Electronic Supplementary Material (ESI) for Molecular Omics.
This journal is © The Royal Society of Chemistry 2022

Towards inferring absolute concentrations from relative abundance in time-course GC-MS metabolomics data

Justin Y. Lee[1], Yue Han[1], Mark P. Styczynski[1]

[1] School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Supplementary Information

Table S1: Penalties used in the optimization approach of MetaboPAC

Penalty	**Description**	**Reasoning**
Mass balance	Calculate the sum of squared residuals between the inferred change in absolute concentration over time calculated from the raw relative abundance data (i.e. the change in relative abundance over time divided by the predicted response factor) and the inferred change in absolute concentration over time calculated from the stoichiometry of the system and inferred fluxes (i.e. Equation 1 in the Methods).	If the change in absolute concentration over time is very different between the two calculations (i.e. the sum of squared residuals is much greater than zero), the predicted response factors have failed to produce inferred absolute concentration and flux profiles that are consistent with mass balances in the system.
Maximum concentration	If the inferred absolute concentration for any metabolite is above 5 mM or 50 mM for synthetic and biological systems, respectively, add a penalty equal to the maximum value of all inferred concentrations.	It is reasonable to assume that for many metabolites, there can be a general *a priori* estimate for a maximum concentration that is biologically feasible, either due to limits in production or cell toxicity. Here, we use a single threshold for all metabolites, but imposing individual maximum thresholds could lead to better response factor predictions.

Correlation for mass action reaction with a single substrate and no other regulation	Calculate the correlation between the substrate metabolite and inferred target flux. The correlation is expected to be positive (because metabolites induce mass action reactions); the penalty for each reaction with no regulation other than mass action equals the calculated correlation minus one.	If a reaction is only controlled by a single metabolite, the reaction rate should increase as the concentration of the metabolite increases (assuming the reaction kinetics do not exhibit any behavior similar to substrate inhibition).
Curve fit for mass action reaction with a single substrate and no other regulation	Calculate the fit of a second-order polynomial to the substrate metabolite and target flux data. The penalty for each reaction equals one minus the adjusted R^2 of the fit (adjusted for the number of parameters).	A second-order polynomial should fit the data reasonably well if a reaction is controlled by a single metabolite (e.g. if the data is well-modeled by a Michaelis-Menten saturation curve).
Fit to BST kinetic equations	For each reaction in a system, fit the inferred absolute concentration and flux data to a BST equation [1] representing the reaction rate. Calculate the sum of squared residuals of the fit.	A generic BST kinetic equation should fit reasonably well to correctly inferred absolute concentration and flux data.
Deviation from steady-state flux distribution*	For each flux in a system, calculate the summed percentage deviation from steady state for the last 25% of timepoints. In this work, the flux distribution of the last timepoint was assumed to be the steady-state distribution due to challenges in applying flux balance analysis in synthetic or smaller pathways.	Most biological systems eventually converge to a steady state, and these flux distribution can be reasonably estimated using flux balance analysis without much *a priori* knowledge.

*The steady-state penalty was not applied to the determined system, which does not reach steady-state due to the nature of the pathway.

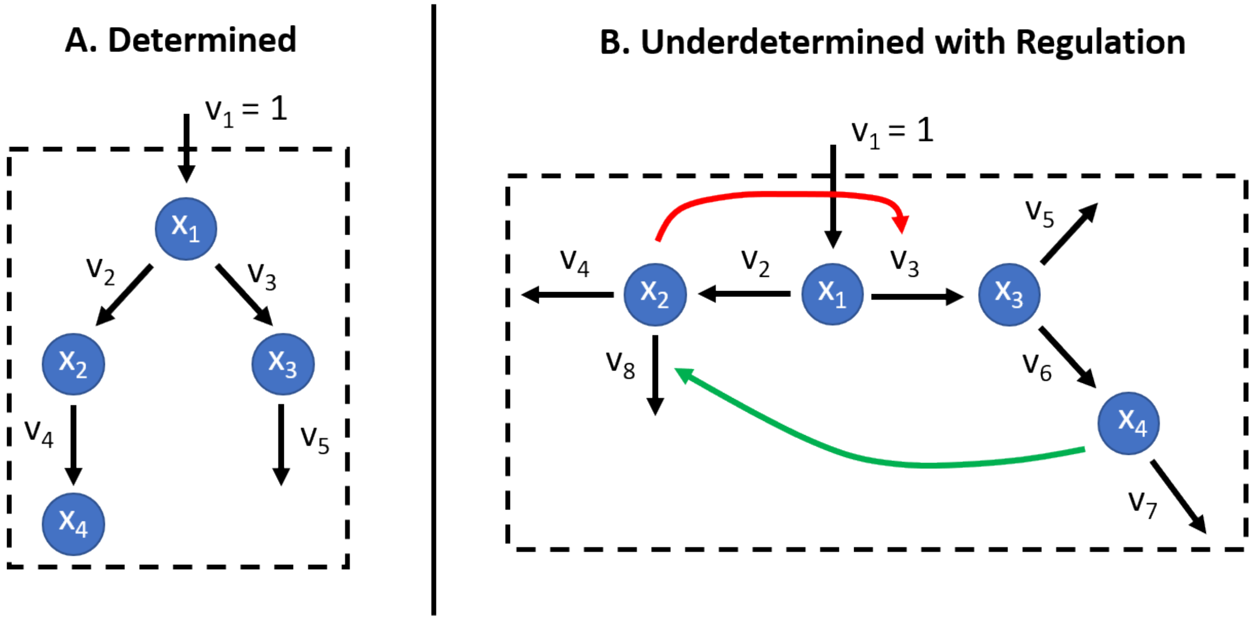

Figure S1. Synthetic systems tested with MetaboPAC. We built one determined synthetic system and one underdetermined synthetic system with regulation using Michaelis-Menten kinetics for each reaction. x_i represents the *i*th metabolite and v_j represents the *j*th flux. In both systems, flux v_1 is assumed to be constant and known.

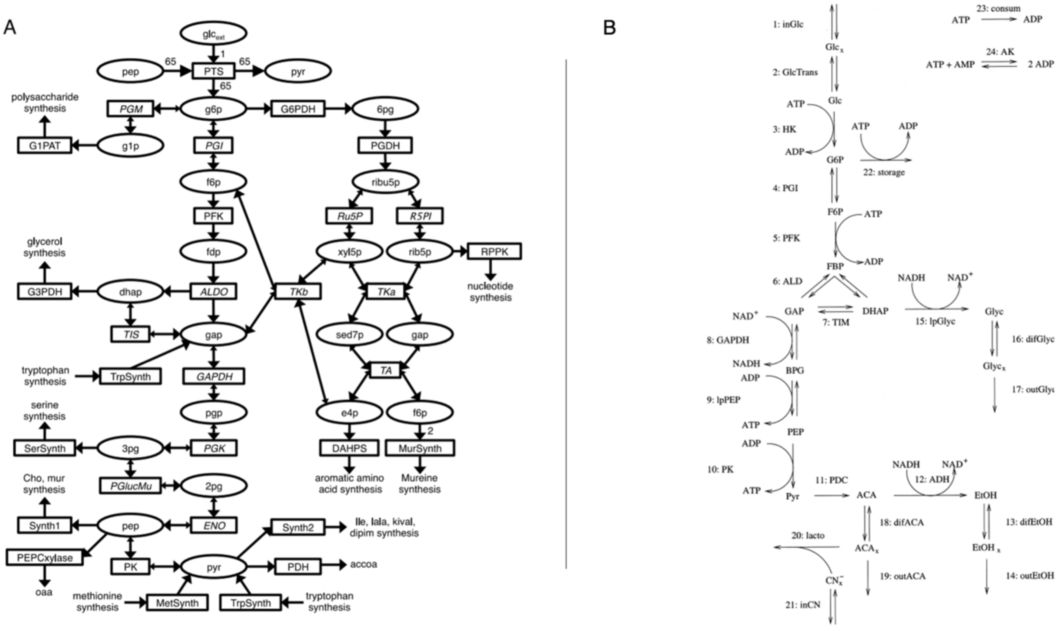

Figure S2. Biological pathways used in this study (A) central carbon metabolism in *E. coli* (adapted from Fig. 2 in [2]) (B) glycolysis pathway in *S. cerevisiae* (adapted from Fig. 2 in [3]).

