Bayesian inference of metabolic kinetics from genome-scale multiomics data

Peter C. St. John[1], Jonathan Strutz[2], Linda J. Broadbelt[2], Keith E.J. Tyo[2], Yannick J. Bomble[1,*]

[1]Biosciences Center, National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden CO 80401, USA
[2]Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Evanston IL 60208, USA
[*]Email: Yannick.Bomble@nrel.gov

Summary

Modern biological tools generate a wealth of data on metabolite and protein concentrations that can be used to help inform new strain designs. However, integrating these data sources to generate predictions of steady-state metabolism typically requires a kinetic description of the enzymatic reactions that occur within a cell. Parameterizing these kinetic models from biological data can be computationally difficult, especially as the amount of data increases. Robust methods must also be able to quantify the uncertainty in model parameters as a function of the available data, which can be particularly computationally intensive. The field of Bayesian inference offers a wide range of methods for estimating distributions in parameter uncertainty. However, these techniques are poorly suited to kinetic metabolic modeling due to the complex kinetic rate laws typically employed and the resulting dynamic system that must be solved. In this paper, we employ linear-logarithmic kinetics to simplify the calculation of steady-state flux distributions and enable efficient sampling and variational inference methods. We demonstrate that detailed information on the posterior distribution of kinetic model parameters can be obtained efficiently at a variety of different problem scales, including large-scale kinetic models trained on multiomics datasets. These results allow modern Bayesian machine learning tools to be leveraged in understanding biological data and developing new, efficient strain designs.

Introduction

Optimizing the metabolism of microorganisms for maximum yields and titers is a critical step in improving bioprocess economics (Davis et al., 2013). Towards this goal, characterization of engineered strains has become increasingly detailed with the growing availability of transcriptomic, proteomic, and metabolomic analysis techniques (Zhang et al., 2009). These methods, collectively termed multiomics, measure relative changes in gene, protein, or metabolite concentrations across different strains or growth conditions (Hackett et al., 2016). However, utilizing multiomics data to make informed decisions about future strain improvements remains a major challenge in modern bioengineering (Marcellin and Nielsen, 2018). Cellular metabolism is controlled by a vast network of enzymes with complex and nonlinear kinetics, while regulatory effects (both allosteric and transcriptional) add additional layers of complexity to these metabolic systems. While the field of systems biology has developed modeling frameworks that can describe these types of interactions in great detail, parameterizing these models from indirect, *in vivo* data

is challenging and typically infeasible at the genome-scale (Saa and Nielsen, 2017). There is therefore a need for mechanistic modeling frameworks that can handle the large amounts of data generated through multiomics experiments to yield actionable insight for strain engineers.

Traditionally, multiomics data have been understood through statistical approaches by searching for genes, proteins, and metabolites whose activity levels are correlated with improved product production (Yoshikawa et al., 2012). While useful in identifying potential targets, statistical methods that examine multiomics data without consideration for their interconnected nature may miss important trends. Conversely, metabolic modeling frameworks that readily reconcile connections between metabolites, fluxes, and proteins can have difficulty in using multiomics data to improve their predictions. Stoichiometric models of metabolic networks have proved among the most successful techniques for incorporating existing knowledge of genomic and biochemical networks into strain designs (Orth et al., 2010). Instead of attempting to estimate the parameters of detailed rate rules for each enzymatic reaction in the cell, constraint-based models assume that metabolic reactions reach a pseudo-steady state with respect to the longer time scales of cell growth and substrate depletion. These methods there investigate feasible steady-state phenomena in a parameter-free approach by placing constraints on reaction fluxes in accordance with stoichiometric (Orth et al., 2010), thermodynamic (Henry et al., 2007), enzymatic (Sánchez et al., 2017), and regulatory (reviewed in (Saha et al., 2014)) rules. The resulting models are invaluable for predicting ways to restrict cell physiology to specific regions (*e.g.* forcing growth-coupled product production through gene knockouts (Burgard et al., 2003)). However, their parameter-free construction renders the direct inclusion of measured metabolite and enzyme concentrations difficult, and their lack of kinetic information makes them poorly suited to recommending enzyme expression changes to optimize pathway flux. Additionally, constraint-based techniques typically assume growth as a cellular objective, making them poorly suited to *in vitro* or other non-growth associated bioprocesses.

While some studies have used constraint-based techniques to interpret multiomics data (Brunk et al., 2016; Cotten and Reed, 2013; O'Brien et al., 2014; Sánchez et al., 2017; Yizhak et al., 2010), building and parameterizing kinetic models will likely be essential in utilizing these types of data to predict metabolic interventions that will improve flux through a given pathway. Kinetic models of metabolism typically describe the interior cellular environment through systems of coupled, nonlinear ordinary differential equations with metabolite concentrations as the state variables. Metabolite concentrations change with time according to the kinetics of enzyme-mediated reactions, reaching a stable steady-state for constant concentrations of extracellular substrates. Understanding systems-level effects is the main motivation of metabolic control analysis (MCA), which links effects of local perturbations (*i.e.*, changes to enzyme expression) to changes in the resulting steady-state concentrations and fluxes (Ehlde and Zacchi, 1997; Visser and Heijnen, 2002). MCA defines local coefficients, or *elasticities*, which are local approximations to the reaction rate rule and relate reaction flux to substrate concentration. Through linearization, MCA also solves for global *control coefficients*, which relate steady-state fluxes and concentrations to enzyme levels. Parameters estimated via MCA are directly relevant to strain engineering goals, as they allow the prediction of which enzymes to over- or underexpress to achieve a desired change in pathway flux. As a result, a number of computational frameworks have been developed to perform MCA with incomplete data (Chakrabarti et al., 2013; Delgado and Liao, 1991; Wang et al., 2004; Wu et al., 2004). Most MCA approaches require an accurate dynamic model of metabolism that must be informed from experimental measurements. However, direct measurements of enzyme kinetics *in vivo* are difficult, and measurements *in vitro* do not always accurately reflect *in vivo* dynamic behavior (Teusink et al., 2000; Zotter et al., 2017). Experimental characterization of genome-scale enzyme kinetics is particularly challenging (Nilsson et al., 2017). As a result, the most readily available data for parameterizing kinetic metabolic models is obtained by repeatedly perturbing enzyme expression or external metabolite concentration and characterizing the resulting strain's pseudo-steady-state behavior through multiomics experiments.

The process for constructing a kinetic metabolic model therefore consists of choosing an appropriate functional form for the rate rule of each reaction and estimating parameter values from experimental data. A number of different frameworks for describing the kinetics of enzyme-catalyzed reactions have been proposed, covering a spectrum of computational efficiency and mechanistic fidelity (reviewed in (Heijnen, 2005; Saa and Nielsen, 2017)). However, regardless of the framework chosen, estimating uncertainty in fitted

parameter values is a challenging process due to the dimensionality of the resulting parameter space. One approach is to find an ensemble of possible parameter values that when passed through the kinetic model closely reproduce the observed experimental data (Tran et al., 2008). These distributions in parameter values can then be updated as more data is collected and used to predict enzyme targets that give the most likely chance of improving performance (Contador et al., 2009). This technique has more recently been formalized as Bayesian inference (Saa and Nielsen, 2016), where parameter values are modeled as probability distributions. In Bayesian inference, a *likelihood* model, $p(y|\theta)$, is constructed for the probability of observing the measured data, y, given particular values for each parameter, θ. Combined with a *prior* distribution, $p(\theta)$, for each parameter that represents generally feasible values, numerical approaches use Bayes theorem,

$$p(\theta|y) \propto p(y|\theta)\ p(\theta)$$

to estimate the *posterior* parameter distribution $p(\theta|y)$: the probability a parameter takes the given value after accounting for the observed data.

A major limitation of ensemble-based modeling has been its ability to scale both to larger datasets as well as larger kinetic models. Computation times for metabolic ensemble modeling (MEM) on the order of hours per parameter sample have been encountered even for relatively small models and datasets in previous studies (Saa and Nielsen, 2016; Tran et al., 2008). Despite improvements to the computational efficiency of MEM (Greene et al., 2017; Zomorrodi et al., 2013), applications of ensemble modeling have been restricted to small datasets and/or kinetic models, without the ability to scale to genome-scale kinetic representations fit with data from multiomics workflows. This limitation stems from the need to perform computationally intensive integration of underlying ODEs to find steady-state concentrations and fluxes, combined with a relatively inefficient rejection sampler approach used to estimate the posterior distribution. Modern and efficient inference algorithms (Hoffman and Gelman, 2014; Kucukelbir et al., 2017) require information on the gradient of the likelihood function with respect to the kinetic parameters, which can only be obtained for ODE models through numerically intensive ODE sensitivity analysis (Li and Petzold, 2000).

In this study, we present a scalable method for inferring posterior distributions in kinetic parameters of large metabolic models with multiomics datasets. We sidestep many of the previously discussed computational bottlenecks through the use of linear logarithmic (linlog) kinetics as an approximate reaction rate rule (Visser and Heijnen, 2003; Visser et al., 2004). Linlog kinetics is derived using the thermodynamic concept that reaction rate is proportional to reaction affinity near equilibrium (Onsager, 1931). While many biochemical reactions are far from equilibrium, this relationship remains linear over a wide range of reaction affinities for enzymatic reactions (Meer et al., 1980; Rottenberg, 1973), As an approximation, linlog kinetics does not describe enzyme-mediated kinetics as faithfully as more mechanistic frameworks (Saa and Nielsen, 2017). However, linlog kinetics has been shown to be accurate up to 20-fold changes in metabolite concentrations (Visser and Heijnen, 2003), and for 4 to 6-fold changes in enzyme concentration relative to a reference state (Visser et al., 2004). As a result, linlog kinetics has been used as a framework for estimating flux control coefficients from a range of data sources (Heijnen et al., 2004; Kresnowati et al., 2005; Nikerel et al., 2006, 2009). Most importantly, this kinetic formalism allows steady-state fluxes and metabolite concentrations as a function of enzyme expression to be determined directly via linear algebra, without the need to explicitly integrate the dynamic system until a steady-state is reached (Visser et al., 2004). We are therefore able to leverage modern Bayesian inference and machine learning algorithms, including Hamiltonian Monte Carlo (HMC) (Neal, 2010) and variational inference (Blei et al., 2017) to fully characterize the posterior space. Additionally, this framework naturally lends itself to directly incorporating relative changes in metabolite and protein concentrations between experimental conditions, without requiring absolute quantification.

