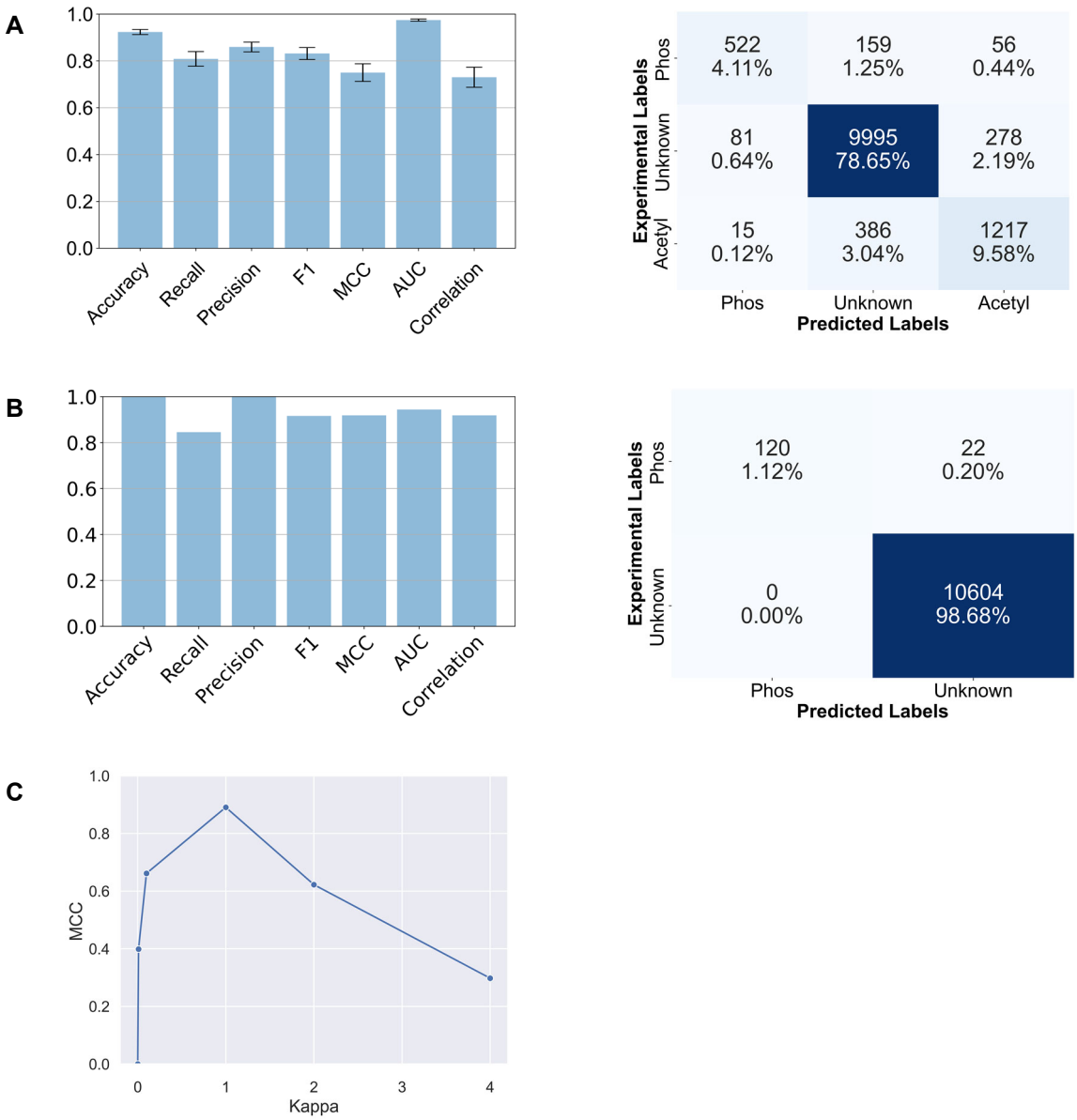


Figure S24. Impact of adjusting flux-related feature parameters on the ML results, Related to Figures 3-5. For the ML analysis, the Vmax and Vmin features were constrained to magnitudes below 100 in order to reduce the effect of unconstrained reactions and the variability across organism types. Here we show that the CAROM-ML model is robust to increasing the threshold to the 900 mmol/gDW/hr value used for the ANOVA testing. A supplementary model was trained on the *E. coli*, yeast, HeLa and G0 phase data after adjusting this threshold. **A.** Results from training the model using 5-fold cross-validation. Bar graph shows the mean scores across the 5 folds with a 95% confidence interval. **B.** The model was used to predict on the cell cycle validation dataset, which includes the G1, S and G2 phases. **C.** The cell cycle metabolic models were generated using dynamic flux analysis (DFA) with a default value of 1 for kappa, the optimization weight that is applied to the metabolomics data relative to the biomass objective. Changes to kappa therefore affect the flux- and growth-related features. The ML model's performance on the cell cycle dataset was fairly robust as kappa was incrementally changed from 1e-3 to 4, however the default value of 1 provided the best results. Setting it 0 or very low values results in the model not learning any differences between the cell cycle phases as expected. At very high values, the DFA model overfits to the metabolomics and is affected by noise in the measurement. The default value of 1 provides a good trade off in separating signal from noise.

