Parameter balancing in kinetic models of cell metabolism

Supplementary information: Parameter balancing - method and formulae

Timo Lubitz, Marvin Schulz, Edda Klipp, and Wolfram Liebermeister

Humboldt-Universität zu Berlin, Institut für Biologie
Invalidenstraße 42, D-10115 Berlin
E-mail: wolfram.liebermeister@biologie.hu-berlin.de

Contents

Overview

Parameter balancing is a method to determine complete, consistent sets of model parameters from potentially incomplete and contradictory data. Here we summarize the formulae needed for kinetic models of cell metabolism. We start with the basic concept and then discuss additional features that address typical problems of kinetic data (incompatible measurement conditions, few data available). Finally, we outline potential extensions of the method.

1 Parameter balancing – basic method

The parameters of kinetic models can be mutually dependent, either by their definitions or because they are constrained by thermodynamics. For many kinetic models, all relevant quantities can be split into two subsets: a set of mutually independent *basis quantities* that can be chosen arbitrarily, and a set of derived quantities that can be computed from these basis quantities. This partitioning of the parameter set is exploited by parameter balancing. A key component is the dependence matrix Q, which relates both groups of quantities.

Of course, the shape of this matrix depends on the parameters appearing in a model, which in turn depend on the choice of enzymatic rate laws. To be specific, we refer in this article to the modular rate laws, a family of rate laws for reactions with arbitrary stoichiometries [1], but the same approach also works for most other reversible rate laws. The modular rate laws share the form

$$v_l(c) \quad = \quad u_l \, f_l(c) \frac{T_l(c)}{D_l^*(c)} \tag{1}$$

where v_l is the rate of the l^{th} reaction, u_l is the enzyme level (concentration or amount, depending on the definition), and f_l and D_l^* are positive terms describing the regulating or saturating influence of metabolites with concentrations c_i (for detailed formulae, see [1]). The term that is most relevant for parameter balancing is the numerator

$$T_l \quad = \quad k_l^{\text{cat}+} \prod_i \left(\frac{c_i}{k_{li}^{\text{M}}} \right)^{h_l \, n_{il}^+} - k_l^{\text{cat}-} \prod_i \left(\frac{c_i}{k_{li}^{\text{M}}} \right)^{h_l \, n_{il}^-}, \tag{2}$$

containing the forward and backward catalytic constants $k_l^{\text{cat}\pm}$ (in 1/s) and the reactant constants k_{li}^{M} (in mM). The cooperativity factors h_l were introduced in [1] to describe sigmoid kinetics, for non-sigmoid kinetics they have a value of 1. The symbols n_{il}^+ and n_{il}^- denote the (positive) stoichiometric coefficients of substrates and products, respectively. As we shall see, the thermodynamic laws that are incorporated in the term T_l (and in corresponding terms appearing in many other rate laws) are responsible for important dependencies among kinetic parameters.

These dependencies lead to dependence matrices which can be constructed from the choice of quantities used in the model, the metabolic network structure, and the available data. Each row of the matrix corresponds to one quantity and shows how it is computed from the basis quantities. To explain this in detail, we will now consider the various types of quantities and their interrelations.

1.1 Dependencies between model quantities

For metabolites in an ideal solution, the transformed equilibrium constant[1] k^{eq} (in standard notation: $= K'$) of a biochemical reaction can be computed from the transformed standard chemical potentials of its reactants. Its natural logarithm reads

$$\ln k_l^{\text{eq}} \quad = \quad -\frac{1}{RT} \sum_i n_{il} \, \mu_i'^{\circ}, \tag{3}$$

[1] The model reactions are assumed to involve biochemical reactants (e.g. ATP) rather than individual protonation states (like ATP^{4-}). This implies that the equilibrium constants or chemical potentials considered in a model are actually *transformed* thermodynamic quantities, which refer to the biochemical reactants and depend on the pH value [2, 3]. Accordingly, protons (H^+) must not appear in the reaction formulae and all data used have to refer to transformed quantities. For an excellent introduction to this topic, see the review by Alberty [4].

where n_{il} is the stoichiometric coefficient of metabolite i in reaction l, $\mu_i'^\circ$ is its transformed standard chemical potential, R is Boltzmann's gas constant ($R \approx 8.31\,\mathrm{J/(mol\,K)}$), and T is the absolute temperature. In contrast to usual practice in kinetic models, we also consider the non-balanced ("external") metabolites. If the reaction network contains thermodynamic cycles (i.e., if the stoichiometric matrix $N = (n_{il})$ has a non-empty kernel matrix K satisfying $N\,K = 0$), Eq. (3) implies that some of the equilibrium constants are related by Wegscheider conditions [5].

Another group of constraints, the Haldane relationships, follows from the fact that reaction rates vanish in equilibrium states. For the modular rate laws [1], the Haldane relationships link the equilibrium constant k^{eq} (dimensionless, but possibly referring to a standard concentration $c^\circ = 1$ mM) with the catalytic constants $k^{\mathrm{cat}\pm}$ (in 1/s) and the Michaelis constants k^{M} (in mM):

$$h_l\,\ln k_l^{\mathrm{eq}} \;=\; \ln k_l^{\mathrm{cat}+} - \ln k_l^{\mathrm{cat}-} + \sum_i h_l\,n_{il}\ln k_{li}^{\mathrm{M}}. \tag{4}$$

After combining Eqs. (3) and (4), the forward and reverse catalytic constants can be expressed as

$$\ln k_l^{\mathrm{cat}\pm} \;=\; \ln k_l^{\mathrm{V}} \mp \frac{h_l}{2}\sum_i n_{il}(\mu_i'^\circ/RT + \ln k_{li}^{\mathrm{M}}) \tag{5}$$

where the velocity constant $k_l^{\mathrm{V}} = \sqrt{k_l^{\mathrm{cat}+}k_l^{\mathrm{cat}-}}$ (the geometric mean of both catalytic constants), has been introduced as a new basis quantity.