We show that this method is capable of providing systems-level insight into metabolic kinetics through estimated distributions in control coefficients for a wide range of kinetic model and dataset sizes. First, we demonstrate the method on a simple *in vitro* example, showing that the method is flexible enough to capture allosteric interactions between metabolites and enzymes. We next show that the method appropriately captures uncertainy in estimated parameters, revealing significant flux control coefficients for only the most likely enzyme perturbations in the case of limited biological data. Finally, we employ the method to integrate thousands of individual metabolomic, proteomic, and fluxomic data-points with a large-scale model of yeast metabolism. We therefore show that the field of metabolic modeling can take full advantage

of recent advances in the fields of probabilistic programming, machine learning, and computational statistics, and that ensemble-based approximate kinetic modeling approaches are capable of scaling to genome-sized models and datasets to provide interpretable and actionable insight for strain engineers.

Results and discussion

Enabling efficient Bayesian inference through linlog kinetics

We begin with a review of the relevant equations from dynamic flux balance analysis and the linear-logarithmic kinetic framework, which together form the theoretical basis for the methodology discussed in the remainder of the study. In flux balance analysis, we assume that metabolite concentrations, x, quickly reach a pseudo-steady state by balancing fluxes v through each reaction.

$$\frac{dx}{dt} = \underset{m\times n}{N}\, v(x) = 0, \tag{1}$$

for n reactions and m metabolites, where N_{ij} indicates the stoichiometry of metabolite i in reaction j. Linlog kinetics approximates a reaction rate $v(x)$ as a sum of logarithms (Visser and Heijnen, 2003). For the reaction $A \rightarrow B + C$, the reaction rate is modeled as

$$v = e(k + a\log[A] + b\log[B] + c\log[C]),$$

for which the coefficients $a > 0$; $b, c < 0$ allow an approximation of Michaelis-Menten-type kinetics (Figure 1). This approximation is most accurate in the vicinity of an introduced reference state, e^*, v^*, x^* (Visser and Heijnen, 2003). As the goal of the proposed method is to tailor enzyme expression to maximize desired fluxes, the reference state is best chosen as the current optimal performing strain. Deviations from this state can be described by the flux expression

$$v(x,y) = \text{diag}\left(\frac{v^*e}{e^*}\right)\left(1_n + \underset{n\times m}{\epsilon_x^*}\log\frac{x}{x^*} + \underset{n\times p}{\epsilon_y^*}\log\frac{y}{y^*}\right) \tag{2}$$

where y is the concentration of p external (independently controllable) metabolite species, and ϵ_x^* and ϵ_y^* are sparse matrices of kinetic parameters describing the effects of changes to metabolite concentrations on reaction rates. Elasticities parameterize the slope of the reaction rate rule near the reference state. Linear-logarithmic kinetics therefore offer a close approximation to standard Michaelis-Menten kinetics in the vicinity of the reference state concentration (x^*). A benefit of the linlog approximation is that enzyme elasticities are direct kinetic parameters. Since these slopes tend to be positive for reactants, negative for products, and not be much larger than 1, reasonable starting guesses and bounds can be generated for all kinetic parameters in the model in a much easier fashion and for rate rules parameterized through traditional enzymatic expressions. Elasticities for linear logarithmic kinetics have typically been estimated in the literature using multiple linear regression (Chen et al., 2017; Wu et al., 2004), where estimated fluxes for each reaction are fitted as a function of their measured metabolite concentrations. However, this approach does not enforce the $Nv = 0$ constraint, nor does it allow for missing data in concentration or flux measurements. We demonstrate that incorporating steady-state constraints is computationally feasible, and that a full characterization of the posterior space can be accomplished using Hamiltonian Monte Carlo.

While linlog kinetics is a close approximation of more mechanistic rate rules, it suffers from a number of notable inconsistencies. One consequence is that fluxes can approach negative infinity as metabolite concentrations approach zero, making the framework unsuitable for describing complete pathway knockouts. However, in practice metabolite concentrations are typically expressed as log-transformed variables which also cannot fall to zero. Other methodological strategies discussed later also prevent fluxes from taking

unrealistic values, including using a least-norm linear solve for steady-state concentrations and clipping data to a finite range. Additionally, as a local approximation, the method will poorly reproduce alternative rate rules at large deviations from the reference state. However, because cellular systems are constrained by homeostasis, metabolite concentrations generally do not change drastically enough to invalidate rate estimates (Ishii et al., 2007).

A key step in dynamic modeling of metabolic networks is solving for the steady-state concentrations and fluxes that arise from a given parameterization. Simulating this perturbation efficiently with the mathematical model is therefore a key step in estimating parameter values for the ϵ^* matrices. In doing so, it is useful to define transformed variables in order to rewrite Equation 2 in a linear form (as demonstrated by Smallbone et al., 2007):

$$\begin{gathered}\chi = \log\frac{x}{x^*}; \quad \gamma = \log\frac{y}{y^*}; \quad \hat{v} = \frac{v}{v^*}; \quad \hat{e} = \frac{e}{e^*} \\ v = \operatorname{diag}(v^*\hat{e})(1_n + \epsilon_x^*\chi + \epsilon_y^*\gamma)\end{gathered} \tag{3}$$

Since log-transformed metabolite concentrations are linearly related to the reaction fluxes, concentrations which yield steady-state behavior can therefore be determined via a linear solve (Visser and Heijnen, 2003) after combining Equation 3 with Equation 1:

$$\begin{gathered}N\operatorname{diag}(v^*\hat{e})(1_n + \epsilon_x^*\chi + \epsilon_y^*\gamma) = 0 \\ \underbrace{N\operatorname{diag}(v^*\hat{e})\epsilon_x^*}_{\mathbf{A}}\chi = \underbrace{-N\operatorname{diag}(v^*\hat{e})(1_n + \epsilon_y^*\gamma)}_{\mathbf{b}} \\ \chi = \mathbf{A}^{-1}\mathbf{b}\end{gathered} \tag{4}$$

This significant result is the key advantage of linlog kinetics over alternative nonlinear rate laws. While determination of steady-state concentrations would typically require a computationally intensive ODE integration, in this approximation they can instead be calculated using a single linear solve. Additionally, it is relatively straightforward to obtain forward and reverse-mode gradients for this operation (changes in steady-state with respect to changes in kinetic parameters), a much more difficult task for ODE integration (Petersen and Pedersen, 2012).

However, in general a metabolic system will contain *conserved moieties*, or metabolite quantities which can be expressed as linear combinations of other metabolites (*e.g.* ATP + ADP = constant). The stoichiometric matrix N, and as a result the $\mathbf{A}$ matrix defined above, will therefore not be full row rank. In effect, this means that Equation 4 has multiple solutions, each of which corresponds to a different total cofactor pool. In metabolic control theory, this problem has traditionally been solved through the introduction of a link matrix, L, and a reduced set of metabolites with conserved moieties removed (Smallbone et al., 2007; Visser and Heijnen, 2002). Through the link matrix, the matrix $\mathbf{A}$ can be transformed to a full-rank, square matrix and a unique steady-state can be determined that corresponds to the dynamic system's true steady-state. However, in most biological experiments, changes to steady-state enzyme expression correspond with separately cultured cell lines for which the assumption that total cofactor pools would remain constant is not necessarily valid. Instead, we propose that a more biologically relevant solution to Equation 4 is one that minimizes $\|\chi\|_2$: *i.e.*, the solution that results in the smallest deviation of metabolite concentrations from the reference state. This assumption has experimental support in that intracellular metabolite concentrations tend to be buffered from drastic changes through feedback circuits at the genetic and enzyme level (Ishii et al., 2007). We therefore calculate steady-state metabolite concentrations through a pseudoinverse,

$$\chi_{ss} = \mathbf{A}^{\dagger}\mathbf{b} \tag{5}$$

A derivation of the forward and reverse-mode gradients for the regularized linear solve operation is included in the supplemental text. We note that in practice, numerical stability is improved if $\mathbf{A}$ can be

made full row-rank prior to the least-norm linear solve. We can therefore replace N with $\tilde{N}$ (by removing rows corresponding to redundant conservation relations) in order to form a wide $\mathbf{A}$ matrix (with more columns than rows) prior to performing the least-norm linear solve in Equation 5. Since a flux vector that satisfies $\tilde{N}\, v = 0$ will also satisfy $N\, v = 0$, this change can be made without affecting the final solution.

Due to the changes to traditional MCA theory introduced by the altered steady-state calculation defined above, we also slightly modify the traditional calculations of metabolite and flux control coefficients (FCCs).

$$C^*_{x,kj} = \frac{e^*_j}{x^*_k}\frac{dx_k}{de_j} = -\left(\tilde{N}\,\mathrm{diag}(v^*)\epsilon^*_x\right)^{\dagger}\tilde{N}\,\mathrm{diag}\,v^*$$

$$C^*_{v,ij} = \frac{e^*_j}{v^*_i}\frac{dv_i}{de_j} = \epsilon^*_x C^*_x + I$$

Since flux and metabolite control coefficient matrices describe the response of the steady-state to changes in enzyme expression, our altered versions describe the flux response at the particular steady-state in which metabolite concentrations are as close as possible to the unperturbed state. In practice, this has the effect of improving the identifiability of FCCs in the numerical experiments described below. A plot of FCC values obtained via both traditional and modified methods for the following genome-scale model is shown in Fig. S1, indicating that either both methods tend to yield a similar result, or that the identifiability of the link-matrix FCC is particularly poor, with the pseudoinverse FCC pulled close to zero.

With a suitable kinetic framework for calculating steady-state fluxes and concentrations as a function of enzyme expression, we next discuss the prior distributions and likelihood function required for Bayesian inference. The prior distributions represent our belief of possible parameter values before any experimental data is collected. For metabolite elasticity matrices we assume that for any given reaction, reactants are likely to be associated with a positive elasticity, while products likely have a negative elasticity (increasing reactant concentration increases reaction rate, while increasing product concentration decreases reaction rate). Alternatively, we assume that if a metabolite does not directly participate in a reaction, it can only regulate the reaction if it appears in the same sub-cellular compartment. We denote the vectors c_m and c_r of metabolite and reaction compartments, respectively. Since regulation of enzymatic reactions by otherwise nonparticipatory metabolites is relatively rare, we place a sparsity-inducing prior on its elasticity value. This distribution encourages elasticities for off-target metabolites to take values near zero, unless strong experimental evidence for a regulator interaction is present. The combined priors for enzyme elasticities can then be expressed through the following functional form, also depicted graphically in Figure 1.

$$\epsilon^*_{x,ji} \sim \begin{cases} \mathrm{sign}(-N_{ij}) \cdot \mathrm{HalfNormal}(\sigma = 1) & \text{if } N_{ij} \neq 0 \\ \mathrm{Laplace}(\mu = 0, b = 0.05) & \text{if } N_{ij} = 0 \text{ and } c_{m,i} = c_{r,j} \\ 0 & \text{if } N_{ij} = 0 \text{ and } c_{m,i} \neq c_{r,j} \end{cases} \tag{6}$$

We note that the assumption that reactants and products must take positive and negative elasticity values, respectively, can be relaxed by replacing the half-normal distribution in Equation 6 with a skew-normal distribution with a positive shape parameter. This distribution reflects the belief that while reactants typically take positive elasticities, rare cases may exist where substrate inhibition results in a negative slope of reaction rate with respect to substrate concentration. In practice, however, this choice of a prior distribution results in less robust convergence to a stable posterior distribution and was avoided in higher-dimensional inference problems.