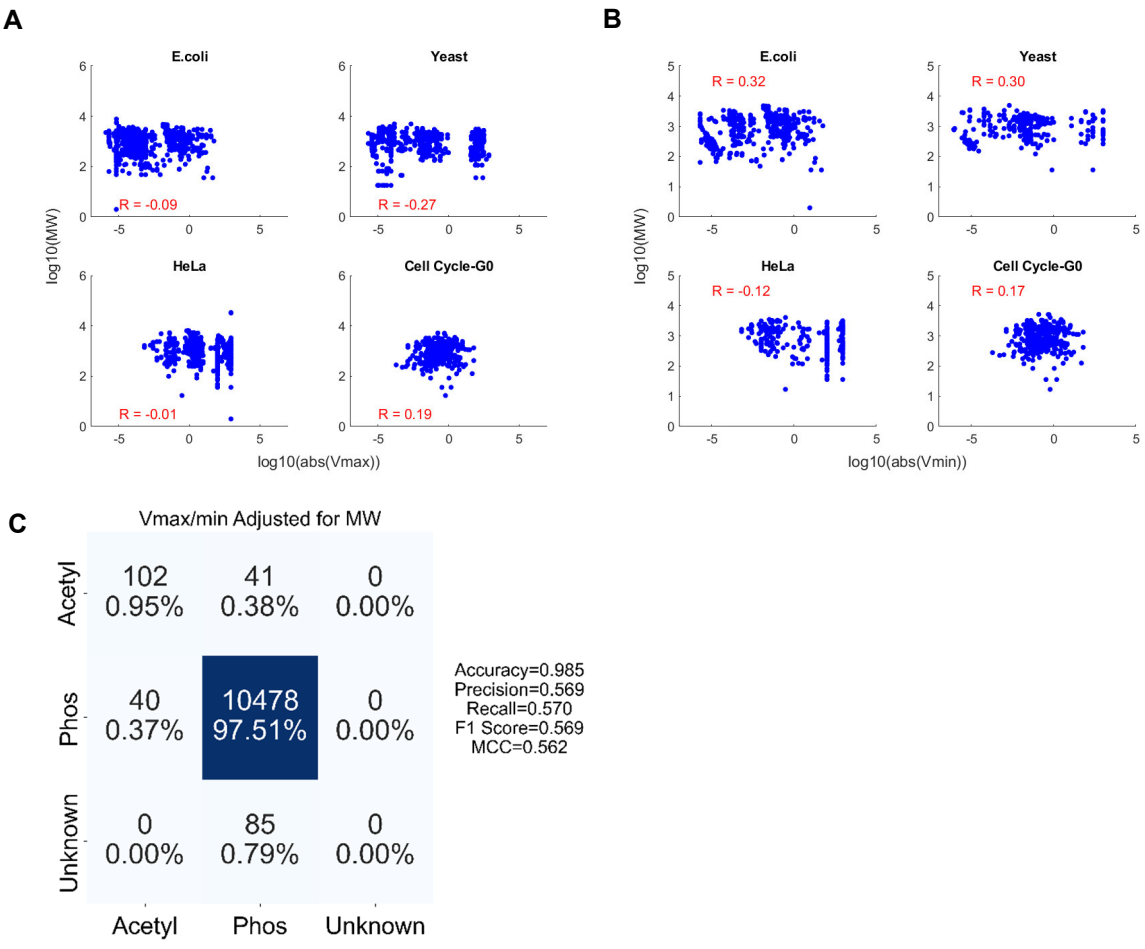


Figure S25. Relationship between flux and MW per reaction, Related to Figures 3-5. Here we address whether the flux features generated from flux variability analysis, Vmax and Vmin, are strongly correlated with the molecular weight (MW) of the metabolites present in the corresponding reactions. We did not find a significant correlation between MW of the metabolites and the predicted fluxes. For this analysis, "MW" represents the sum of MW for all metabolites in a given reaction. The plots show the relationships between **(A)** *Vmax vs. MW* and **(B)** *Vmin vs. MW* for each organism on log scales. The Spearman's correlation shown on each plot suggest there is not a consistent relationship between flux and the MW present in a given reaction. While a negative correlation is expected, in some cases we see a positive correlation. This suggests that there is not a strong relationship between the two. **C.