Apart from the kinetic constants, the set of basis quantities can also comprise metabolite concentrations c_i and enzyme concentrations u_l representing one or several metabolic states, which need not be stationary. From these concentrations, we can derive a number of other state-dependent quantities. First, the forward and backward maximal velocities $v^{\mathrm{max}\pm} = u\,k^{\mathrm{cat}\pm}$ (with enzyme concentration u) can be expressed as

$$\ln v_l^{\mathrm{max}\pm} \;=\; \ln u_l + \ln k_l^{\mathrm{V}} \mp \frac{h_l}{2}\sum_i n_{il}(\mu_i'^\circ/RT + \ln k_{li}^{\mathrm{M}}). \tag{6}$$

Just like the reaction rates in the modular rate laws, the maximal velocities have units of mM/s. An extension to reaction rates measured as amounts per time is discussed in section 4. Second, the transformed chemical potentials μ_i' for ideal mixtures can by expressed by

$$\mu_i' \;=\; \mu_i'^\circ + RT\cdot\ln c_i, \tag{7}$$

where the concentration c_i must be given in units of the standard concentration $c^\circ = 1$ mM. For charged molecules (ions), we consider the electrochemical potential comprising the additive term $z\,F\,\Phi$, with the charge number z, the Faraday constant $F \approx 96485$ C/mol, and the electric potential Φ of the compartment in which the ions are residing. Given the chemical or electrochemical potentials, we can compute their negative difference along a reaction

$$A_l \;=\; -\Delta_{\mathrm{r}}G' = -\sum_i n_{il}\mu_i'^\circ - RT\sum_i n_{il}\,\ln c_i, \tag{8}$$

called the *reaction affinity* [6].

1.2 The dependence matrix Q

If we take a closer look at equations (3) through (8), we easily notice that some of the quantity types – the basis quantities – can be freely chosen (e.g. according to data), while all others are derived from them (see Table 1). Moreover, if we use logarithms for all quantities that do *not* represent energies (in units of kJ/mol) – as we did in the equations – all these dependencies are linear. After collecting the basis quantities in a vector q and all available model quantities in a vector x, we can express all dependencies by a simple equation

$$x = Q\,q, \tag{9}$$

	μ'°	$\ln \mathbf{k}^{\mathrm{V}}$	$\ln \mathbf{k}^{\mathrm{M}}$	$\ln \mathbf{u}$	$\ln \mathbf{c}$
$\ln \mathbf{k}^{\mathrm{eq}}$	✓				
$\ln \mathbf{k}^{\mathrm{cat}\pm}$	✓	✓	✓		
$\ln \mathbf{v}^{\mathrm{max}\pm}$	✓	✓	✓	✓	
μ'	✓				✓
$\mathbf{A}$	✓				✓

Table 1: Dependencies between basis quantities (columns) and derived quantities (rows).

where Q, called dependence matrix, follows from the choice of quantity types to be considered and from the structure of the kinetic model. Given the model, we can construct the dependence matrix Q

$$
\begin{array}{r}
 \\
\mu'^{\circ} \\
\ln k^{\mathrm{V}} \\
\ln k^{\mathrm{M}} \\
\ln k^{\mathrm{A}} \\
\ln k^{\mathrm{I}} \\
\ln u \\
\ln c \\
\ln k^{\mathrm{eq}} \\
\ln k_{+}^{\mathrm{cat}} \\
\ln k_{-}^{\mathrm{cat}} \\
\ln v_{+}^{\mathrm{max}} \\
\ln v_{-}^{\mathrm{max}} \\
\mu' \\
A
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
\mu'^{\circ} & \ln k^{\mathrm{V}} & \ln k^{\mathrm{M}} & \ln k^{\mathrm{A}} & \ln k^{\mathrm{I}} & \ln u & \ln c
\end{array} \\
\left(\begin{array}{ccccc|cc}
\mathbf{I} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \mathbf{I} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \mathbf{I} & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \mathbf{I} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \mathbf{I} & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{I} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{I} \\
\hline\hline
-\frac{1}{RT} N^{\mathrm{T}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{1}{2RT} \tilde{N}^{\mathrm{T}} & \mathbf{I} & -\frac{1}{2} \tilde{N}_{\mathrm{kM}} & \cdot & \cdot & \cdot & \cdot \\
\frac{1}{2RT} \tilde{N}^{\mathrm{T}} & \mathbf{I} & \frac{1}{2} \tilde{N}_{\mathrm{kM}} & \cdot & \cdot & \cdot & \cdot \\
\hline
-\frac{1}{2RT} \tilde{N}^{\mathrm{T}} & \mathbf{I} & -\frac{1}{2} \tilde{N}_{\mathrm{kM}} & \cdot & \cdot & \mathbf{I} & \cdot \\
\frac{1}{2RT} \tilde{N}^{\mathrm{T}} & \mathbf{I} & \frac{1}{2} \tilde{N}_{\mathrm{kM}} & \cdot & \cdot & \mathbf{I} & \cdot \\
\mathbf{I} & \cdot & \cdot & \cdot & \cdot & \cdot & RT \cdot \mathbf{I} \\
-N^{\mathrm{T}} & \cdot & \cdot & \cdot & \cdot & \cdot & -RT \cdot N^{\mathrm{T}}
\end{array}\right)
\end{array}, \tag{10}
$$

with unit matrices $\mathbf{I}$ of different size and the matrix $\tilde{N} = (\tilde{n}_{il})$, where the elements $\tilde{n}_{il} = h_l \, n_{il}$ comprise the stoichiometric coefficients n_{il} and the cooperativity factors h_l. As above, the stoichiometric matrix N refers both to internal and external metabolites. The k^{M} values are matched with the respective reactions by an auxiliary matrix $\tilde{N}_{\mathrm{kM}}$ containing the reordered elements of $\tilde{N}$. For instance, the $\tilde{N}_{\mathrm{kM}}$ matrix for the simple model shown in Figure 1 reads

$$
\begin{array}{r}
 \\
\text{Reaction1} \\
\text{Reaction2}
\end{array}
\begin{array}{c}
\begin{array}{cccc}
\ln k_{11}^{\mathrm{M}} & \ln k_{12}^{\mathrm{M}} & \ln k_{22}^{\mathrm{M}} & \ln k_{23}^{\mathrm{M}}
\end{array} \\
\left(\begin{array}{cccc}
-1 & 1 & \cdot & \cdot \\
\cdot & \cdot & -1 & 1
\end{array}\right)
\end{array} \tag{11}
$$