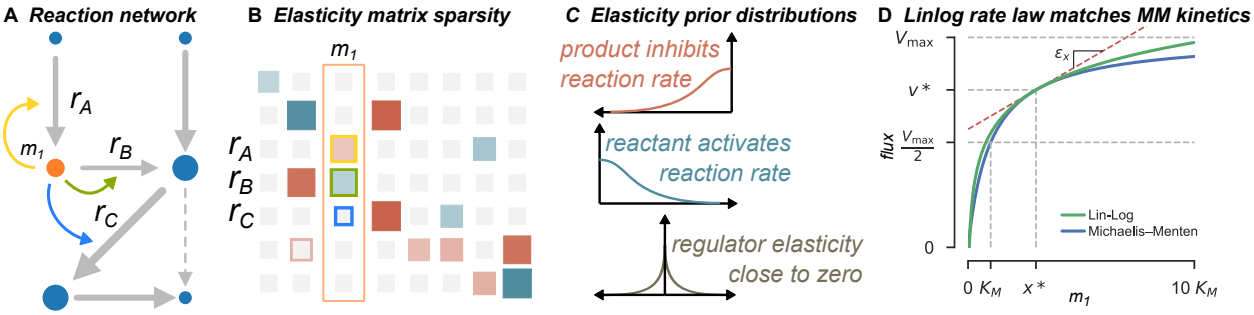

Figure 1: Overview of the modeling framework. (A) The stoichiometry of the reaction network is used to determine prior distributions for the elasticity parameters (B), represented in matrix form. (C) For a metabolite m_1, the prior distribution has predominatly negative support for reactions in which the metabolite is a product (r_A), positive support for reactions in which the metabolite is a reactant (r_B), and a zero-centered, sparsity inducing prior for reactions in which the metabolite does not participate (r_C). (D) The resulting rate law for linlog kinetics closely approximates Michaelis-Menten kinetics in the vacinity of the reference state.

An explicit likelihood function can be formed by constructing a statistical model for the observed data. We assume that observed data are normally distributed around the calculated steady-state metabolite and flux values.

$$\begin{aligned} \chi_{obs} &\sim \text{Normal}(\chi, \sigma_x^2) \\ \hat{v}_{obs} &\sim \text{Normal}(\hat{v}, \sigma_v^2) \end{aligned} \tag{7}$$

Experimental errors, σ_x and σ_v, can either be set explicitly or estimated from the data. For smaller-scale examples, we place half-normal priors on these variables, while for larger datasets we set these values explicitly to improve numerical stability. We also note that for genome-scale multiomics data, computational stability can be improved by fitting log-transformed normalized fluxes,

$$\log \hat{v}_{obs} \sim \text{Normal}(\log \hat{v}, \sigma_v^2),$$

so that flux, metabolite, and enzyme expression data fall on similar orders of magnitude. While this assumption comes at the cost of preventing measured fluxes from reversing directions between perturbed states, this restriction was not significant for the examples considered in this study. However, this framework could be easily extended to handle situations where a measured flux reverses directions between experimental conditions. Most simply, the reversible reaction could be withheld from the log transform and fit in linear space. Alternatively, if separate estimates for the forward and reverse flux could be obtained, as is often the case in ^{13}C labeling studies, the reaction could be decomposed and modeled separately as irreversible forward and reverse reactions.

Once the prior distribution and likelihood model have been specified, the remaining task is to numerically estimate posterior distributions in elasticity parameters Towards this goal, two inference algorithms were used. The No-U-turn sampler (NUTS) (Hoffman and Gelman, 2014), as a variant of Hamiltonian Monte Carlo (HMC), constructs an iterative process (a Markov chain) that eventually converges to the true posterior distribution. Markov chain Monte Carlo methods, while accurate, are computationally intensive and likely limited in application to smaller-scale models and datasets. While the major computational bottleneck in metabolic ensemble modeling (integrating an ODE until steady state) has been removed, calculating the likelihood function still involves a separate linear solve for each steady-state experimental condition. Therefore as model sizes approach the genome-scale, HMC methods quickly become computationally infeasible. Variational methods, however, offer an alternative to Markov chain Monte Carlo methods that can scale to models with thousands of parameters. Automatic differentiation variational inference (ADVI)approximates the posterior distribution by a simple, closed-form probability (typically Gaussian),

then estimates parameters for the approximate posterior to minimize the distance between the true and approximated distribution.

Characterization of an *in vitro* linear pathway

While the primary purpose of the proposed modeling framework is to parameterize genome-scale kinetic models from large, multiomics datasets, we first demonstrate the method on a simple pathway. We re-fit a simple three-reaction model (Wu et al., 2004) to steady-state *in vitro* flux and concentration data for a reconstructed subsection of lower glycolysis (Giersch, 1995). A schematic of the considered pathway is shown in Figure 2A. The model consists of two internal metabolite species, 2-phosphoglycerate (2PG) and phosphoenolpyruvate (PEP), and two metabolites with externally-controllable concentrations, adenosine diphosphate (ADP) and 2,3-bisphosphoglycerate (BPG). The model consists of three reactions in series, phosphoglycerate mutase (PGM), enolase (ENO), and pyruvate kinase (PK); therefore each carries the same flux at steady-state. The dataset consists of 19 separate experiments, each of which contains the enzyme loadings (concentrations) and external metabolite concentrations together with the resulting internal metabolite concentrations and steady-state flux.

Since all metabolites (including external species) are present in the same compartment, all elasticities are allowed to have allosteric interactions normalized with Laplace priors. Measurement errors in fluxes and metabolite concentrations were fit by the inference algorithm by placing a half-normal prior distribution on the σ values in Equation 7. The same reference steady-state was chosen (experiment 2) as was done by Wu et al., 2004.

Using NUTS, stable traces were found across four independent chains, indicating that each trace converged to the true posterior distribution (Figure 2B, Figure 5). For this small-scale example, NUTS sampling took less than 10 minutes on a single computer. Applying ADVI to this example, the evidence lower bound (ELBO), a measure of the closeness of fit between the approximated and true posterior distribution, converged after approximately 10,000 iterations of stochastic gradient descent (Figure 6). A full 25,000 iterations were completed in under 40 seconds on a single computer.

Comparing the results of the two inference methods indicates that both methods yield similar conclusions. ADVI fits a mean-field approximation - i.e., each parameter's posterior is represented by a mean and standard deviation. Comparing the mean and variance of the elasticity posterior distributions from the two different approaches, we notice that while the mean values agree closely, ADVI underestimates the variance for many parameters (Figure 7). This underestimation is typical of mean-field ADVI (Kucukelbir et al., 2017), and might be alleviated in the future through more advanced variational methods (Rezende and Mohamed, 2015). A posterior predictive check for both inference methods indicates that the measured experimental data is well-captured by the model (Figure 2C). Despite normalizing priors on off-target regulation, all elasticity values in the internal metabolite elasticity matrix, ϵ_x^*, were confidently nonzero (as determined by whether the 95% highest posterior density (HPD) interval overlapped zero). Inferred regulatory interactions, which were all consistent between both inference methods, are shown in gray in Figure 2A. These include a strong repression of PGM by PEP, and a weaker repression of PK by 2PG. These off-target regulatory interactions (with similar elasticity values) were also found through the original linear regression approach of Wu *et al.*, 2004. For the external metabolite species, only one of the four possible off-target regulatory interactions, ADP activation of PGM, resulted in a posterior distribution that was confidently nonzero. This relatively weak interaction was rejected by the original linear regression method through a combination of experimental and mathematical reasoning, but underscores that interactions between metabolites and fluxes are inherently difficult to predict from this type of data: direct vs. indirect interactions often look similar, and causality is often impossible to establish. Notably, the posterior distribution as estimated via NUTS contains a rich amount of information on the identifiability of elasticity values (Figure 2D). Strong correlations in estimated parameters typically occur where the two elasticities share either a metabolite or reaction.

The main goal of the method is determining posterior predictive distributions in flux control coefficients, *i.e.*, determining major control points that determine how flux is distributed in the pathway. In this example,

since steady-state fluxes are constrained to be equal for all three reactions, the FCCs are a vector of three coefficients that determine whether increasing enzyme concentration will increase or decrease pathway flux. Figure 2E shows posterior distributions in FCCs as estimated with both inference methods. These are compared against FCC distributions resulting from only the prior distributions on elasticity parameters, without considering any experimental results. Prior distributions are similar between all three enzymes and indicate no structural bias on flux control values. The data therefore indicate that pyruvate kinase (PK) is the limiting enzyme at the reference state. We also compare our FCC estimates against those originally calculated via linear regression, assuming specific allosteric interactions between metabolites and enzymes that differ from those found to be significant through our approach. Our estimates of flux control coefficients closely match those found by Wu *et al.*, 2004, indicating that systems-level properties are relatively insensitive to the particular parameterization of allosteric regulation.

The close agreement of the estimates provided by the approximate ADVI method to the more accurate NUTS sampling in elasticities and flux control coefficients is an important result. As most applications in metabolism involve a larger reaction network, approximate inference methods are likely the only techniques that will scale to biologically-relevant *in vivo* examples. We therefore rely only on these variational techniques for subsequent examples that deal with larger reaction networks.