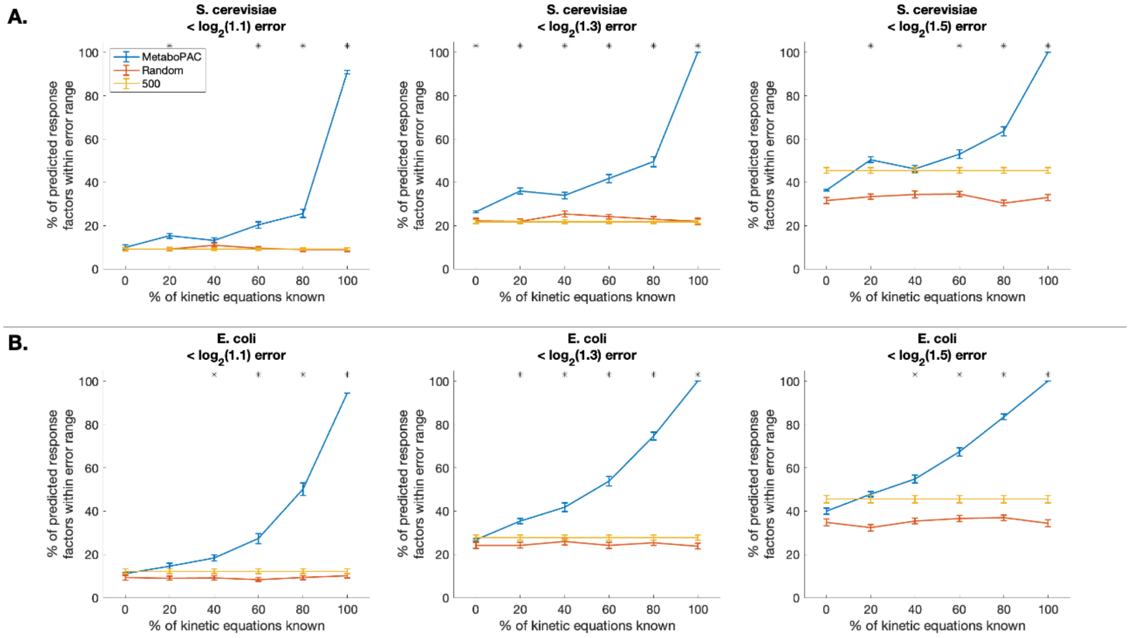

Figure S3. Percentage of response factors predicted by the combined approach without assuming a determined system in the *S. cerevisiae* and *E. coli* systems. MetaboPAC compared to random response factors and response factors of 500 for the A. *S. cerevisiae* and B. *E. coli* systems using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$. Instead of assuming sufficient known kinetics to yield a determined system such that fluxes can be estimated accurately, the Moore-Penrose pseudoinverse approach was used to estimate individual fluxes. Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 50 for different sets of true response factors and different sets of known kinetic equations for 20-80% known kinetics, n=5 for different sets of true response factors for 0% and 100%). Asterisks denote when the combined approach performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$). The combined approach outperforms the two baseline comparators when more than 40% kinetic equations are known.

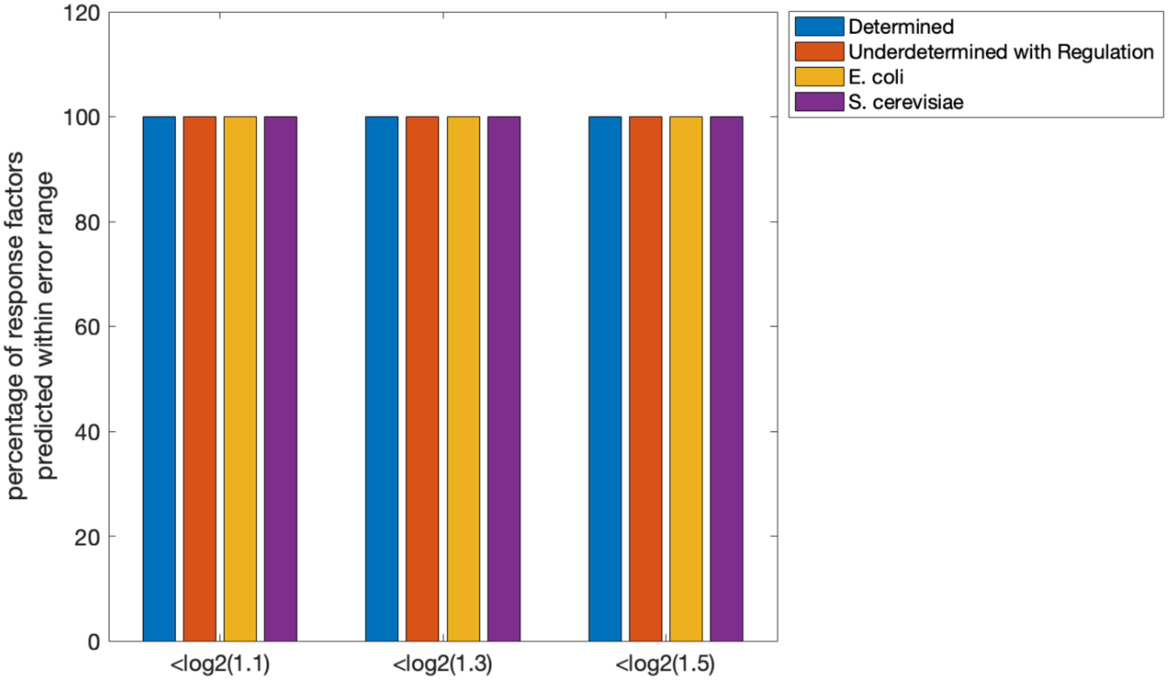

Figure S4. Kinetic equation approach performance on noiseless data at 100% known kinetics with no discretization error. To assess the impact of discretization error on the two biological systems, the change in concentrations is derived directly from *in silico* simulated fluxes rather than from calculating differences in concentrations at two adjacent timepoints (i.e., finite different approximations). The change in metabolite concentration is found by multiplying fluxes by the stoichiometric matrix and then multiplying by true response factors to yield change in relative abundance with no discretization error. All response factors can be predicted accurately in this case.