The rows in the upper part of Q correspond to the basis quantities and form an identity matrix (dots represent zeros), while the derived quantities below are computed as shown by the Eqs. (3) through (8). For simplicity, the dependence matrix Eq. (10) does not show all possible complications. Below, we shall discuss how the dependence of standard chemical potentials and equilibrium constants on pH and temperature can be included into the scheme. Other extensions are discussed in section 4.

1.3 Computing the posterior distribution

Among our model quantities, there are two distinct groups which have to be treated differently: The first group are standard chemical potentials μ'°, chemical potentials μ', and reaction affinities

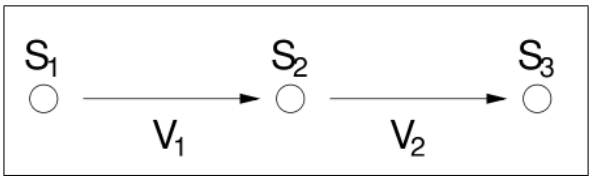

Figure 1: Example network with three substances and two reactions. The four stoichiometric coefficients – corresponding to four k^{M} values – are arranged in the matrix $\tilde{N}_{\mathrm{kM}}$ given by Eq. (11).

A, which are molar energies (in kJ/mol) and can have positive and negative values. The second group comprises all other quantities (in different physical units), which are positive and appear in the equations (3) through (8), and thus in the vectors q and x, as natural logarithms. This distinct representation is called *natural scaling*[2]. Natural scaling allows us to express all dependencies between the model quantities by linear equations, which is the basis of parameter balancing.

Using equation (9), we can compute all model quantities x from the basis quantities q, both given in natural scaling. In parameter balancing, we invert this equation and estimate a set of basis quantities $\bar{q}$ from a (possibly incomplete and contradictory) data vector x^* [7]. By applying Eq. (9) again in forward direction, we then obtain a complete and consistent set $\bar{x}$ of model quantities, supposed to resemble the original data x^*.

The estimation itself is based on Bayesian statistics[3]. We first describe the basis quantities q by a multivariate normal prior distribution (with mean vector $\overline{q}_{\mathrm{prior}}$ and covariance matrix C_{prior}), consider the vector $\overline{x}^*$ as input data, and determine the posterior distribution for q. Their posterior, which is again multivariate normal, shows how plausible certain parameter sets appear in the light of prior and experimental data. To compute it, we need to prepare the following vectors and matrices:

1. The vectors $\overline{x}^*$ and σ_{x} describe the collected kinetic data (arithmetic means and standard deviations in natural scaling). Each entry corresponds to a model quantity, but the data vector may also contain several or no values for some of the model quantities. From the standard errors in σ_x, we obtain the diagonal covariance matrix $C_{\mathrm{x}} = \mathrm{Dg}(\sigma_x)^2$.

2. The prior mean vector $\overline{q}_{\mathrm{prior}}$ and covariance matrix C_{prior} characterize the basis quantities. Usually, we choose an uncorrelated prior with a diagonal covariance matrix $C_{\mathrm{prior}} = \mathrm{Dg}(\sigma_q)^2$. Priors for the different quantity types can be chosen by the user. Our default values, in rough agreement with a statistics over collected data values, are listed in Table 2.

3. We define two dependence matrices: while the usual dependence matrices Q is used for forward prediction, a variant Q^* is needed for parameter estimation. The complete dependence matrix Q follows from the model structure and from the user's choice of relevant quantities. For instance, if only kinetic constants are needed, the columns for metabolite and enzyme concentrations $\ln c$ and $\ln u$ can be omitted. By duplicating and omitting certain rows of Q, we construct the data dependence matrix Q^* with rows corresponding to the entries of x^*.

By maximizing the logarithmic posterior we obtain the posterior mean vector and the posterior

[2]In natural scaling, we could also employ a scaling factor for all energy-like quantities to keep the variance of all quantity types in a similar range. Of course, this scaling factor would then also appear in the formula for the dependence matrix Q.

[3]To become familiar with Bayesian concepts like prior and posterior distribution, we recommend the excellent book on Bayesian data analysis by Gelman et al. [8].

covariance matrix[4]

$$\begin{aligned} \overline{q}_{\text{post}} &= C_{\text{post}} \cdot \left(Q^{*\text{T}}\, C_{\text{x}}^{-1}\, \overline{x}^{*} + C_{\text{prior}}^{-1}\, \overline{q}_{\text{prior}} \right) \qquad (12) \\ C_{\text{post}} &= \left(C_{\text{prior}}^{-1} + Q^{*\text{T}}\, C_{\text{x}}^{-1}\, Q^{*} \right)^{-1}. \qquad (13) \end{aligned}$$

Since both C_{x} and C_{prior} have positive eigenvalues, Eq. (12) implies that the posterior will be narrower than the prior. This gain in information can be quantified, for instance, by the differential Shannon entropies[5]. Finally, the posterior mean and covariance matrix for the complete parameter vector x can be computed by $\bar{x}_{\text{post}} = Q\, \bar{q}_{\text{post}}$ and $C_{\text{x,post}} = Q\, C_{\text{post}}\, Q^{\text{T}}$.