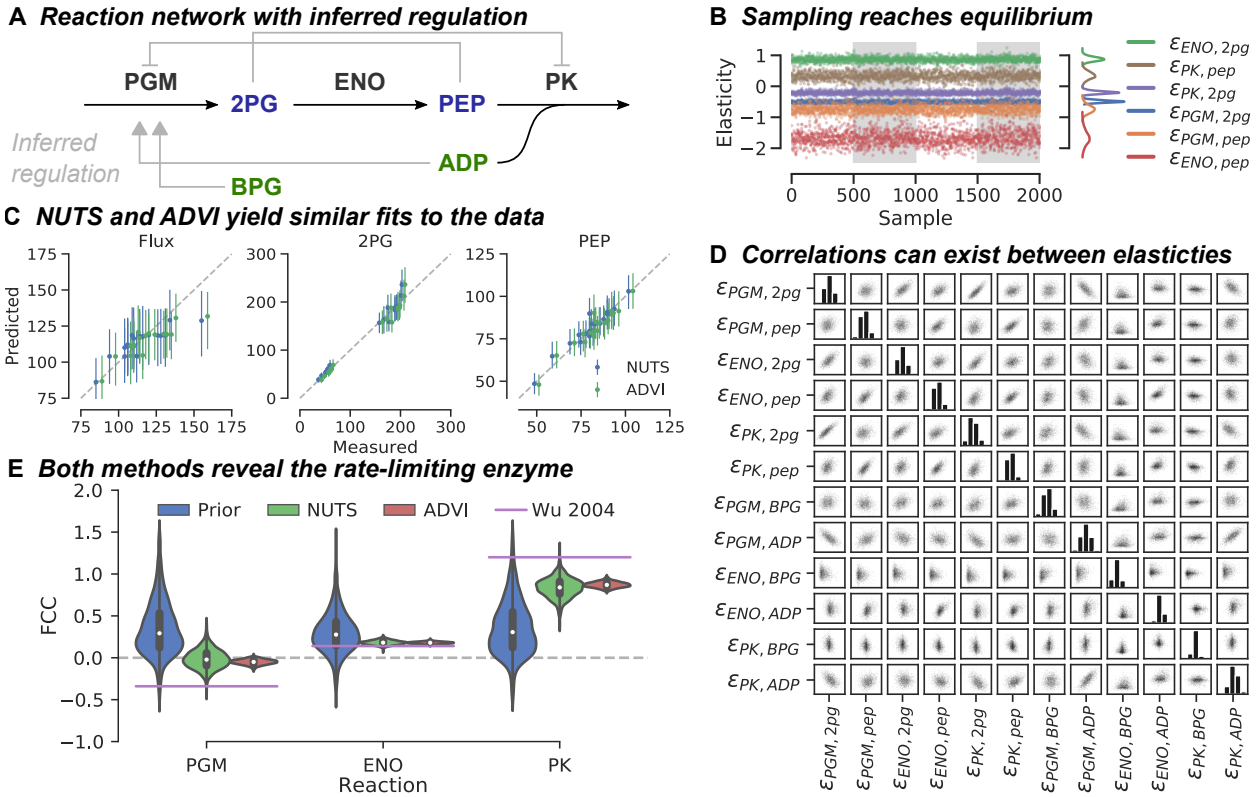

Figure 2: *In vitro* pathway inference. (A) Schematic of the considered pathway. Inferred allosteric interactions are shown in gray, in which arrows indicate an activation, while bar-headed lines indicate inhibition. (B) Traces for ϵ_x^* values as estimated by NUTS. Samples come from four parallel chains stacked together as indicated by the shaded regions. Resulting posterior densities are indicated by the inset on the right. (C) Posterior predictive distributions of steady-state flux and metabolite concentrations. Points represent medians of the posterior predictive distributions, with lines extending to cover the 95% HPD interval. Slight jitter was added to differentiate the distributions as estimated by NUTS and ADVI. (D) Pairplot of the posterior distributions of elasticity variables as estimated via NUTS. Strong correlations can exist between fitted parameters, which are missed by the mean-field ADVI approximation. (E) Violin plot of distributions in posterior flux control coefficients. Median and inner quartile range are indicated by the inner box plots, overlaid on a kernel density plot of each distribution.

Determining optimal enzyme targets from limited data

We next demonstrate how the inference framework can be used to suggest enzyme targets in a many-reaction network, including branched reaction networks and conserved metabolite pools. The problem we consider was previously examined through ensemble metabolic modeling (Contador et al., 2009), and involves predicting what manipulations might further increase lysine production in engineered *E. coli* strains. We therefore replicate the previous ensemble modeling assumptions as closely as possible in order to allow a direct comparison of resulting predictions. The experimental data consists of six sequential enzyme overexpression experiments, all of which were observed to improve l-lysine yields (Kojima et al., 1993). The metabolic model used for inference comprises 44 reactions and 44 metabolites covering central carbon metabolism and lysine production, taken from Contador *et al.*, 2009. A schematic of the reaction network is shown in Figure 3A.

As the goal of the inference approach is to estimate targets for subsequent lysine flux improvement, we chose the reference state for linlog kinetics to be the final, optimized strain with 5 overexpressed enzymes.

Following the assumptions made in (Contador et al., 2009), we also assumed each overexpression doubled the concentration of the respective enzyme. Since the reference state was chosen to be the final, optimized strain, perturbed strains had lower relative enzyme concentrations and lysine flux. Reference state fluxes were also taken from previously published values, corresponding to a total lysine yield of 11.2% (Contador et al., 2009).

When analyzed with metabolic ensemble modeling, each successive enzyme overexpression was required to increase lysine flux over the previous base strain. However in our framework, we require a continuous and differentiable likelihood function. We therefore assume that each enzyme overexpression increases lysine flux relative to the wild-type strain by an additional 20% on average (with standard deviation 0.5%). Target relative fluxes after normalizing to the new reference are shown in Table 1. Prior distributions in enzyme elasticities were specified as described in Equation 6, and since the dataset did not include changes in external metabolites, no ϵ_y values were needed. Posterior distributions were estimated using ADVI, with the optimization taking under three minutes. The posterior predictive distribution for each strain closely matches the target lysine fluxes, indicating the model is capable of reproducing the desired behavior (Figure 3B).

Table 1: Assumed relative lysine fluxes for each considered strain, relative to both the wild-type flux and the chosen reference strain (dapA, lysC, dapB, dapD, dapE). Strain designs taken from (Kojima et al., 1993).

Strain	Lysine Flux (relative to WT)	Lysine Flux (relative to reference)
Wild Type (WT)	1.0	0.5
dapA	1.2	0.6
dapA, lysC	1.4	0.7
dapA, lysC, dapB	1.6	0.8
dapA, lysC, dapB, dapD	1.8	0.9
dapA, lysC, dapB, dapE	1.8	0.9
dapA, lysC, dapB, dapD, dapE	2.0	1.0

Using a half-normal distribution with $\sigma = 1$, prior distributions in elasticities associated with stoichiometric metabolite-reaction pairs had a 95% HPD that spanned from 0 to 2. Of these 133 'kinetic' elasticity terms, only twelve were constrained by the experimental data to a 95% highest posterior density that spanned less than 0.75 elasticity units. In addition to these kinetic terms, three regulatory elasticities were identified as confidently nonzero (with a 95% HPD that did not include 0). These regulations include both feedback and feedforward connections, likely used by the model to fine tune the lysine expression to the desired 20% target in response to doubling of enzyme concentration. Posterior distributions for these elasticities are shown in Figure 3C, and confidently inferred regulatory interactions are shown in gray in Figure 3A. Unsurprisingly, nearly all of these elasticities involve reactions and metabolites in the lysine synthesis pathway, the only portion of the model for which overexpression results were provided.

Prior and posterior distributions in flux control coefficients were also calculated. Because only a limited selection of data was available to constrain the elasticity values, only five of the 44 reactions had a flux control coefficient for lysine export whose 95% HPD did not overlap zero. However, these five reactions were the same set of dapA, lysC, dapB, dapD, and dapE previously specified as successful modifications for improving lysine flux. Prior and posterior distributions in FCC values for lysine export are shown in Figure 3D. While previous ensemble modeling results indicated several enzyme overexpressions that might increase lysine pathway flux, our reimplementation demonstrates that the observed sequential overexpression experiments can be recreated through a wide variety of possible parameterizations, with a resulting wide distribution in possible flux responses. These results show that the method generalizes well to the case where insufficient data is provided to constrain model predictions and underscores the importance of rigorously characterizing posterior parameter space to determine the full range of possible model responses.

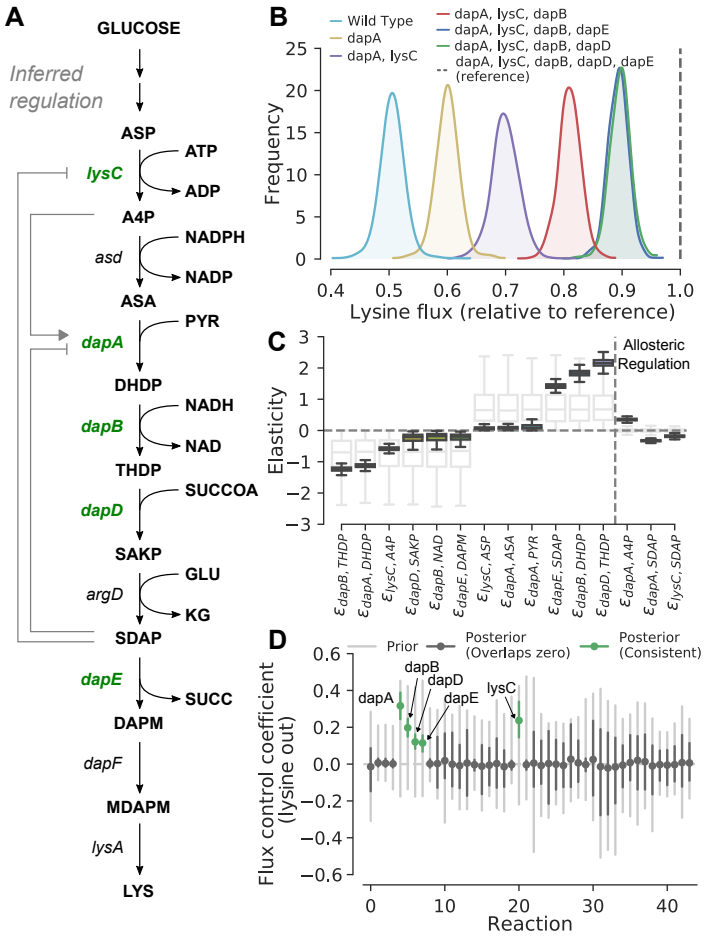

Figure 3: Inference on a medium-scale metabolic network with limited data. (A) Schematic of a portion of the considered metabolic network corresponding to lysine biosynthesis. Reactions shown in green were experimentally determined to improve lysine yields. Regulatory elasticities that were confidently inferred by the model are shown in gray. (B) Posterior predictive distributions for the enzyme overexpression experiments. Since the reference state was chosen as the highest-producing strain, all other strains have a relative lysine flux less than one. (C) Distributions of elasticities informed by the experimental results. Prior distributions for these elasticities are shown in light gray. Allosteric elasticities with Laplace priors that are confidently inferred are shown as the last four entries. (D) Flux control coefficients for each reaction in the model. Prior distributions (light gray) are mostly centered around zero. Posterior distributions (dark gray) are highlighted in green if their 95% HPD does not overlap zero. All lines indicate 95% HPD ranges, dots indicate median.