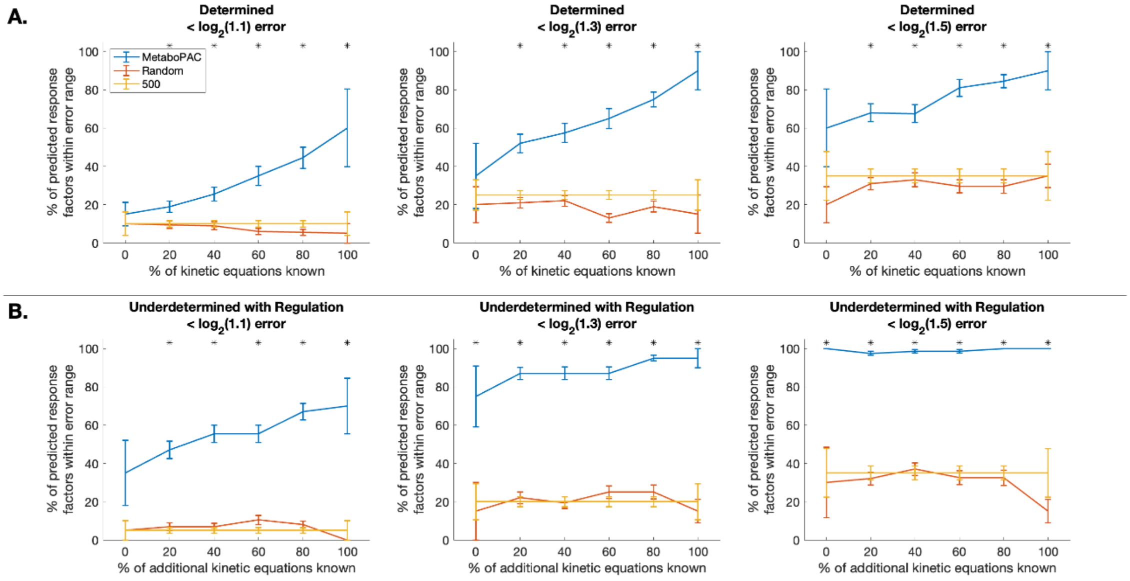

Figure S5. Performance of the optimization approach on noiseless data for synthetic models. The optimization approach is compared to random response factors and response factors of 500 for the A. determined and B. underdetermined with regulation synthetic models using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$. Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 50 for different sets of true response factors and known kinetic equation terms for 20-80% known kinetics, n=5 for different sets of true response factors for 0% and 100%). Asterisks denote when the optimization approach performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with $\alpha = 0.05$).

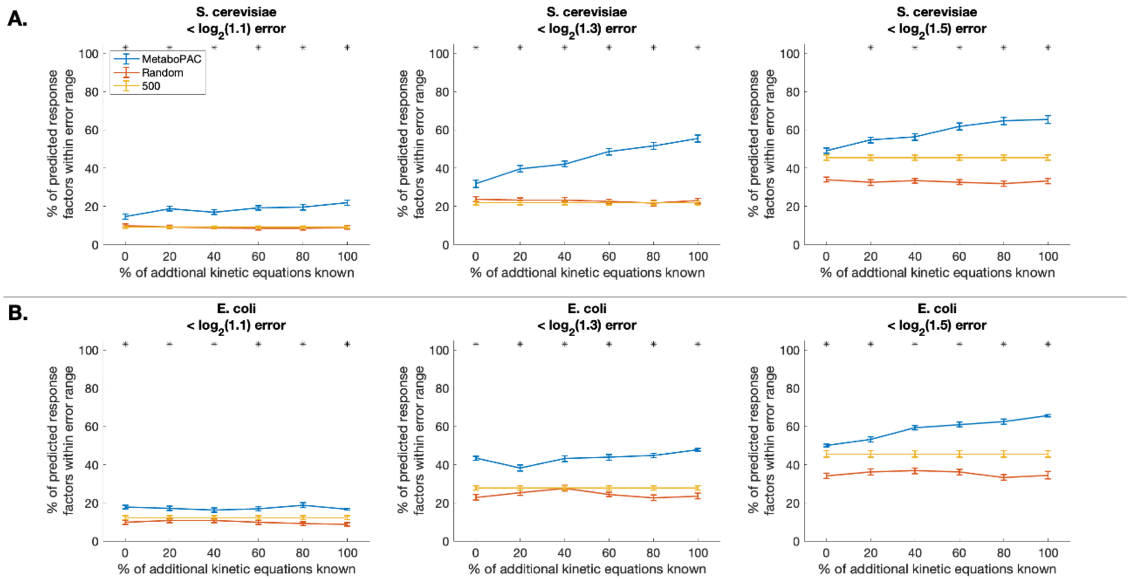

Figure S6. Performance of the optimization approach on noiseless data for biological models. The optimization approach is compared to random response factors and response factors of 500 for the A. *S. cerevisiae* and B. *E. coli* models using error ranges of $\log_2(1.1)$, $\log_2(1.3)$, and $\log_2(1.5)$. Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n = 50 for different sets of true response factors and different sets of known kinetic equation terms for 20-80% known kinetics, n=5 for different sets of true response factors for 0% and 100%). Asterisks denote when the optimization approach performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with α = 0.05).

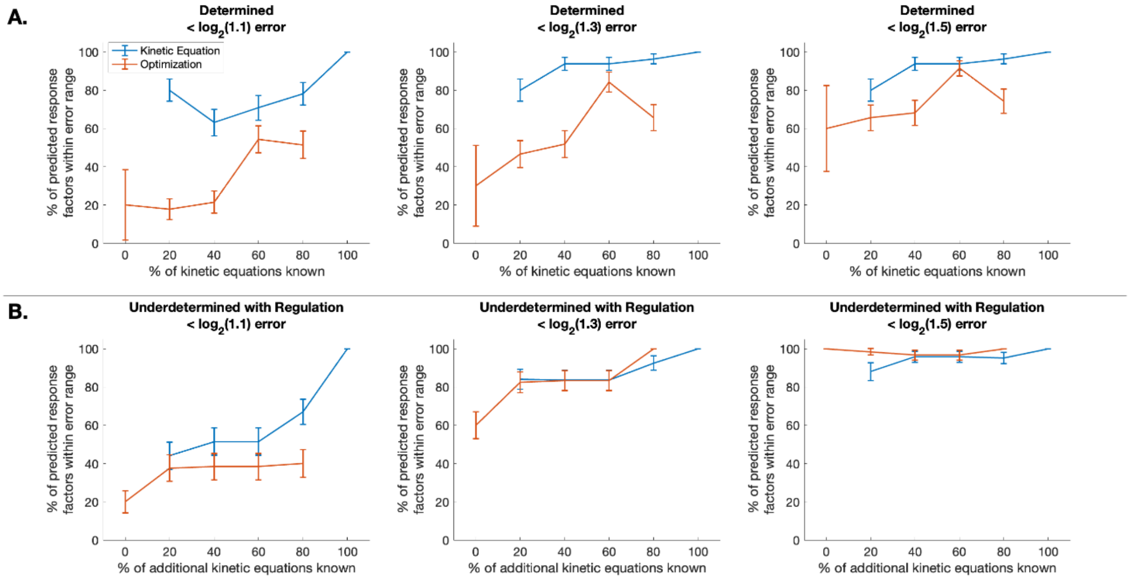

Figure S7. Percent of response factors predicted by the kinetic equation and optimization steps of MetaboPAC within each $\log_2$ error range for the synthetic models when using noiseless data. The kinetic equation approach often outperforms the optimization approach in predicting response factors within log2(1.1) and log2(1.3) error range. The performance of the optimization and kinetic equation approaches generally increases with increasing percentage of known kinetic equations. Error bars represent the standard error of the mean (n = 50 for different sets of true response factors and different sets of known kinetic equations for 20-80% known kinetics, n=5 for different sets of true response factors for 0% and 100%. Number of response factors predicted by kinetic equation or optimization approach varies based on the percentage of kinetic equations known (Figure S8))**.**