1.4 Conversion between non-logarithmic and logarithmic scale

In natural scaling, all quantities are described by multivariate normal distributions. When preparing the data vector x^* or when inserting the balanced values into a model, we need to convert the quantities between natural and non-logarithmic scale. The energy quantities (in kJ/mol) are always described by normal distributions, but the other quantities will be distributed log-normally[6]. In fact, a statistics of kinetic constants in Brenda shows that log-normality is a fairly realistic assumption, at least for the prior distributions (compare data histograms in [9]).

When converting probability distributions between non-logarithmic and logarithmic scale, we need to pay special attention to the conversion formulae. On logarithmic scale, the arithmetic mean, median, and maximum point of the distribution are identical. By taking the exponential function of this number, we obtain the (non-logarithmic) median, which is identical to the geometric mean. The (non-logarithmic) arithmetic mean, however, can be much larger. The arithmetic mean $\langle \cdot \rangle$ and the variance σ are given by the formulae

$$\begin{aligned} \langle x \rangle &= e^{\langle \ln x \rangle + \frac{1}{2}\sigma^2_{\ln x}}, \\ \sigma_x^2 &= (e^{\sigma^2_{\ln x}} - 1) e^{2\langle \ln x \rangle + \sigma^2_{\ln x}} \qquad (14) \end{aligned}$$

and the conversion in the other direction (to logarithmic scale) reads

$$\begin{aligned} \langle \ln x \rangle &= \ln(\langle x \rangle) - \frac{1}{2} \ln\left(1 + \frac{\sigma_x^2}{\langle x \rangle^2} \right), \\ \sigma^2_{\ln x} &= \ln\left(1 + \frac{\sigma_x^2}{\langle x \rangle^2} \right). \qquad (15) \end{aligned}$$

It is important to use these formulae carefully and to clear distinguish between mean and median values, in particular for the priors and pseudo values, which can show large variances.

After parameter balancing, the basis quantities in natural scaling show a multivariate normal posterior with mean vector $\bar{q}^{\text{post}}$ and covariance matrix C^{post}. After converting the values back to non-logarithmic scale, the vector of median values can be directly inserted into an SBML model as a consistent parameter set. In contrast, the vector of arithmetic mean values will usually not satisfy the constraints and is therefore not a valid parameter set. Alternatively, we can sample parameter sets from the posterior by the formula

$$x^{\text{sample}} = \bar{x}^{\text{post}} + Q\, C_{\text{post}}^{1/2}\, \xi \qquad (16)$$

where $C_{\text{post}}^{1/2}$ denotes the matrix square root of the posterior covariance matrix and ξ is an independent standard normal random vector. Each such parameter set can be converted back to

[4] Computing the inverse of C_{prior} and C_{x} is easy because both matrices are usually diagonal. When computing $\overline{q}_{\text{post}}$ by Eq. (12), the inverse of $C_{\text{prior}}^{-1} + Q^{*\text{T}}\, C_{\text{x}}^{-1}\, Q^{*}$ need not be computed explicitly; instead, we can employ a left matrix division for sparse matrices, which can be computed efficiently by Gaussian elimination.

[5] The differential Shannon entropy of an n-dimensional multivariate normal distribution with mean $\bar{x}$ and covariance matrix with determinant $|C|$ is given by $S = \frac{1}{2} \ln\left((2\pi\mathrm{e})^n\, |C|\right) = \frac{n}{2} \ln(2\pi\mathrm{e}) + \frac{1}{2} \sum_j \ln \lambda_j$, where the λ_j are the eigenvalues of C.

[6] By definition, a random variable X is log-normally distributed if its logarithm $\ln X$ follows a normal distribution.

non-logarithmic form by applying the exponential function (where appropriate). By creating different models with sampled parameters, we can explore the range of potential model dynamics in agreement with our present knowledge about the kinetic constants.

2 Data and prior distributions

2.1 How to choose the prior distributions

In any Bayesian approach, the choice of the prior distribution is of crucial importance. In the case of kinetic models, the priors for basis quantities (see Table 2) can be chosen according to the ranges of collected data values [10, 11]. But this is not the only possibility; in general, any multivariate Gaussian prior may be used. Since our normal distribution act as conjugate priors, one could also reuse a posterior, obtained from parameter balancing, as a prior for another round of parameter balancing with new data. As a practical application, one could balance the equilibrium constants and standard chemical potentials for a large metabolic network and then use the resulting posterior to define consistent correlated priors for kinetic models of different subsystems.

Name	Symbol	Available values	Median value	Std. dev. of log10	Mean value	Standard deviation	Unit
Standard chemical potentials	μ'°	45	-	-	-880	680	kJ/mol
Velocity constants	$\mathbf{k^V}$	-	10	1	141.7	2002	1/s
Michaelis constants	$\mathbf{k^M}$	62740	0.1	1	1.417	20.02	mM
Inhibitory constants	$\mathbf{k^I}$	12827	0.1	1	1.417	20.02	mM
Activation constants	$\mathbf{k^A}$	-	0.1	1	1.417	20.02	mM
Metabolite concentrations	$\mathbf{c}$	755	0.1	1.5	38.94	1516	mM
Enzyme concentrations	$\mathbf{E}$	912	0.0001	1.5	0.0389	15.16	mM
Equilibrium constants	$\mathbf{k^{eq}}$	2088	1	1.5	389.4	1.516e+05	-
Catalytic rate constants	$\mathbf{k^{cat}}$	12083	10	1.5	3894	1.516e+06	1/s
Maximal velocities	$\mathbf{v^{max}}$	-	0.001	2	40.29	1.623e+06	mmol/s
Reaction affinities	$\mathbf{A}$	-	-	-	0	10	kJ/mol
Chemical potentials	μ'	-	-	-	-880	680	kJ/mol

Table 2: Prior distributions (top) and pseudo values (bottom) used in parameter balancing. Distribution parameters were chosen in rough agreement with a data collection from Brenda [11], Sabio-RK [12], NIST [13], [14]), and Gibbs free energies published by Alberty [10] (see also [9]). For energy quantities, the arithmetic means and standard deviations were chosen directly. For all other quantities, the median values and the spread of their decadic logarithms were chosen and the arithmetic means and standard deviations for non-logarithmic values were computed from them.