Informing strain design through multiomics

The main strength of the proposed method is its ability to constrain kinetic parameters using multiomics data, even for large-scale metabolic systems. We therefore demonstrate the method using literature data on metabolomics, proteomics, and quantification of exchange fluxes for 25 different chemostat experiments with yeast (Hackett et al., 2016). The dataset comprises 5 different media conditions, each of which was run at 5 different dilution rates. We adapt a large-scale metabolic model of yeast metabolism that includes many of the genes, metabolites and boundary fluxes of interest (from (Jol et al., 2012)). The adapted model contains 203 metabolites and 240 reactions and was obtained by removing blocked metabolites and reactions under growth on glucose. As the goal in this example is to demonstrate that linlog kinetics are

able to consume large amounts of multiomics data, a reference state near the center of the considered data was chosen, specifically the chemostat with phosphate-limiting media at a 0.11 hr^{-1} dilution rate. Reference fluxes (v^*) were calculated by minimizing error with the experimental boundary measurements while enforcing a nonzero flux through each reaction. In total, the experimental data consists of 1800 metabolite measurements, 792 boundary flux measurements, and 3480 enzyme measurements (omitting the reference state). Since the linlog inference framework only uses relative changes to enzyme, flux, and metabolite concentrations with respect to a reference state, it can naturally ingest large-scale multiomics datasets without the need for absolute quantification. In this example, relative metabolite concentrations are given as log2-transformed values (Boer et al., 2010). Even with an unknown pre-exponential constant A, relative concentrations χ can be calculated from log2-transformed concentrations a and b:

$$\chi = \log\left(\frac{x}{x*}\right) = \log\left(\frac{A2^a}{A2^b}\right) = (a-b)\log 2.$$

Distributions of the transformed data are shown in Figure 4A, indicating that the majority of data falls within one order of magnitude from the reference state value (values shown are natural logs).

In fitting the observed steady-state phenotypes, the model has to account for not only experimental error in measured enzyme concentrations, but also for potential changes in gene expression in unmeasured enzymes. Allowing all enzyme concentrations to vary induces a trade-off where steady-state fluxes are controlled through changes to enzyme expression instead of changes to steady-state metabolite concentrations. While Hackett *et al.*, 2016 have previously shown that metabolic control is mainly determined by metabolite concentrations, some mechanism for adjusting enzyme levels is required to buffer against errors in model formulation and experimental measurements. We therefore place prior distributions on log enzyme concentrations for each condition that drive enzyme changes towards their measured values, or, if the reaction is not measured, towards zero (unchanged):

$$\log \hat{e}_i \sim \begin{cases} \text{Normal}(\mu = \log(\hat{e}_{i,obs}, \sigma = 0.2) & \text{if } e_i \text{ measured} \\ \text{Laplace}(\mu = 0, b = 0.1) & \text{if } e_i \text{ unmeasured} \\ 0 & \text{if reaction } i \text{ uncatalyzed} \end{cases}$$

By placing a Laplace prior on unmeasured enzymes, we create a regularizing effect that penalizes an over-reliance on enzymatic control. Thus, we allow unmeasured enzyme concentrations to deviate from zero only if there is sufficient evidence. The model also has to consider changes in the external metabolite concentrations between media formulations and dilution rates. We therefore place vague priors on the external concentrations of imported substrates, including glucose, phosphate, sulfate, nitrogen, and oxygen:

$$\gamma \sim \text{Normal}(\mu = 0, \sigma = 10).$$

The model parameters therefore include 915 elasticities associated with direct kinetic regulation, 23,684 elasticities associated with potential off-target allosteric regulation, 4,680 enzyme expression levels (195 enzymes over 24 experiments), and 192 external metabolite concentrations (8 metabolites over 24 experiments), for a total of 29,471 parameters. While this number is far greater than the number of experimental data points, regularization forces many of these parameters to be zero.

Observed steady-state metabolite concentrations and fluxes are incorporated through a likelihood model that assumes experimental error is normally distributed around log-transformed metabolite and boundary flux data. Standard deviations were chosen as $\sigma_x = 0.2$ for the metabolite data and $\sigma_v = 0.1$ for the log-transformed fluxes. To improve numerical stability, we also clip the log-transformed, relative experimental data to ± 1.5, such that log-transformed experimental data and model predictions greater than 1.5 or less than -1.5 are replaced by ± 1.5. This process has the effect of reducing the influence of extreme points, especially in regimes far from the reference state that are unlikely to be fit well by the linlog approximation. However, the model is still required to predict the directionality and high-magnitude of these points correctly. Fitting the model using ADVI required 40,000 iterations of stochastic gradient descent, taking approximately five hours on a single compute node (Figure 8). The model is able to recapture a vast majority of the variance seen in the experimental fluxes, enzymes, and metabolites. Median absolute errors

between the model predictions (median of the posterior predictive distribution) and experimental data points are 0.124, 0.0952, and 0.0186 for log-transformed metabolite, flux, and enzymes, respectively, for normalized points that fall within the $[-1.5, 1.5]$ window.

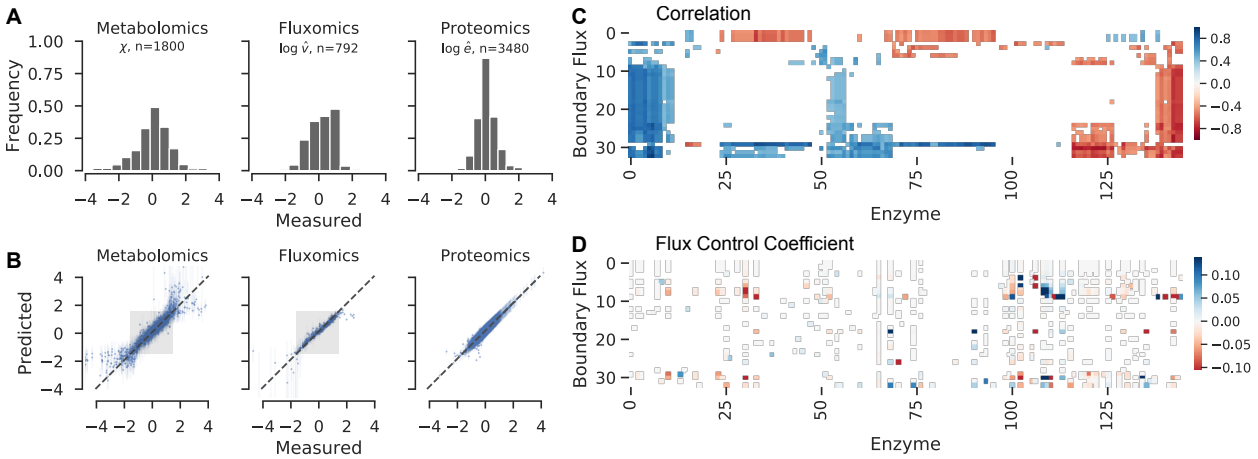

Figure 4: Parameterizing a genome-scale kinetic model with multiomics data (A) Distributions in log-transformed experimental data after normalizing with respect to the phosphate-limited reference state. (B) Posterior predictive distributions after fitting with ADVI. Higher weight was given to experimental datapoints close to the reference state (± 1.5) as indicated by the gray boxes. (C) Heat map of correlation coefficient between experimental enzyme measurements (x-axis) and experimental boundary flux measurements. Boundary fluxes and enzymes are sorted with hierarchical clustering. (D) Heat map of flux control coefficients as estimated from posterior parameter distributions. Boundary flux and enzyme ordering match those determined in (C). Colors represent medians of the posterior predictive distributions, FCCs with a direction that could not be confidently determined (having a 95% HPD that crossed zero) are colored white.

Posterior distributions in fitted parameter values indicate that the model is able to fit the observed experimental data while using relatively few of the additional regulatory parameters. Of the 23,684 regulatory elasticities, only 153 (0.65%) were confidently nonzero. However, we note that determining mechanistically accurate regulatory interactions from observations of steady-state flux behavior is inherently difficult. For instance, for a regulatory pathway in which A regulates B and B regulates C, identifiability issues might cause the pathway to be modeled as A regulates B and A directly regulates C. While poorly identifiable, the impacts of these alternative regulatory topologies on flux control coefficients are largely similar. Of the 50 unmeasured enzymes, only half were nonzero in at least one experimental condition. Overall, only 35% of the available unmeasured enzyme expressions differed from their reference state value.

We next look at what the model is able to learn about the systems-level control of yeast metabolism. A common goal in strain engineering is to find gene targets for increasing the yield of a given metabolic product. We therefore look at relationships between enzymes with measured protein concentrations and measured boundary fluxes. In a traditional statistical approach, correlations between enzyme levels and metabolite fluxes might be used to further enhance production of a desired metabolite. Figure 4C shows a heat map of Pearson correlation coefficients between enzyme expression levels (as determined through proteomics) and measured metabolite boundary fluxes. A permutation test was performed to determine correlations significant at the $\alpha = 0.05$ confidence level; non-significant correlations were masked from the array. In this map, hierarchical clustering is used to reveal clear groups of metabolites and enzymes that vary together in the experimental data. A larger version of this image, with labelled axes, is shown in Figure 9. However, correlations between proteins and metabolite boundary fluxes do not necessarily imply that a particular enzyme is involved in directing flux to a particular product. For instance, several of the highest correlations exist between methionine synthase and relatively distant amino acid products alanine,

arginine, and histidine. The top ten enzyme-boundary flux correlations are shown in Table 2.

Table 2: Largest significant correlations between measured enzymes and measured boundary fluxes.

Enzyme	Boundary	ρ
Methionine synthase	L-Alanine	0.910
Glycine hydroxymethyltransferase, reversible	L-Alanine	0.884
Glycine hydroxymethyltransferase, reversible	Pyruvate	0.866
3',5'-bisphosphate nucleotidase	Succinate	0.854
Methionine synthase	L-Arginine	0.850
Argininosuccinate lyase	Succinate	0.844
Phosphoserine transaminase	Succinate	0.844
Asparagine synthase (glutamine-hydrolysing)	Succinate	0.839
Methionine synthase	L-Histidine	0.835
Imidazole-glycerol-3-phosphate synthase	Succinate	0.835

Flux control coefficients as estimated through the proposed method therefore offer an alternative approach for determining potential enzyme targets that more systematically considers the effects of metabolic stoichiometry and kinetics. Before considering posterior distributions in flux control coefficients, we first look at whether the prior assumptions on enzyme elasticities and model stoichiometry esult in any confidently nonzero values. From the prior predictive distribution, only 6 enzyme-boundary flux pairs have a significantly nonzero FCC, and typically involve reactions directly associated with metabolite production. For instance, a positive flux control coefficient is associated with asparagine synthase and valine transaminase on asparagine and valine export, respectively. A heat map of FCCs calculated from the fitted posterior elasticity matrix is shown in Figure 4C, in which FCCs that have a 95% HPD that includes zero are colored white. Unlike the map of correlation coefficients, FCCs result in a much sparser matrix of inferred connections between enzyme concentration and steady-state flux. However, these coefficients are much more interpretable as direct causality between enzyme expression and increased downstream flux. The top 10 largest, identifiable flux control coefficients are shown in Table 3. Some pairs of enzymes and boundary fluxes, *i.e.* glycerol-3-phosphate dehydrogenase enhancing glycerol production, are direct upstream enzymes for the boundary flux in question. However, since linear pathways can have an uneven distribution of flux control coefficients, determining the rate-limiting step in biosynthesis pathways is an important result. Other confident FCCs represent more indirect effects, for instance the consumption of the upstream phosphoenolpyruvate in 3-phosphoshikimate 1-carboxyvinyltransferase reducing the export of pyruvate.