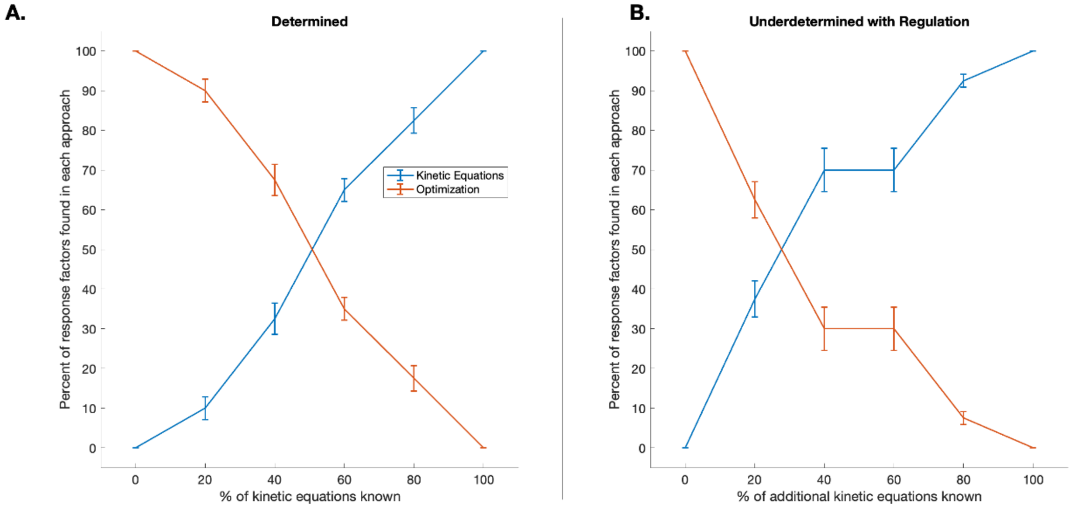

Figure S8. Percentage of response factors predicted overall by the kinetic equation and optimization steps of MetaboPAC in the synthetic models. As the percentage of known kinetic equations increases, it is more likely for response factors to be solvable using the kinetic equation approach. Error bars represent the standard error of the mean (n = 50 for different sets of true response factors and different sets of known kinetic equations for 20-80% known kinetics, n=5 for different sets of true response factors for 0% and 100%).

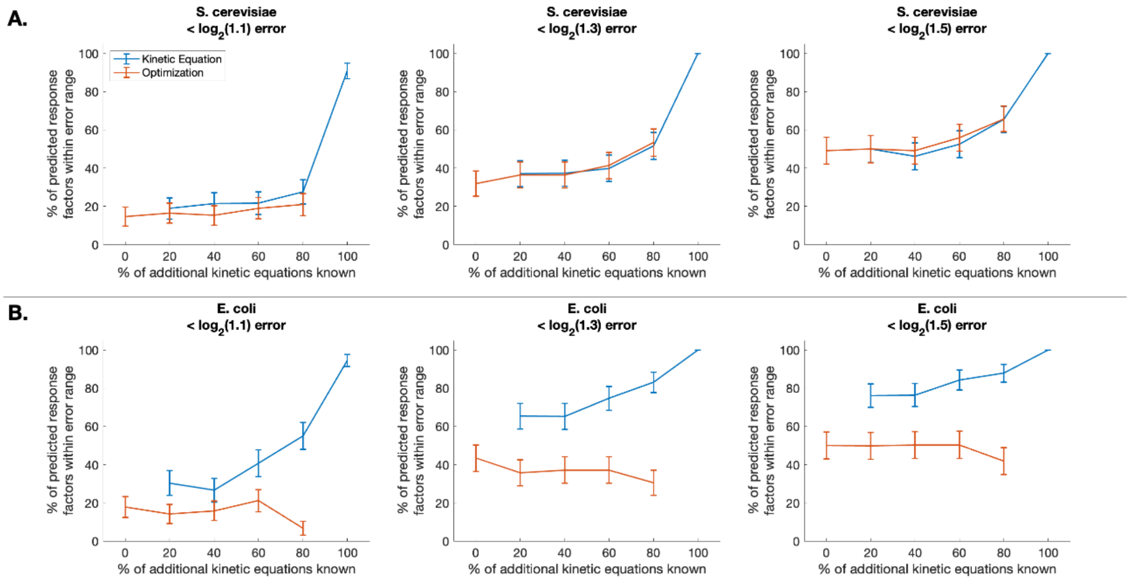

Figure S9. Percent of response factors predicted by the kinetic equation and optimization steps of MetaboPAC within each log$_2$ error range for the *S. cerevisiae* and *E. coli* systems when using noiseless data. There is a substantial improvement in performance for the kinetic equation approach at a higher percentage of known kinetic equations, but not for the optimization approach as more kinetic equations are known. Error bars represent the standard error of the mean (n = 50 for different sets of true response factors and different sets of known kinetic equations for 20-80% known kinetics, n=5 for different sets of true response factors for 0% and 100%. Number of response factors predicted by kinetic equation or optimization approach varies based on the percentage of kinetic equations known (Figure S10)**).**

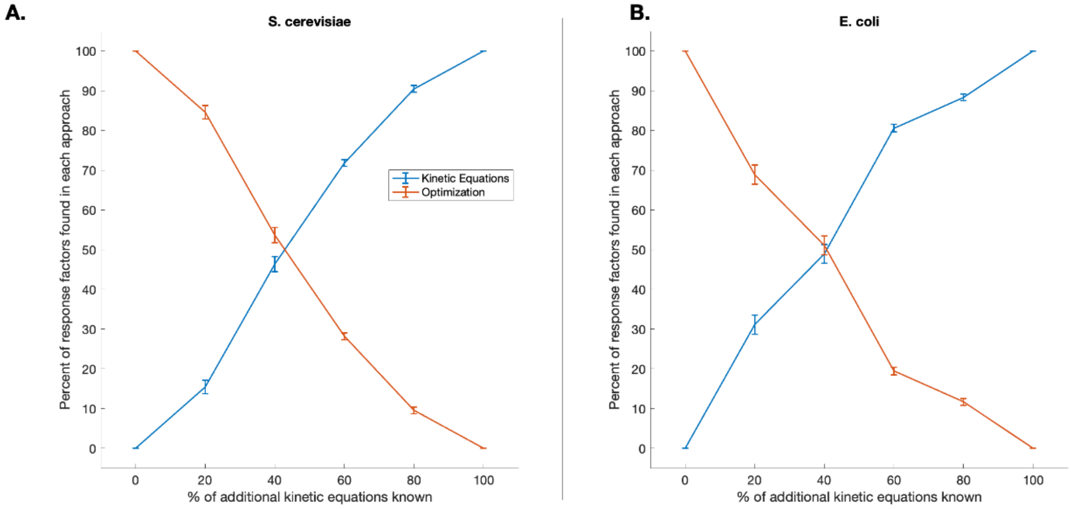

Figure S10. Percentage of response factors predicted overall by the kinetic equation and optimization steps of MetaboPAC in the *S. cerevisiae* and *E. coli* systems. As the percentage of known kinetic equations increases, it is more likely for response factors to be solved using the kinetic equation approach. Error bars represent the standard error of the mean (n = 50 for

different sets of true response factors and different sets of known kinetic equations for 20-80% known kinetics, n=5 for different sets of true response factors for 0% and 100%).