A possible further improvement would be to use priors that are specific for certain enzyme classes or organisms, obtained from a previous analysis of variance or regression models [15, 16], or from physical predictions based on molecular structures [17].

2.2 Data and model parameters

If we are using heterogeneous data from the literature, it is important that the data values match the definition of our kinetic constants. For instance, the IC50 constants describe an inhibitor concentration that would lead to half-maximal inhibition. This exactly matches the definition of the k^{I} constants for complete inhibition in the modular rate laws. Other kinetic constants, like the Michaelis constants, refer to specific rate laws, which are also used to determine their values in enzyme assays. In parameter balancing, we use such Michaelis constants as proxies for the k^{M} values in the modular rate laws. The justification for this is that both quantities describe substrate concentrations leading to a half-maximal rate in the absence of product. Nevertheless, they appear in different rate laws and are therefore not completely equivalent.

If the input data table contains more than one value for the same quantity, we could represent these data values by several rows in the dependence matrix Q^*. However, to reduce the computational effort, we average over these values (for non-energy quantities, on log scale) using the formulae $\hat{\sigma}^2 = 1/(\sum_i 1/\sigma_i^2)$ and $\hat{x} = \hat{\sigma}^2 \sum_i x_i/\sigma_i^2$, which eventually leads to the same results. Of course, this averaging is not applied to pH- and temperature-dependent data, which we intend to keep as separate values.

3 Parameter balancing – additional features

Parameter balancing is a fairly flexible approach, which can be adapted to a variety of basis and derived quantities. The only restriction is that all dependencies between quantities (possibly, in their logarithmic form) need to be linear. Here we added some new aspects to the original parameter balancing method [7]. We extend the range of model quantities from kinetic constants to state-dependent quantities describing one or more metabolic states (chemical potentials, reaction affinities), account for measurement conditions affecting the model quantities (pH, temperature, electric potential), and propose data augmentation to implement further prior knowledge.

3.1 Reaction rates and and reaction affinities

By introducing metabolite and enzyme concentrations as independent basis quantities, we can capture state-dependent quantities such as the chemical potentials and reaction affinities. Unfortunately, the most important quantities – namely the reaction rates themselves – cannot be included into the scheme because the kinetic rate laws do not have the linear form required for Eq. (9). However, we can account for them indirectly *via* the forward and backward reaction rates (see section 4) or *via* the reaction affinities.

Depending on their reactant concentrations, chemical reactions can be in equilibrium (where the reaction rate is zero) or far from equilibrium (almost irreversible reactions). The distance from equilibrium can be expressed by the reaction affinity, which also determines the reaction direction [6]

$$v_l \neq 0 \Rightarrow \text{sign}(v_l) = \text{sign}(A_l), \tag{17}$$

the exchange fluxes [18], and the ratio of forward and backward rates [1]. The relation (17) between reaction affinities and reaction directions is exploited by existing methods for flux and metabolome analysis such as Network-Embedded Thermodynamic (NET) analysis [19] [20], Energy Balance Analysis [21], and other related flux-balance methods [22]. In parameter balancing, known reaction affinities can be used as input data for parameter estimation. Moreover, known flux directions imply inequality constraints for the reaction affinities, which could be used in the future to improve parameter balancing (see section 4).

3.2 Correction for incompatible pH values and temperatures

Kinetic constants can significantly depend on the measurement conditions, in particular on temperature and the pH value. If data values stem from different experimental conditions, this could lead to inconsistencies in parameter balancing. To avoid this problem, we account for the measurement conditions in the estimation procedure. We will illustrate this approach for the pH-dependence of standard chemical potentials and equilibrium constants, but it applies to other types of quantities as well.

A standard chemical potential μ'° is defined as the chemical potential of a substance in aqueous solution at a standard concentration of $c^\circ = 1$ mM and at given pressure and temperature. Dissolved molecules may assume different protonation states and in biochemical models, such states are usually summarized in a single *reactant* variable. It is convenient to describe this reactant by a *transformed* standard chemical potential μ'°, which effectively summarizes different protonation states and varies with the pH (see [2, 3, 4]). The dependency on other conditions (e.g., salt concentrations) can be treated in a similar manner. Since parameter balancing can only handle linear

dependencies, we approximate the impact of temperature and pH value on the transformed μ'° values by a linear expansion

$$\mu'^{\circ}(\mathrm{p_H}, \mathrm{T}) \quad \approx \quad \mu'^{\circ}({\mathrm{p_H}}^{\mathrm{ref}}, \mathrm{T}^{\mathrm{ref}}) + \frac{\partial \mu'^{\circ}}{\partial \mathrm{p_H}} \Delta \mathrm{p_H} + \frac{\partial \mu'^{\circ}}{\partial \mathrm{T}} \Delta \mathrm{T}. \tag{18}$$