Table 3: Largest flux control coefficients for the modulation of measured enzymes on measured boundary fluxes. FCC ranges represent upper and lower bounds of the 95% highest posterior density. Enzyme-boundary pairs that also appear as confident predictions prior to including experimental data are omitted.

Enzyme	Boundary	FCC Range
Glycerol 3 phosphate dehydrogenase (NAD)	Glycerol	$[+0.661, +0.867]$
Triose-phosphate isomerase	Glycerol	$[-0.529, -0.375]$
Threonine aldolase	Glycine	$[+0.323, +0.420]$
Pyruvate decarboxylase	Pyruvate	$[-0.379, -0.308]$
Pyruvate kinase	Pyruvate	$[+0.207, +0.281]$
Phosphofructokinase	Glycerol	$[+0.150, +0.278]$
ATPase cytosolic	Pyruvate	$[+0.178, +0.242]$
Pyruvate kinase	Ethanol	$[+0.184, +0.226]$
Fructose-bisphosphate aldolase	Pyruvate	$[-0.244, -0.156]$

Enzyme	Boundary	FCC Range
3-phosphoshikimate 1-carboxyvinyltransferase	Pyruvate	$[-0.219, -0.157]$

Methods

All simulations were performed in Python using the pymc3 library (Salvatier et al., 2016). Additional code to initialize the elasticity prior matrices and calculate the steady-state metabolites and fluxes is provided at `github.com/pstjohn/emll`, along with jupyter notebooks detailing the use cases described above.

Conclusion

In this study we demonstrate how kinetic models of microbial metabolism can be analyzed through modern probabilistic programming frameworks. In doing so, we have invoked approximate formalisms for enzymatic kinetics; however, we note that similar trade-offs between modeling fidelity and computational efficiency are common throughout biology and chemistry. For instance, while small-scale pathways might be better modeled at a higher level of kinetic theory, a complete kinetic description of a genome-scale kinetic model is likely currently infeasible given available data and computational resources. As biological experiments are becoming increasingly easy to iterate with modeling results, a complete kinetic description of a given pathway may not be as valuable as a reasonable guess as to how to improve a desired phenotype. Computational methodologies that quickly converge to generate a list of potential targets, such as the one proposed in this study, may therefore be essential in keeping up with the growing ease of multiomics experiments. The proposed method can also be run efficiently on consumer-grade hardware, a important factor for applications in industrial microbiology where access to large-scale high performance computing resources is limited.

As the field of variational inference is rapidly evolving, this technique could likely be made more robust or efficient through the use of alternative inference algorithms. For instance, correlations between elasticities were demonstrated through a Hamiltonian Monte Carlo trace but were missed by the corresponding mean-field Gaussian approximation. While fitting a full-rank Gaussian is likely impractical at larger data set sizes, reduced-rank approximations (Rezende and Mohamed, 2015) might offer a suitable compromise between posterior accuracy and computational efficiency. Additionally, inference approaches which only consider a subset of the experimental data might also prove useful. Since each perturbed state involves a new linear solve in calculating the likelihood, stochastic variational inference (Hoffman et al., 2013) or firefly MCMC (Maclaurin and Adams, 2014) might reduce the cost of approximating or drawing samples from the posterior.

Acknowledgement

This work was authored in part by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding was provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Bioenergy Technologies Office via the Agile BioFoundry to PCSJ. We thank Jay Fitzgerald at DOE, Jacob Hinkle, and members of the Agile BioFoundry for helpful discussions.

Supplementary Material

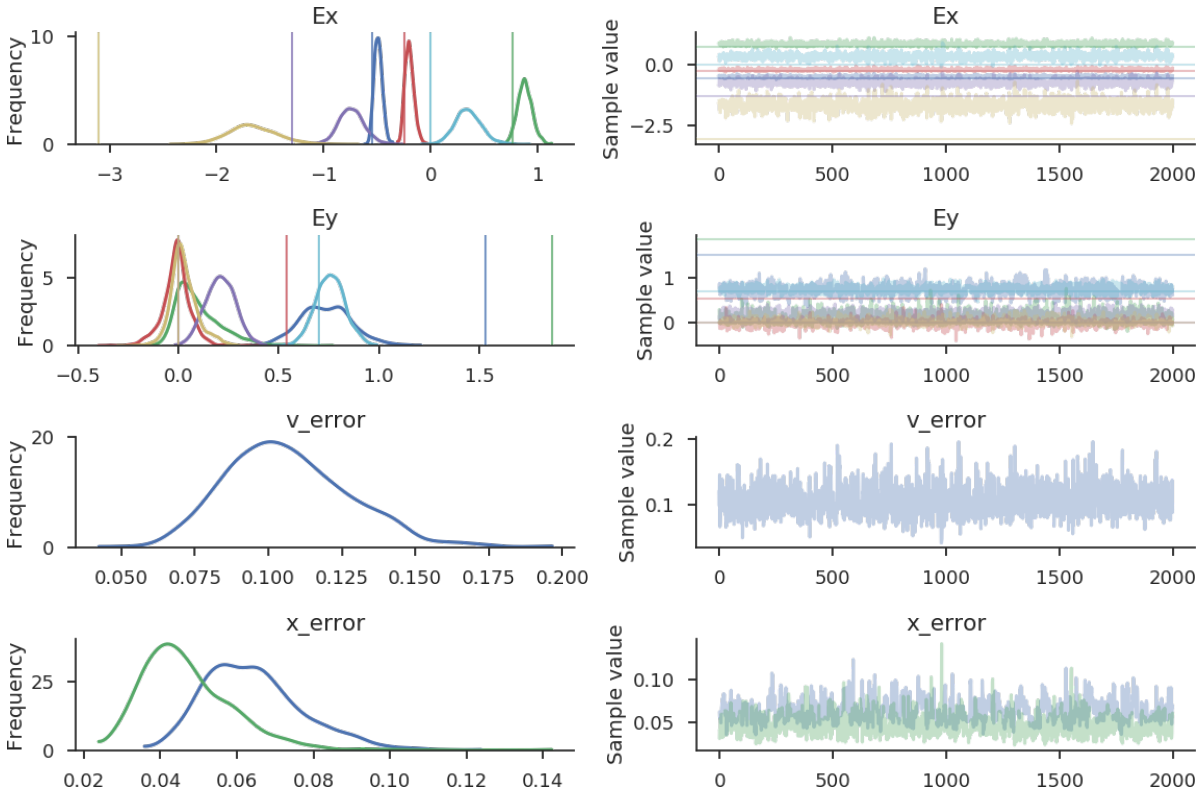

Figure 5: Trace of the NUTS sampler for the *in vitro* dataset. (left) kernel density estimates of each parameter. Vertical bars indicate the values obtained using the multiple linear regression technique of (Wu et al., 2004). (right) Samples from the MCMC sampler

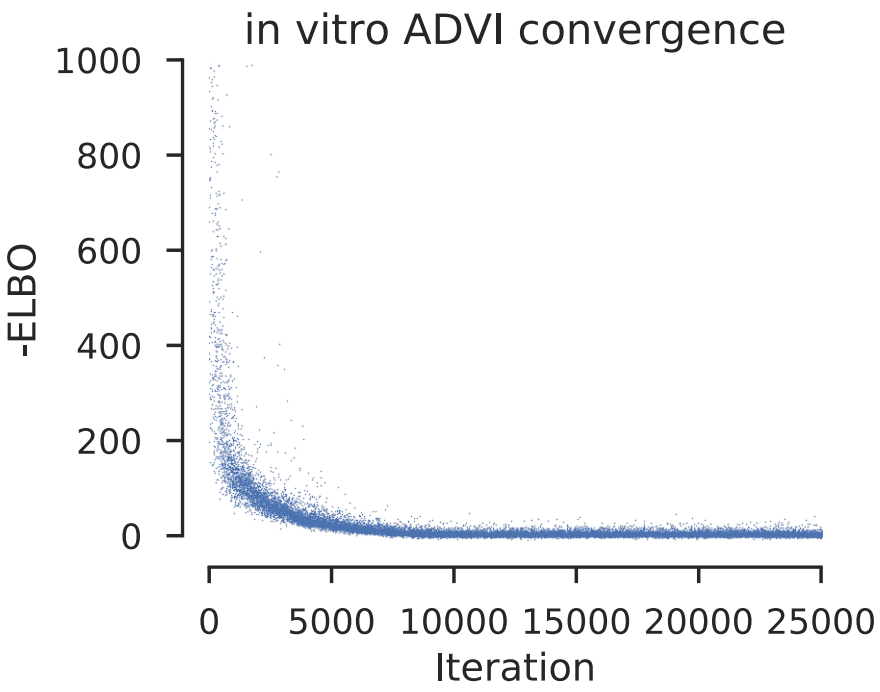

Figure 6: Convergence of the Evidence Lower Bound (ELBO) for the *in vitro* dataset.