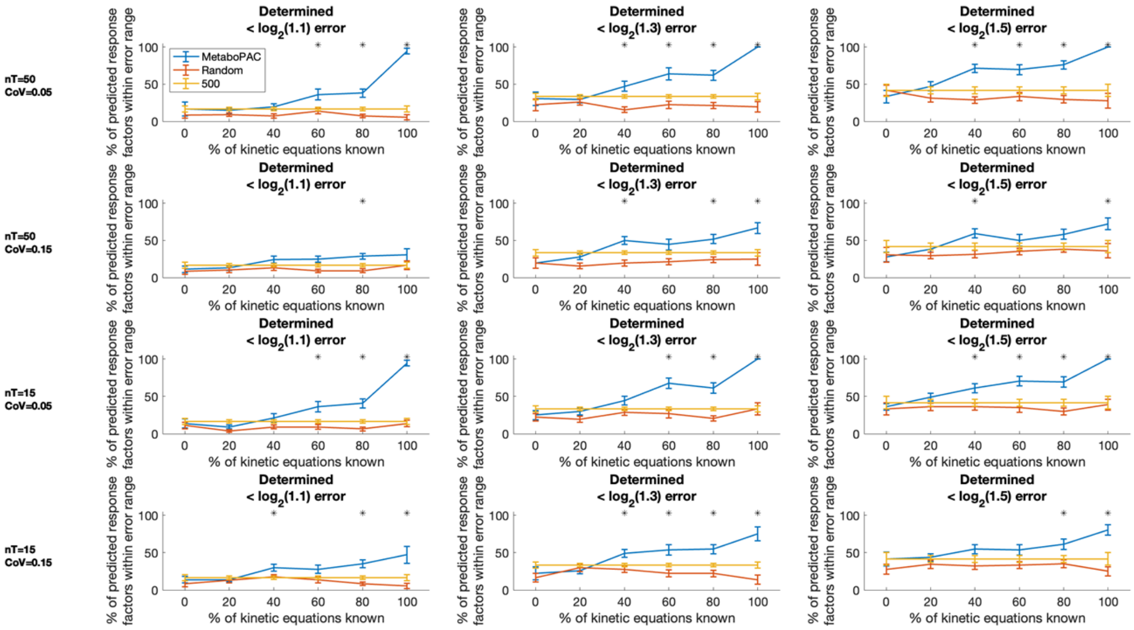

Figure S11. MetaboPAC performance on various conditions of noisy data for the Determined model. MetaboPAC is compared to random response factors and response factors of 500 for the determined model using error ranges of log2(1.1), log2(1.3), log2(1.5) on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n=9 for 3 different sets of true response factors and 3 replicates of noisy data for 0% and 100% known kinetic equations, n = 27 for 3 different sets of true response factors, 3 different subsets of known kinetic equations, and 3 replicates of noisy data for the rest). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with α = 0.05).

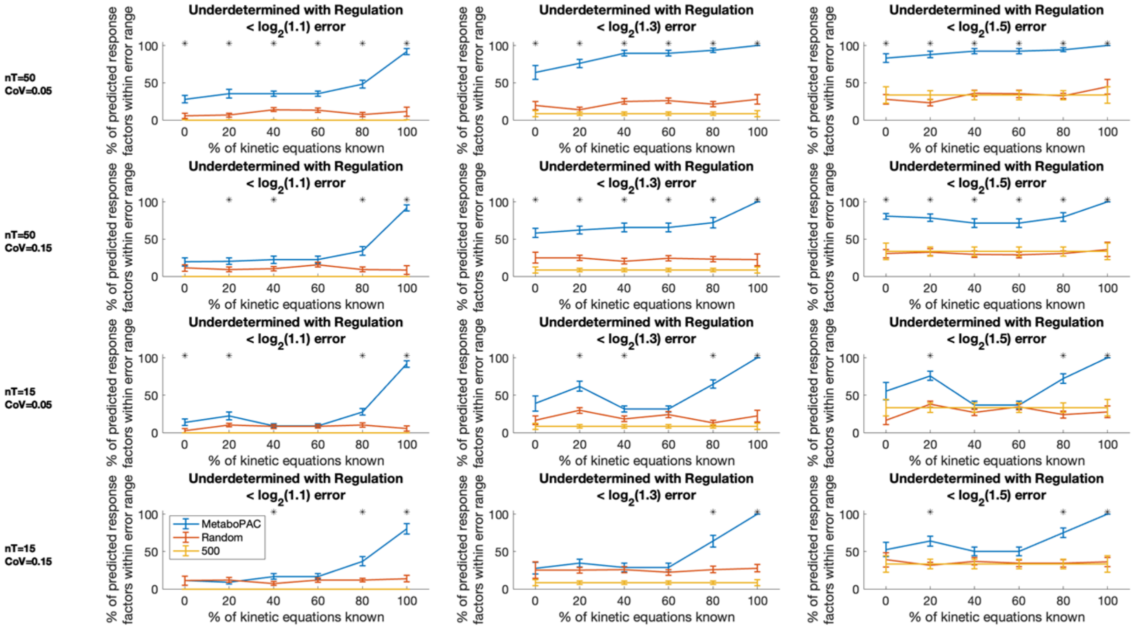

Figure S12. MetaboPAC performance on various conditions of noisy data for the Underdetermined with Regulation model. MetaboPAC is compared to random response factors and response factors of 500 for the underdetermined with regulation model using error ranges of log2(1.1), log2(1.3), log2(1.5) on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n=9 for 3 different sets of true response factors and 3 replicates of noisy data for 0% and 100% known kinetic equations, n = 27 for 3 different sets of true response factors, 3 different subsets of known kinetic equations, and 3 replicates of noisy data for the rest). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with α = 0.05).

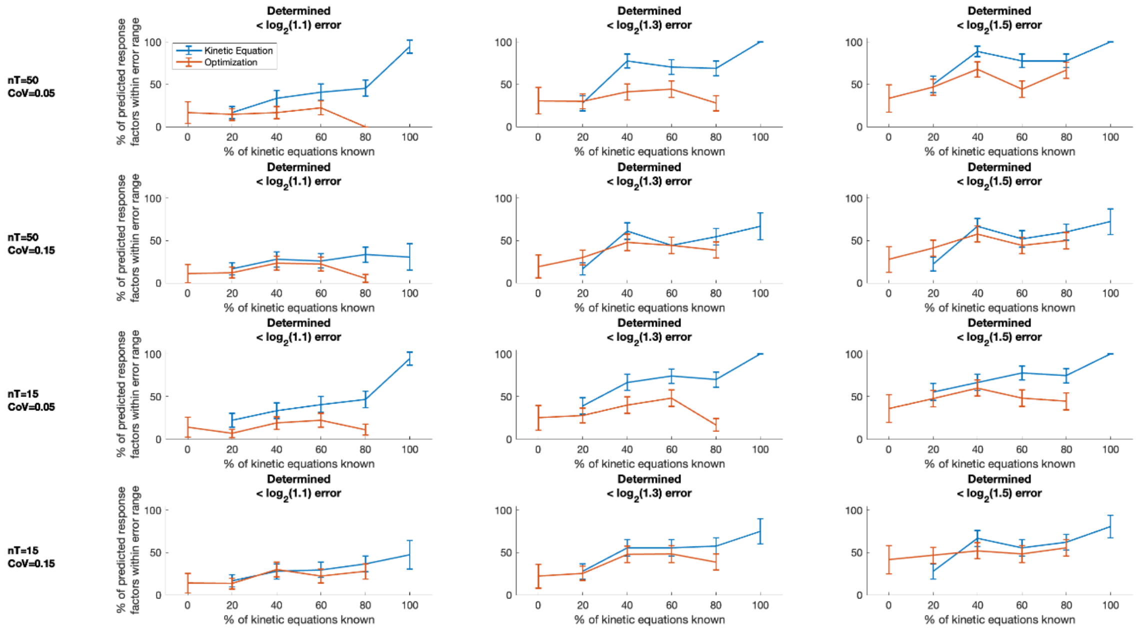

Figure S13. Percent of response factors predicted by the kinetic equation and optimization steps of MetaboPAC within each $\log_2$ error range for the determined system when using noisy data. The kinetic equation approach generally predicted more accurate response factors than the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure S8))**.**