$\mathrm{T}^{\mathrm{ref}}$ and ${\mathrm{p_H}}^{\mathrm{ref}}$ are reference values for temperature and pH value, and the differences between actual values and reference values are denoted by "Δ". In practice, T and $\mathrm{p_H}$ represent the actual measurement conditions for a certain data value μ'°, and $\mathrm{T}^{\mathrm{ref}}$ and ${\mathrm{p_H}}^{\mathrm{ref}}$ are the *in-vivo* conditions described in the model. To include Eq. (18) in our dependence matrix, we introduce the derivatives $\mu'^{\circ}_{\mathrm{pH}} = \frac{\partial \mu'^{\circ}}{\partial p_H}$ and $\mu'^{\circ}_{\mathrm{T}} = \frac{\partial \mu'^{\circ}}{\partial T}$ (i.e., the negative molar entropies) as new basis quantities and insert the terms $\Delta \mathrm{p_H}$ and $\Delta \mathrm{T}$ into the respective columns of Q^*. The transformed equilibrium constants, which also depend on the pH value, are treated accordingly. From Eqs. (18) and (3), we obtain the analogous formula

$$\ln k^{\mathrm{eq}}(\mathrm{p_H}, \mathrm{T}) \quad \approx \quad -\frac{1}{RT} N^{\mathrm{T}} \mu'^{\circ}({\mathrm{p_H}}^{\mathrm{ref}}, \mathrm{T}^{\mathrm{ref}}) - \frac{1}{\mathrm{RT}} \mathrm{N}^{\mathrm{T}} \frac{\partial \mu'^{\circ}}{\partial \mathrm{p_H}} \Delta \mathrm{p_H} - \frac{1}{\mathrm{RT}} \mathrm{N}^{\mathrm{T}} \frac{\partial \mu'^{\circ}}{\partial \mathrm{T}} \Delta \mathrm{T}. \tag{19}$$

The resulting dependence matrix Q^* has the form (only relevant rows and columns shown)

$$\begin{array}{c} \\ \mu'^{\circ} \\ \ln k^{\mathrm{eq}} \end{array} \begin{array}{c} \begin{array}{cc} \mu'^{\circ} & \qquad \mu'^{\circ}_{\mathrm{T}} \qquad\qquad \mu'^{\circ}_{\mathrm{pH}} \end{array} \\ \left(\begin{array}{c|cc} \mathbf{I} & \Delta \mathrm{T} & \Delta \mathrm{p_H} \\ \hline -N^{\mathrm{T}}/RT & -N^{\mathrm{T}} \cdot \Delta \mathrm{T}/\mathrm{RT} & -N^{\mathrm{T}} \cdot \Delta \mathrm{p_H}/\mathrm{RT} \end{array} \right). \end{array} \tag{20}$$

Note that $\Delta \mathrm{T_j}$ and $\Delta \mathrm{p_{H_j}}$ can have different values in each row, depending on the measurement conditions of the respective data points x^*_j. In parameter balancing, the matrix Q^* is used for *estimating* the basis quantities q, which are supposed to describe *in-vivo* temperature and pH. The matrix Q, on the contrary, is used to recompute all model quantities x, so we set the deviations $\Delta \mathrm{p_H}$ and $\Delta \mathrm{T}$ to zero or just omit the quantities $\mu'^{\circ}_{\mathrm{T}}$ and $\mu'^{\circ}_{\mathrm{pH}}$ and the corresponding matrix columns. Alternatively, we can also employ the $\mu'^{\circ}_{\mathrm{pH}}$ column to describe the pH differences between cell compartments. This becomes important if the model contains the same species in compartments with different pH values.

It is clear that the linear expansion Eq. (18) is only a rough approximation, chosen not on thermodynamic grounds, but for its simplicity. For a better approximation, we could expand the true pH dependency into a higher-order Taylor series and introduce terms like $\partial^2 \mu'^{\circ} / \partial p_H^2$ as additional basis quantities. This, however, will only be worth the effort if enough data are available. The proposed handling of of measurement conditions is not restricted to pH and temperature, but could be used for other measurement conditions like salt concentrations. As pointed out before, it is also applicable to other quantities, e.g. to pH-dependent k^{M} values.

3.3 Data augmentation: pseudo values for derived quantities

In parameter balancing, we employ priors to keep the balanced quantities within a plausible order of magnitude, even if few data are available. However, since priors can only be formulated for the basis quantities, the derived quantities could escape this control and assume unreasonable values. We address this problem by data augmentation and exemplify this here for equilibrium constants.

If no data value is available for an equilibrium constant, its balanced value will follow mostly from the priors and from the data values of the standard chemical potentials. Since the prior distributions are broad and small errors in the μ'° values can have a large effect on an equilibrium constant, the balanced k^{eq} value may be unreasonably high or low and show large uncertainty ranges. If the equilibrium constants themselves were basis quantities, we could delimit them by their prior distribution. However, this is not directly possible. We might control them, in fact, *via* the priors of the μ'° values, but this would require complicated, correlated priors. Therefore, we use data augmentation [23] as a more simple alternative.

Data augmentation exploits the fact that a conjugate prior is proportional to the likelihood contribution from a set of prior data. Accordingly, prior terms can be replaced by fictitious "pseudo" data values. In our case, we augment our data table with pseudo equilibrium constants, representing the distribution of typical equilibrium constants in our data collection. The same procedure can also be used for all other derived quantities, in parallel to the use of prior distributions for the basis quantities.

In the case of equilibrium constants, the pseudo value and its standard error are chosen from a symmetry consideration. If we exchange the substrates and products of a reaction – which, of course, is just a matter of definition – the equilibrium constant is replaced by its inverse, and its logarithmic value switches its sign. Therefore, the distribution of the logarithmic pseudo equilibrium constant should be centered around zero ($\overline{\ln k^{\mathrm{eq}}} = 0$), while its width $\sigma_{\ln k^{\mathrm{eq}}}$ represents the range of logarithmic equilibrium constants in our data collection. For simplicity, we assume a standard deviation of 1.5 for the decadic logarithms[7]. The non-logarithmic pseudo values represent a log-normal distribution with a median of 1 but, due to its large width, the mean value $\langle k^{\mathrm{eq}} \rangle$ is much larger.