** A separate XGBoost model was trained after adjusting the Vmax/min features for the MW per reaction by multiplying the fluxes with the MW. The model's performance slightly worsened compared to the main CAROM-ML model.

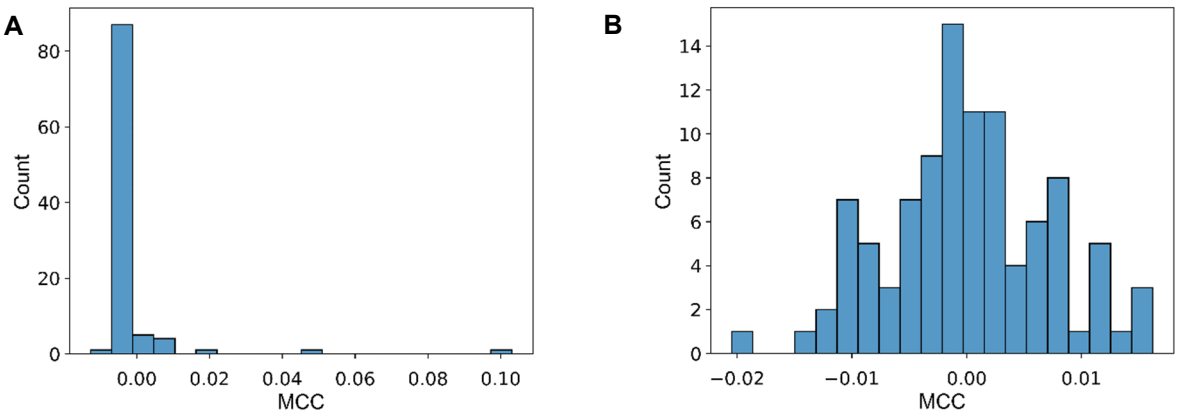


Figure S26. Random permutation models as benchmark for the CAROM-ML model, Related to Figures 3-5. We generated 100 random permutations of the class labels for the CAROM-ML training dataset. The models generated with these permutations achieved scores close to MCC=0, as expected for a random model. **A.** For each permutation, an XGBoost model was trained using the CAROM-ML feature dataset and the shuffled class labels, then used to predict on the cell cycle G1/S/G2 dataset. **B.** For each permutation, a subset of the shuffled class labels from the training dataset was used to guess the G1/S/G2 class labels.