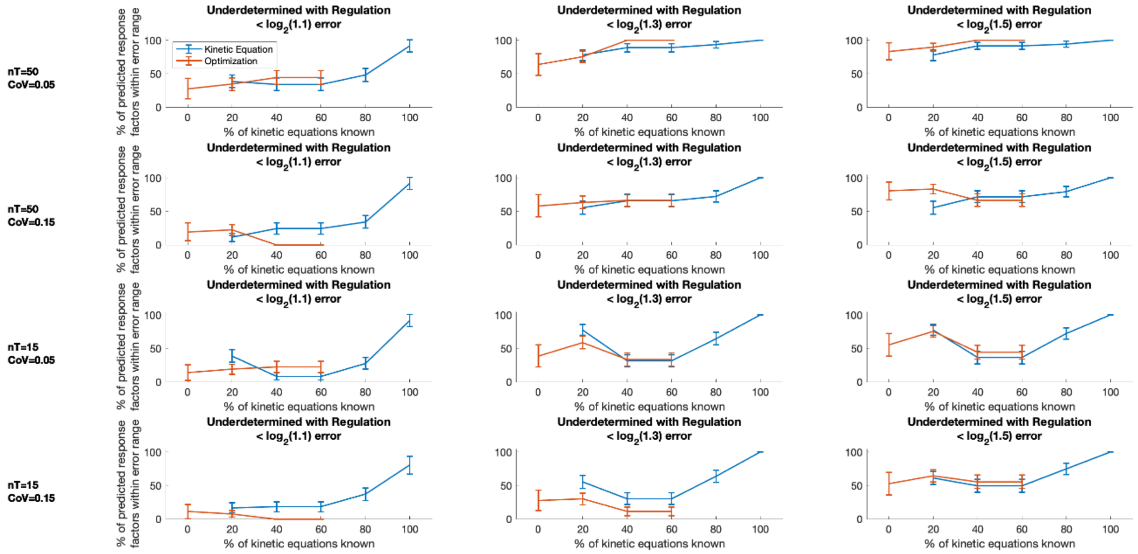

Figure S14. Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the underdetermined system with regulation

when using noisy data. At a lower sampling frequency, there is a decrease in performance of the kinetic equation approach, which is coupled to compromised performance of the optimization approach, causing an overall decrease of prediction accuracy for response factors at the middle known kinetics. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure S8))**.**

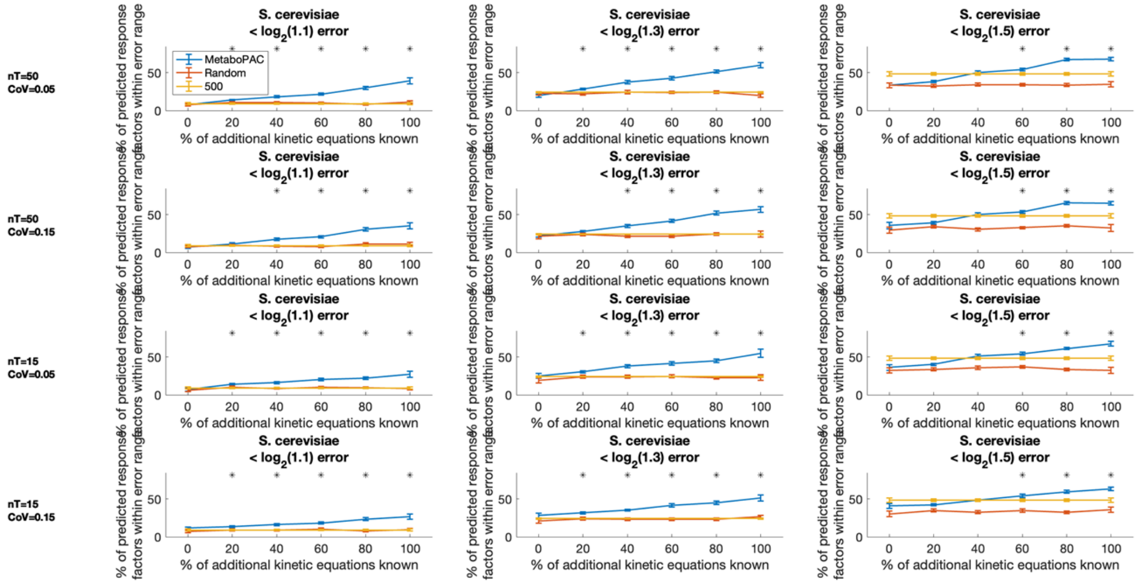

Figure S15. MetaboPAC performance on various conditions of noisy data for the *S. cerevisiae* model. MetaboPAC is compared to random response factors and response factors of 500 for the *S. cerevisiae* model using error ranges of log2(1.1), log2(1.3), log2(1.5) on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n=9 for 3 different sets of true response factors and 3 replicates of noisy data for 0% and 100% known kinetic equations, n = 27 for 3 different sets of true response factors, 3 different subsets of known kinetic equations, and 3 replicates of noisy data for the rest). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with α = 0.05).