4 Future prospects

Finally, we list a couple of possible extensions that have not been implemented yet in the parameter balancing workflow.

4.1 Extending the dependence scheme to other biochemical quantities

To support a broader range of quantities in the data and a more detailed description of biochemistry, the dependence matrix Eq. (10) can be extended to additional basis and derived quantities.

1. **Metabolite amounts.** To describe metabolite amounts (instead of concentrations) as derived quantities, one could introduce new rows containing a localization matrix with elements $\mathcal{L}_{ij} = 1$ where $\mathcal{L}_{ij} = 1$ if compound i is localized in compartment j and $\mathcal{L}_{ij} = 0$ otherwise. The compartment volumes could further be split into $\ln \Omega_{j(l)} = \ln \Omega_{\mathrm{cell}} + \ln \Delta\Omega_{j(l)}$ where $\Delta\Omega_{j(l)}$ is the volume of a cell compartment divided by the cell's total volume. This would allow to handle parameter uncertainties and correlations caused by an uncertain total cell size.

2. **Enzyme amounts and reaction rates measured as amounts per time.** Above, reaction velocities (and accordingly, maximal velocities) were supposed to have the physical dimension of concentration per time (mM/s). To obtain maximal velocities given as amounts/time (in mmol/s) instead, we need to add the term $\ln \Omega_{j(l)}$ to Eq. (5), where $\Omega_{j(l)}$ is the volume of the cell compartment (with index $j(l)$) in which the enzyme of reaction l is residing. To realize this in the dependence scheme, one could introduce the compartment volumes as new basis quantities. In the rows for maximal velocities (in mmol/s), the dependence matrix would be augmented by an enzyme localization matrix with elements $\mathcal{L}^{\mathrm{enz}}_{lj} = 1$ if enzyme l is localized in compartment j and $\mathcal{L}^{\mathrm{enz}}_{lj} = 0$ otherwise.

3. **Electrochemical potential** To account for the electrochemical potential of charged molecules, one could introduce the electric potentials of all model compartments (in their scaled form $F\,\Phi$, in kJ/mol) as basis quantities and augment the dependence matrix Q by a block matrix $\mathrm{Dg(z)}\,\mathcal{L}$ with the charge numbers z_i and the localization matrix $\mathcal{L}$.

4. **Arrhenius equation for rate constants** The temperature dependence of kinetic rate constants can be approximated by the Arrhenius equation $k(T) = k_0\, \mathrm{e}^{-E_{\mathrm{act}}/RT}$ with a hypothetical molar activation energy E_{act}. If we assume such a relationship for the velocity

[7] Since the unit of the equilibrium constant depends on the reaction stoichiometry (precisely, on the difference Δn between the numbers of substrate and product molecules), one should actually choose different widths for different groups of reactions depending on the value of Δn. For simplicity, we just use the same value of $\sigma_{\ln k^{\mathrm{eq}}}$ for all reactions.

constants k^{V}, we obtain an additional basis parameter E_{act} for each enzymatic reaction and the dependence matrix rows describing velocity constants, catalytic constants, and maximal velocities will contain elements $-1/RT$.

5. **Species types** SBML models can contain different `<Species>` elements referring to the same chemical entity (possibly defined by a `<SpeciesType>` element in SBML level 2), but localized in different compartments. For quantities which do not depend on a molecule's localization, only a single basis quantity value needs to be listed[8], but several copies of it will appear in the derived quantities. This also applies to the transformed standard chemical potentials because their dependence on pH and electrical potential is already taken into account in the dependence scheme.

6. **Independent equilibrium constants** Instead of the standard chemical potentials, we may introduce a subset of independent equilibrium constants as new basis quantities. All remaining equilibrium constants can be derived from them *via* Wegscheider conditions as described in [24].

7. $\mathbf{k}^{\mathrm{cat}}$ **over** $\mathbf{k}^{\mathrm{M}}$ The ratios $k^{\mathrm{cat}\pm}/k^{\mathrm{M}}$ for enzyme-reactant pairs could be introduced as additional derived quantities. By providing data or pseudo values for these ratios, one could delimit the parameters in a more plausible way than by directly using the $k^{\mathrm{cat}\pm}$ values.

8. **Combined kinetic constants** In some of the modular rate laws, several kinetic constants can be combined into a single parameter (see Table A.5 in the supplement of [1]). For instance, to translate the power-law modular rate law into a simple mass-action rate law, the enzyme concentration, catalytic constants, and Michaelis constants can be replaced by parameters $k_l'^{\mathrm{m}\pm} = u_l k^{\mathrm{cat}\pm} / \prod_i (k_{li}^{\mathrm{M}})^{h_l n_{il}^{\pm}}$. Even if such quantities are not available as data, they could be included into the dependence matrix Q for forward prediction.

9. **Several metabolic states** We can also consider more than one metabolic state in the dependence scheme; in this case, rows and columns of all state-dependent quantities will be duplicated.

10. **Forward and backward rates** For reversible reactions, the reaction rate can be split into forward and backward rates $v = v^+ - v^-$. The modular rate laws have the form $v_l = \frac{f_l}{D_l^*}[\tilde{v}^+ - \tilde{v}^-]$ where f_l describes allosteric regulation, D_l^* describes enzyme saturation depending on the rate law, and the non-regulated, non-saturable rates $\tilde{v}_l^{\pm}$ read

$$\tilde{v}_l^{\pm} \quad = \quad u_l \, k_{\pm l}^{\mathrm{cat}} \prod_i \left(\frac{c_i}{k_{c_i}^{\mathrm{M}}} \right)^{h_l \, n_{il}^{\pm}} . \tag{21}$$

The symbols n_{il}^+ and n_{il}^- denote the (positive) stoichiometric coefficients for reaction substrates (+) and products (-). The logarithmic values follow the linear equation

$$\ln \tilde{v}_l^{\pm} \quad = \quad \ln u_l + \ln k_{\pm l}^{\mathrm{cat}} + \sum_i h_l \, n_{il}^{\pm} \left(\ln c_i - \ln k_{c_i}^{\mathrm{M}} \right) \tag{22}$$

and can easily be included into the dependence scheme. Even if the rates $\tilde{v}_l^{\pm}$ are not directly available as data, we could guess their values from given flux data v_l and use them as input data for parameter balancing. Afterward, the initial guess may be iteratively improved based on the balancing results (f_l and D_l^* are computed from c, k^{M}, k^{A}, and k^{I}; the individual fluxes $v^{\pm}$ follow from v and A).