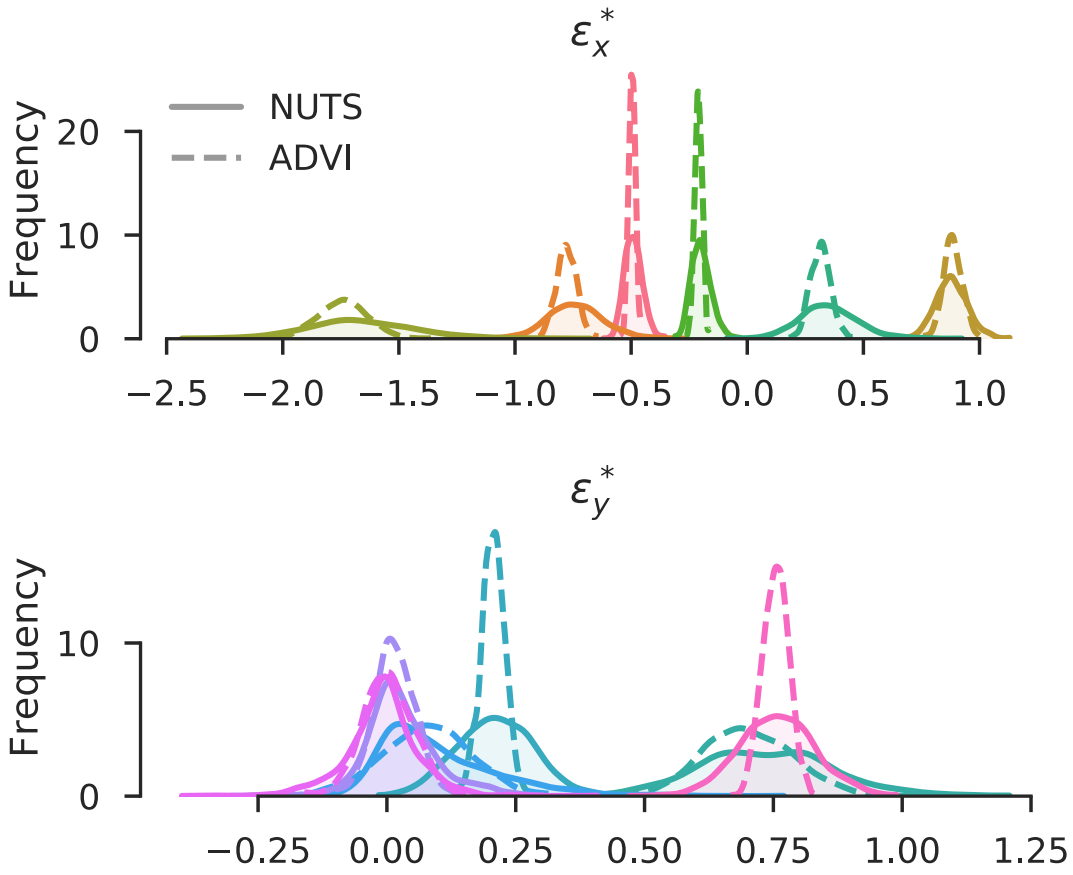

Figure 7: Comparison between posterior distributions for the *in vitro* dataset as estimated by NUTS (solid lines) or ADVI (dashed lines). ADVI posteriors have a similar mean but smaller variance.

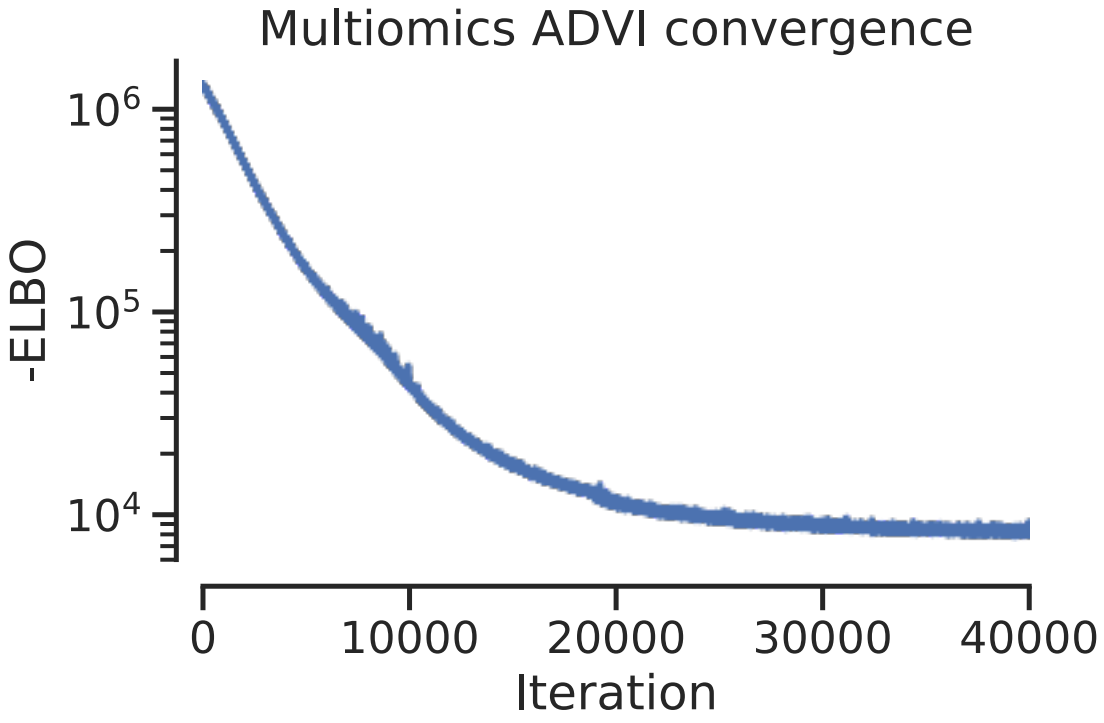

Figure 8: Convergence of the Evidence Lower Bound (ELBO) for the multiomics dataset and yeast metabolic model

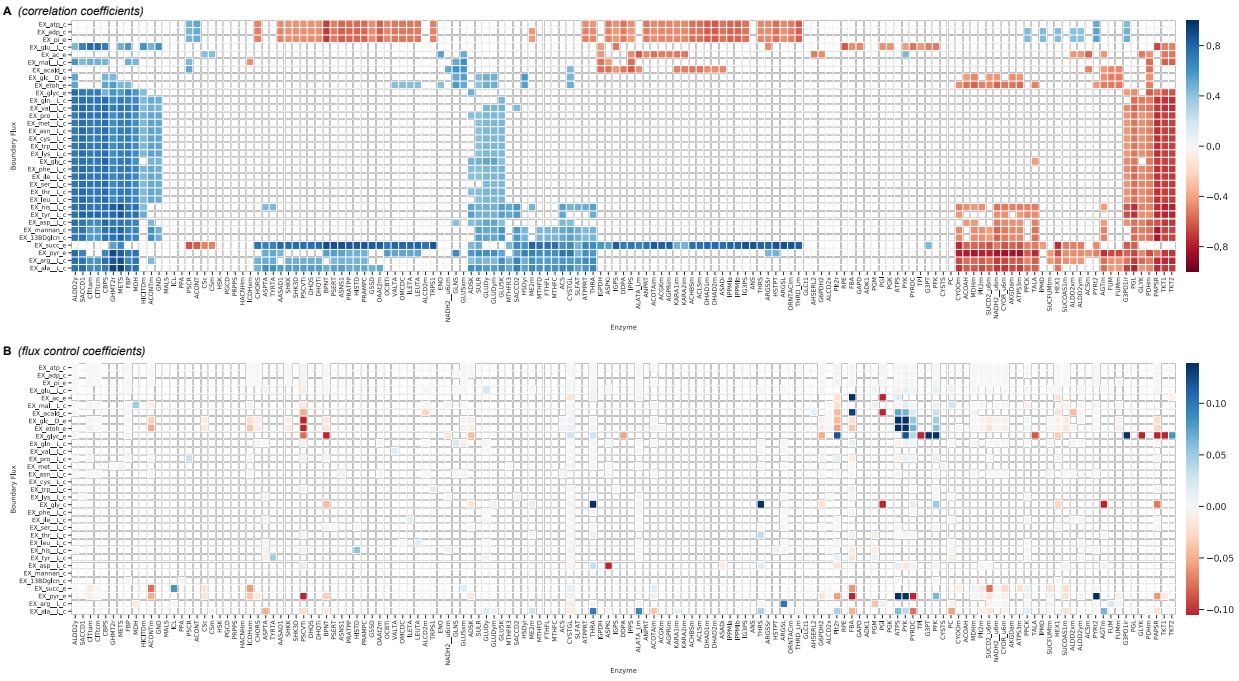

Figure 9: Full heatmaps (with labeled boundary fluxes and enzymes) for the multiomics dataset. Full names for the reaction IDs shown can be found in the detailed model description.

Calculating reverse-mode gradients for regularized linear solve

In order to efficiently perform many inference approaches, forward and reverse-mode gradients for the likelihood function are required. In this method, the least-squares linear solve is a particularly tricky operation for which gradients in some automatic differentiation packages are not automatically supplied. In this section, we therefore derive the necessary matrix equations to calculate forward and reverse mode gradients for the least-norm linear solve, $\chi_{ss} = \mathbf{A}^{\dagger}\mathbf{b}$. In practice, it is much more efficient to calculate this least-norm solution directly (*i.e.*, using the LAPACK routine `dgelsy`) instead of explicitly calculating the pseudoinverse matrix.

Gradients for the least-norm solution are derived by first calculating those for Tikhonov regularization, and subsequently taking the limit as $\lambda \to 0$. Definitions for matrix derivatives are taken from (Giles, 2008). Similar to example 2.3.1 in Giles (2008), the forward derivative for a Tikhonov-regularized linear takes the form

$$C = (\underbrace{A^T A + \lambda I}_{D})^{-1} \underbrace{A^T B}_{E}$$

$$\begin{aligned} dC &= D^{-1}(dE - dD\, C) \quad \text{from } C = D^{-1}E \\ dD &= dA^T A + A^T dA \\ dE &= dA^T B + A^T dB \end{aligned}$$

Substituting into the equation for dC,

$$\begin{aligned} dC &= (A^T A + \lambda I)^{-1}\left(dA^T B + A^T dB - (dA^T A + A^T dA)C\right) \\ dC &= (A^T A)^{-1}\left(dA^T B + A^T dB - (dA^T A + A^T dA)C\right) \end{aligned}$$

Also following Giles (2.3.1), the reverse mode gradient can be found:

$$\begin{aligned} \mathrm{Tr}(\bar{C}^T dC) &= \mathrm{Tr}(\bar{C}^T D^{-1} dE) - \mathrm{Tr}(\bar{C}^T D^{-1} dD\, C) \\ &= \mathrm{Tr}(\bar{C}^T D^{-1} dA^T B) + \mathrm{Tr}(\bar{C}^T D^{-1} A^T dB) \\ &\quad - \mathrm{Tr}(\bar{C}^T D^{-1} dA^T A\, C) - \mathrm{Tr}(\bar{C}^T D^{-1} A^T dA\, C) \\ &= \mathrm{Tr}\left((B - AC)\bar{C}^T D^{-1} dA^T\right) - \mathrm{Tr}(C\bar{C}^T D^{-1} A^T dA) \\ &\quad - \mathrm{Tr}(\bar{C}^T D^{-1} A^T dB) \\ &= \mathrm{Tr}\left[\left(D^{-T}\bar{C}(B - AC)^T - C\bar{C}^T D^{-1} A^T\right) dA\right] \\ &\quad - \mathrm{Tr}(\bar{C}^T D^{-1} A^T dB) \end{aligned}$$

therefore,

$$\begin{aligned} \bar{B} &= \left(\bar{C}^T D^{-1} A^T\right)^T = A D^{-T}\bar{C} \\ \bar{A} &= \left(D^{-T}\bar{C}(B - AC)^T - C\bar{C}^T D^{-1} A^T\right)^T \\ &= (B - AC)\bar{C}^T D^{-1} - \underbrace{A D^{-T}\bar{C}}_{\bar{B}} C^T \\ &= (B - AC)\bar{C}^T D^{-1} - \bar{B} C^T \end{aligned}$$

Since $D = A^T A + \lambda I = (A^T A + \lambda I)^T$, these gradients can be further simplified using the relations

$$\begin{aligned} Dx &= \bar{C} \\ x &= D^{-1}\bar{C} \\ x^T &= \bar{C}^T D^{-1} \end{aligned}$$

After substituting into the equations for $\bar{A}$ and $\bar{B}$, we are left with

$$\begin{aligned} \bar{B} &= Ax \\ \bar{A} &= (B - AC)x^T - \bar{B}C^T \end{aligned}$$

as the reverse-mode gradients for the least-norm solve $C = A^{\dagger}B$.