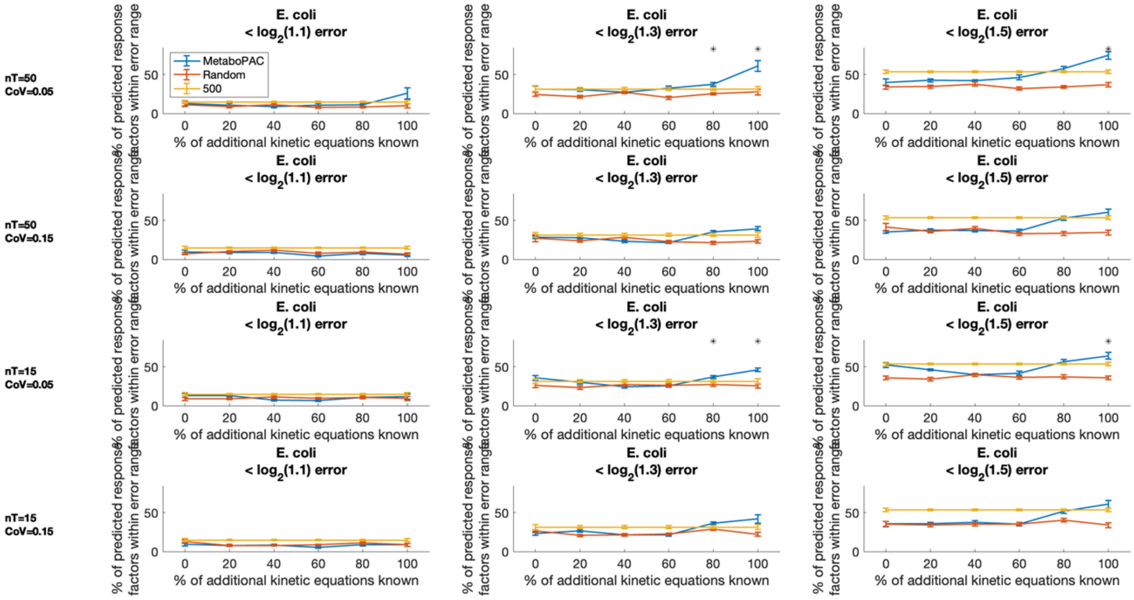

Figure S16. MetaboPAC performance on various conditions of noisy data for the *E. coli* model. MetaboPAC is compared to random response factors and response factors of 500 for the *E. coli* model using error ranges of log2(1.1), log2(1.3), log2(1.5) on data with different sampling frequencies (nT = 50 or 15) and noise added (CoV = 0.05 or 0.15). Lines represent the mean percent of predicted response factors within the error ranges for each method. Error bars represent the standard error of the mean (n=9 for 3 different sets of true response factors and 3 replicates of noisy data for 0% and 100% known kinetic equations, n = 27 for 3 different sets of true response factors, 3 different subsets of known kinetic equations, and 3 replicates of noisy data for the rest). Asterisks denote when MetaboPAC performed significantly better at predicting response factors than both of the other two methods (two-sample t-test with α = 0.05).

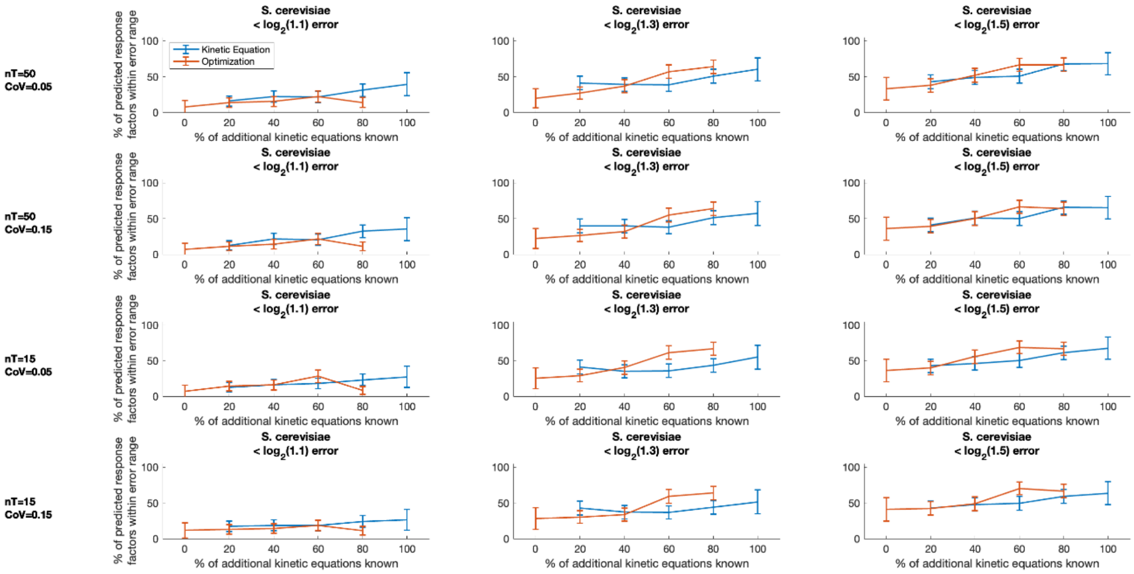

Figure S17. Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the *S. cerevisiae* system when using noisy data. As more kinetic equations are known, the optimization approach predicts more response factors within the $\log_2(1.3)$ error range and the $\log_2(1.5)$ error range, but there is no significant improvement in the performance of the kinetic equation approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure S10)).

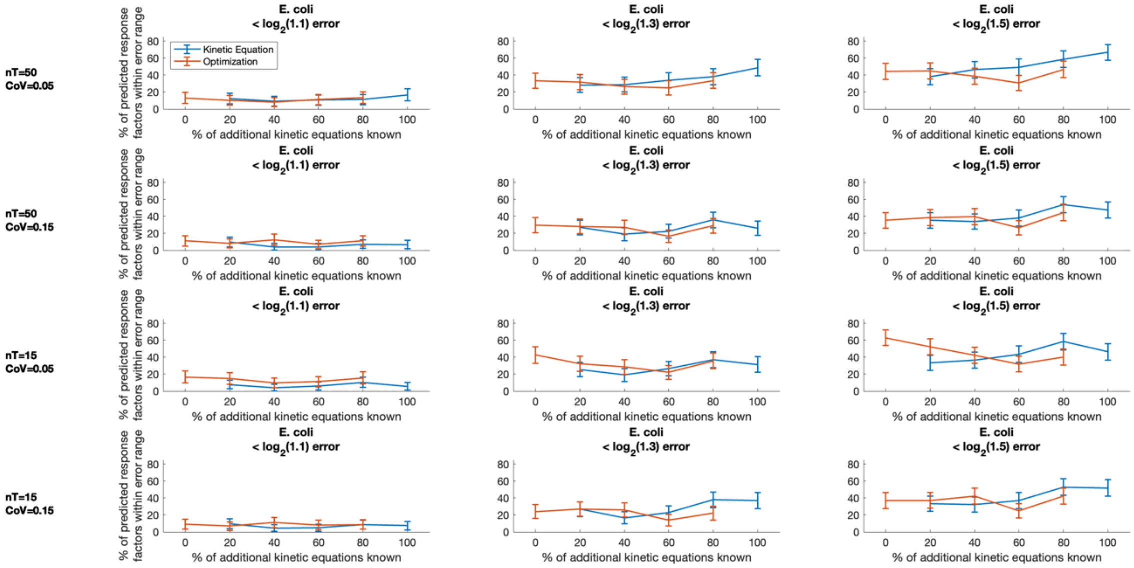

Figure S18. Percent of response factors predicted by the kinetic equation and optimization approaches within each $\log_2$ error range for the *E. coli* system when using noisy data. The two approaches have similar performance. There is a slight improvement in the performance of the kinetic equation approach as more kinetic equations are known, but no significant improvement in the performance of the optimization approach. Error bars represent the standard error of the mean (number of samples varies based on the percentage of kinetic equations known (Figure S10)).

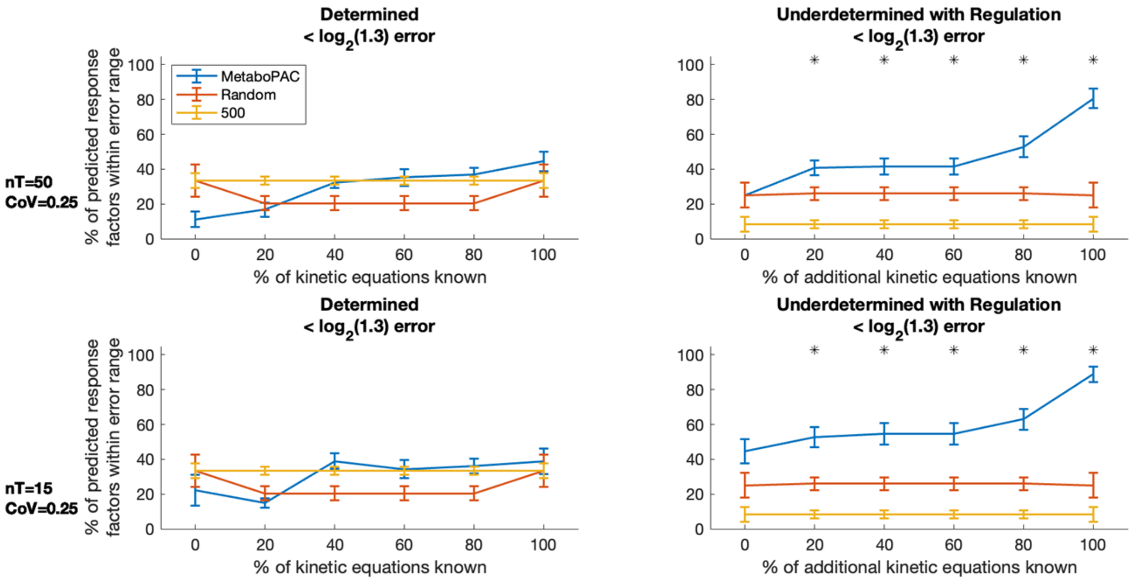

Figure S19. MetaboPAC performance on the synthetic models with data with more noise. In addition to CoV=0.05 and CoV=0.15 as in the main text, CoV=0.25 is used to assess the robustness of MetaboPAC to higher experimental noise. As discussed in the main text, the performance of MetaboPAC on the determined pathway is sensitive to the addition of noise while that on the underdetermined with regulation pathway seems to be robust to noise.

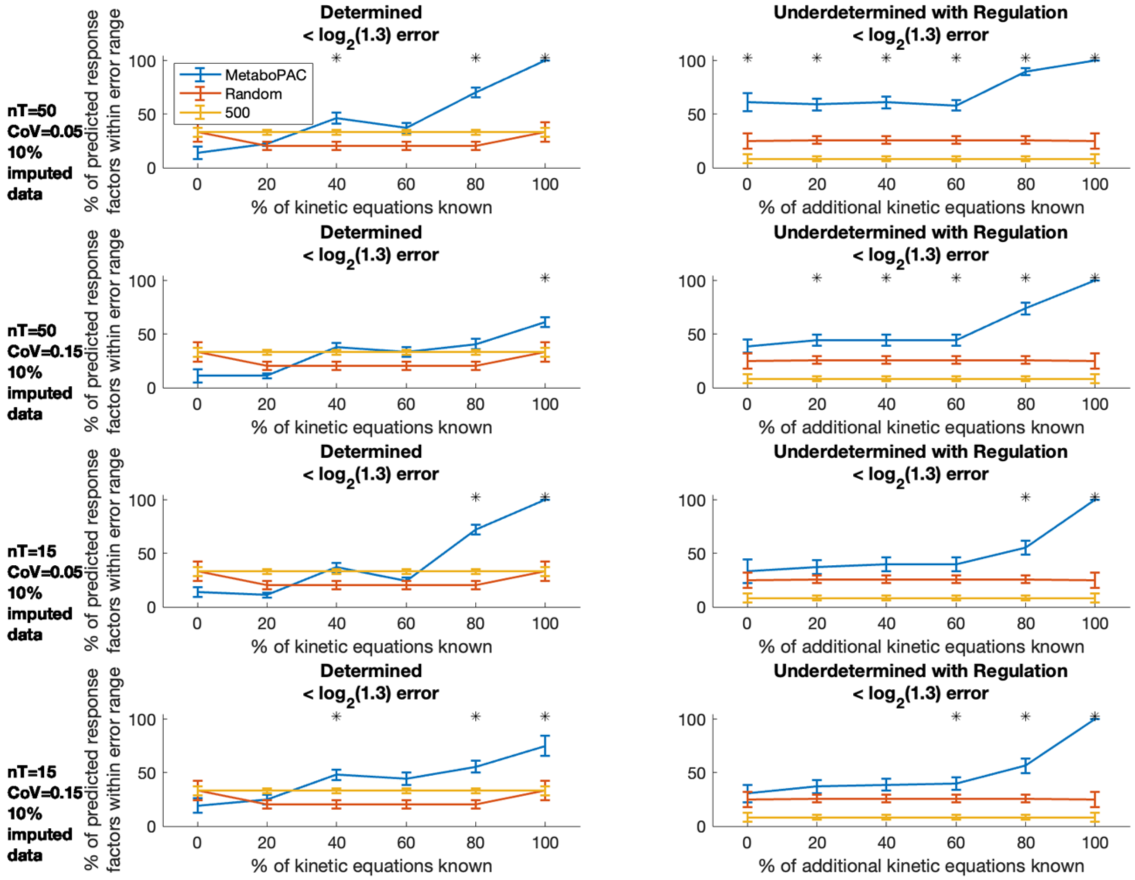

Figure S20. MetaboPAC performance on various conditions of noisy data with 10% missing values for the synthetic models. To assess the robustness of our approach to missing values in data, we removed 10% of the data using a realistic representation of missing value distribution in metabolomics datasets, where the level of missingness is assumed to be different for different metabolites, and that metabolites with lower abundance are more prone to have missing values. Missing data were removed with parameter values of I = 70%, II = 70%, and III = 15%; details about these parameters of missingness are described in [4]. The missing values were then imputed using the modified k-nearest neighbors approach in [4] and MetaboPAC was performed on the resulting relative abundance metabolomics data. Compared to Figure 5, there is a slight decrease in performance across all noise conditions, but MetaboPAC was still significantly better than both the random and 500 response factors approaches when 100% kinetic equation terms are known.

Supplementary Methods

Accurate flux estimation assumption for underdetermined system

To obtain a unique and accurate flux estimation from time-series metabolomics data, the metabolic pathway needs to be determined (having the same number of metabolites and reactions). In the case of underdetermined systems, some kinetic equations are assumed to be known such that the systems will be determined. The known kinetic equations for reactions are selected such that the number of remaining reactions with unknown kinetic equations is equal to the number of metabolites and that the stoichiometric equations are not linearly dependent. The percentage of additional known kinetic equations in figure legends then refers to any known kinetic equations in addition to those assumed to be known to ensure the systems are determined.

Penalty weight optimization

The penalty terms in Table S1 are incorporated into the optimization approach as soft constraints such that the sum of these terms is minimized. Since each penalty term may be of a different order of magnitude, different penalties may have varying impacts on the optimization problem in each system if used in an unweighted fashion. To account for this, we added a penalty weight vector and multiplied that by the penalty terms such that all penalties except the mass balance penalty are brought to the same order of magnitude. To determine the weight vector for each system, 10,000 random sets of response factors were sampled, and the minimum value of each penalty was found and used as the benchmark for the weight vector. Since the mass balance penalty is considered to be a centrally important constraint, it is given a larger penalty weight (100x) than the others.

References

1. Voit, E.O., *Biochemical Systems Theory: A Review.* ISRN Biomathematics, 2013. **2013**: p. 1-53.
2. Chassagnole, C., et al., *Dynamic modeling of the central carbon metabolism of Escherichia coli.* Biotechnol Bioeng, 2002. **79**(1): p. 53-73.
3. Hynne, F., S. Danø, and P.G. Sørensen, *Full-scale model of glycolysis in Saccharomyces cerevisiae.* Biophysical Chemistry, 2001. **94**(1-2): p. 121-163.
4. Lee, J.Y. and M.P. Styczynski, *NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data.* Metabolomics, 2018. **14**(12): p. 153.