References

Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association *112*, 859–877.

Boer, V.M., Crutchfield, C.A., Bradley, P.H., Botstein, D., and Rabinowitz, J.D. (2010). Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. Molecular Biology of the Cell *21*, 198–211.

Brunk, E., George, K.W., Alonso-Gutierrez, J., Thompson, M., Baidoo, E., Wang, G., Petzold, C.J., McCloskey, D., Monk, J., Yang, L., et al. (2016). Characterizing strain variation in engineered e. coli using a multi-omics-based workflow. Cell Systems *2*, 335–346.

Burgard, A.P., Pharkya, P., and Maranas, C.D. (2003). Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng *84*, 647–657.

Chakrabarti, A., Miskovic, L., Soh, K.C., and Hatzimanikatis, V. (2013). Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. Biotechnology Journal *8*, 1043–1057.

Chen, X., Zhang, C., Zou, R., Stephanopoulos, G., and Too, H.-P. (2017). In vitro metabolic engineering of amorpha-4,11-diene biosynthesis at enhanced rate and specific yield of production. ACS Synthetic Biology *6*, 1691–1700.

Contador, C.A., Rizk, M.L., Asenjo, J.A., and Liao, J.C. (2009). Ensemble modeling for strain development of l-lysine-producing escherichia coli. Metab Eng *11*, 221–233.

Cotten, C., and Reed, J.L. (2013). Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. BMC Bioinformatics *14*, 32.

Davis, R., Tao, L., Tan, E.C.D., Biddy, M.J., Beckham, G.T., Scarlata, C., Jacobson, J., Cafferty, K., Ross, J., Lukas, J., et al. (2013). Process design and economics for the conversion of lignocellulosic biomass to hydrocarbons: Dilute-acid and enzymatic deconstruction of biomass to sugars and biological conversion of sugars to hydrocarbons (Office of Scientific; Technical Information (OSTI)).

Delgado, J.P., and Liao, J.C. (1991). Identifying rate-controlling enzymes in metabolic pathways without kinetic parameters. Biotechnol Progr *7*, 15–20.

Ehlde, M., and Zacchi, G. (1997). A general formalism for metabolic control analysis. Chem Eng Sci *52*, 2599–2606.

Giersch, C. (1995). Determining elasticities from multiple measurements of flux rates and metabolite concentrations. application of the multiple modulation method to a reconstituted pathway. Eur J Biochem *227*, 194–201.

Giles, M.B. (2008). Collected matrix derivative results for forward and reverse mode algorithmic differentiation. Lecture Notes in Computational Science and Engineering 35–44.

Greene, J.L., Wäechter, A., Tyo, K.E., and Broadbelt, L.J. (2017). Acceleration strategies to enhance metabolic ensemble modeling performance. Biophysical Journal *113*, 1150–1162.

Hackett, S.R., Zanotelli, V.R.T., Xu, W., Goya, J., Park, J.O., Perlman, D.H., Gibney, P.A., Botstein, D., Storey, J.D., and Rabinowitz, J.D. (2016). Systems-level analysis of mechanisms regulating yeast metabolic flux. Science *354*, aaf2786–aaf2786.

Heijnen, J.J. (2005). Approximative kinetic formats used in metabolic network modeling. Biotechnol Bioeng *91*, 534–545.

Heijnen, J., Gulik, W. van, Shimizu, H., and Stephanopoulos, G. (2004). Metabolic flux control analysis of branch points: An improved approach to obtain flux control coefficients from large perturbation data.

Metabolic Engineering *6*, 391–400.

Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. Biophys J *92*, 1792–1805.

Hoffman, M.D., and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research *15*, 1593–1623.

Hoffman, M.D., Blei, D.M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. The Journal of Machine Learning Research *14*, 1303–1347.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M., Hirai, K., Hoque, A., et al. (2007). Multiple high-throughput analyses monitor the response of e. coli to perturbations. Science *316*, 593–597.

Jol, S.J., Kümmel, A., Terzer, M., Stelling, J., and Heinemann, M. (2012). System-level insights into yeast metabolism by thermodynamic analysis of elementary flux modes. PLoS Computational Biology *8*, e1002415.

Kojima, H., Ogawa, Y., Kawamura, K., and Sano, K. (1993). Method of producing L-lysine by fermentation (US Patent US6040160A).

Kresnowati, M.P., Winden, W.A. van, and Heijnen, J.J. (2005). Determination of elasticities, concentration and flux control coefficients from transient metabolite data using linlog kinetics. Metabolic Engineering *7*, 142–153.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D.M. (2017). Automatic differentiation variational inference. Journal of Machine Learning Research *18*, 14:1–14:45.

Li, S., and Petzold, L. (2000). Software and algorithms for sensitivity analysis of large-scale differential algebraic systems. Journal of Computational and Applied Mathematics *125*, 131–145.

Maclaurin, D., and Adams, R.P. (2014). Firefly Monte Carlo: exact MCMC with subsets of data. Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence 543–552.

Marcellin, E., and Nielsen, L.K. (2018). Advances in analytical tools for high throughput strain engineering. Current Opinion in Biotechnology *54*, 33–40.

Meer, R. van der, Westerhoff, H., and Van Dam, K. (1980). Linear relation between rate and thermodynamic force in enzyme-catalyzed reactions. Biochimica et Biophysica Acta (BBA) - Bioenergetics *591*, 488–493.

Neal, R.M. (2010). MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo *54*, 113–162.

Nikerel, I.E., Winden, W.A. van, Gulik, W.M. van, and Heijnen, J.J. (2006). A method for estimation of elasticities in metabolic networks using steady state and dynamic metabolomics data and linlog kinetics. BMC Bioinformatics *7*, 540.

Nikerel, I.E., Winden, W.A. van, Verheijen, P.J., and Heijnen, J.J. (2009). Model reduction and *a priori* kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics. Metabolic Engineering *11*, 20–30.

Nilsson, A., Nielsen, J., and Palsson, B.O. (2017). Metabolic models of protein allocation call for the kinetome. Cell Systems *5*, 538–541.

Onsager, L. (1931). Reciprocal relations in irreversible processes. I. Physical Review *37*, 405–426.

Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat Biotechnol *28*, 245–248.

O'Brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R., and Palsson, B.O. (2014). Genome-scale models of

metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol *9*, 693–693.

Petersen, K.B., and Pedersen, M.S. (2012). The matrix cookbook (Technical University of Denmark).

Rezende, D., and Mohamed, S. (2015). Variational inference with normalizing flows. In Proceedings of the 32nd International Conference on Machine Learning, F. Bach, and D. Blei, eds. (Lille, France: PMLR), pp. 1530–1538.

Rottenberg, H. (1973). The thermodynamic description of enzyme-catalyzed reactions. Biophysical Journal *13*, 503–511.

Saa, P.A., and Nielsen, L.K. (2016). Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. Scientific Reports *6*.

Saa, P.A., and Nielsen, L.K. (2017). Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. Biotechnology Advances *35*, 981–1003.

Saha, R., Chowdhury, A., and Maranas, C.D. (2014). Recent advances in the reconstruction of metabolic models and integration of omics data. Curr Opin Biotech *29*, 39–45.

Salvatier, J., Wiecki, T.V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. PeerJ Computer Science *2*, e55.

Sánchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P., Kerkhoven, E.J., and Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Molecular Systems Biology *13*, 935.

Smallbone, K., Simeonidis, E., Broomhead, D.S., and Kell, D.B. (2007). Something from nothing - bridging the gap between constraint-based and kinetic modelling. FEBS Journal *274*, 5576–5585.

Teusink, B., Passarge, J., Reijenga, C.A., Esgalhado, E., Weijden, C.C. van der, Schepper, M., Walsh, M.C., Bakker, B.M., Dam, K. van, Westerhoff, H.V., et al. (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. European Journal of Biochemistry *267*, 5313–5329.

Tran, L.M., Rizk, M.L., and Liao, J.C. (2008). Ensemble modeling of metabolic networks. Biophys J *95*, 5606–5617.

Visser, D., and Heijnen, J.J. (2002). The mathematics of metabolic control analysis revisited. Metab Eng *4*, 114–123.

Visser, D., and Heijnen, J.J. (2003). Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. Metab Eng *5*, 164–176.

Visser, D., Schmid, J.W., Mauch, K., Reuss, M., and Heijnen, J.J. (2004). Optimal re-design of primary metabolism in *Escherichia coli* using linlog kinetics. Metab Eng *6*, 378–390.

Wang, L., Birol, I., and Hatzimanikatis, V. (2004). Metabolic control analysis under uncertainty: Framework development and case studies. Biophys J *87*, 3750–3763.

Wu, L., Wang, W., Winden, W.A.V., Gulik, W.M.V., and Heijnen, J.J. (2004). A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics. Eur J Biochem *271*, 3348–3359.

Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E., and Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Method Biochem Anal *26*, i255–i260.

Yoshikawa, K., Furusawa, C., Hirasawa, T., and Shimizu, H. (2012). Design of superior cell factories based on systems wide omics analysis. In Systems Metabolic Engineering, (Springer Netherlands), pp. 57–81.

Zhang, W., Li, F., and Nie, L. (2009). Integrating multiple omics analysis for microbial biology: Application and methodologies. Microbiology+ *156*, 287–301.

Zomorrodi, A.R., Lafontaine Rivera, J.G., Liao, J.C., and Maranas, C.D. (2013). Optimization-driven

identification of genetic perturbations accelerates the convergence of model parameters in ensemble modeling of metabolic networks. Biotechnology Journal *8*, 1090–1104.

Zotter, A., Bäuerle, F., Dey, D., Kiss, V., and Schreiber, G. (2017). Quantifying enzyme activity in living cells. Journal of Biological Chemistry *292*, 15838–15848.