[8] The mapping between species and species types has to be implemented in the dependence matrix Q

4.2 Reducing the numerical effort

In the article, we demonstrate the method with a single reaction, but it works equally well for complex networks. The number of parameters is determined by the number of metabolites (e.g., standard chemical potentials, concentrations), the number of reactions (e.g., catalytic constants, enzyme levels), reaction/reactant edges (Michaelis constants), or allosteric interactions (activation and inhibition constants). Since reactions have a typical, limited number of reactants and regulators, the number of model parameters is approximately proportional to the number of reactions in the network. The computational effort, which involves sparse matrix inversions, grows more than quadratically, so computer speed will become limiting for large models. A possible way to reduce the effort is to subdivide a model into smaller parts (maybe, even individual reactions); however, to ensure compatible balancing results between submodels, all dependencies between the submodels will have to be eliminated beforehand. One possibility is to exactly predefine all equilibrium constants. In this case, parameters could be balanced separately for each reaction, and consistence could also be reached by safe parametrizations like $v(a,b) \sim k^{\mathrm{cat}} a - k^{\mathrm{cat}}/k^{\mathrm{eq}} \cdot b$ for mass-action rate laws. If the dependence scheme contains metabolite concentrations, their values need to be fixed at the submodel boundaries as well. In a more advanced approach, one could first balance the equilibrium constants and the standard chemical potentials (but no other quantities) for the entire model. The posterior obtained from this first run (describing only the standard chemical potentials) could later be used to construct correlated priors for the individual smaller submodels.

4.3 Inequality constraints

Another possible extension of parameter balancing would be to impose upper and lower bounds on individual model parameters or on their ratios or products. After conversion to natural scaling, this would yield linear inequality constraints, defining a feasible region in parameter space. In parameter balancing, a point estimate of the model parameters could be obtained by maximizing the logarithmic posterior, which is a quadratic function. With the new inequality constraints, the posterior will be restricted to the feasible region, which reduces uncertainties and leads to more realistic balancing results, but also increases the computational effort. To compute the posterior mode (i.e., the maximum point), we need to run quadratic programming with linear constraints. On the same occasion, we could impose equality constraints for prescribing model quantities with zero uncertainty, e.g., some predefined equilibrium constants. When sampling from the posterior distribution, we now need to reject all unfeasible points (for a more efficient sampling algorithm, see [25]). In both cases, the calculations are much more expensive than without the constraints. In addition, it may also happen that the constraints do not have a solution. An important application of such inequality constraints would be to restrict the signs of reaction affinities and thereby to enforce parameter sets that match certain known flux directions.

References

[1] W. Liebermeister, J. Uhlendorf, and E. Klipp. Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. *Bioinformatics*, 26(12):1528, 2010.

[2] R.A. Alberty. Degrees of freedom in biochemical reaction systems at specified pH and pMg. *Journal of Physical Chemistry*, 96:9614–9621, 1992.

[3] R.A. Alberty. Equilibrium calculations on systems of biochemical reactions at specified pH and pMg. *Biophysical Chemistry*, 42(2):117–131, 1992.

[4] R.A. Alberty. Biochemical thermodynamics (a review). *Biochemica et Biophysica Acta*, 1207:1–11, 1994.

[5] R. Wegscheider. Über simultane Gleichgewichte und die Beziehungen zwischen Thermodynamik und Reactionskinetik homogener Systeme. *Zeitschrift für Physikalische Chemie*, 39:257–303, 1902.

[6] T. de Donder and P. Van Rysselberghe. *Thermodynamic theory of affinity*. Stanford university press, 1936.

[7] W. Liebermeister and E. Klipp. Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theoretical Biology and Medical Modelling*, 3(1):42, 2006.

[8] A. Gelman, J. B. Carlin, H. S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, 1997.

[9] S. Borger, W. Liebermeister, J. Uhlendorf, and E. Klipp. Automatically generated model of a metabolic network. *Genome Informatics Series*, 18(1):215–224, 2007.

[10] R.A. Alberty. Calculation of Standard Transformed Gibbs Energies and Standard Transformed Enthalpies of Biochemical Reactants. *Archives of Biochemistry and Biophysics*, 353(1):116–130, 1998.

[11] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32, Database Issue:D431, 2004.

[12] U. Wittig, M. Golebiewski, R. Kania, O. Krebs, S. Mir, A. Weidemann, S. Anstein, J. Saric, and I. Rojas. *Data integration in the life sciences*, chapter SABIO-RK: integration and curation of reaction kinetics data. Springer, 2006.

[13] R. N. Goldberg. Thermodynamics of enzyme-catalyzed reactions: Part 6 - 1999 update. *Journal of Physical and Chemical Reference Data*, 28:931, 1999.

[14] R. N. Goldberg, Y.B. Tewari, and T.N. Bhat. Thermodynamics of enzyme-catalyzed reactions - a database for quantitative biochemistry. *Bioinformatics*, 20(16):2874–2877, 2004.

[15] W. Liebermeister. Predicting physiological concentrations of metabolites from their molecular structure. *J Comp Biol*, 12(10):1307–1315, 2005.

[16] S. Borger, W. Liebermeister, and E. Klipp. Prediction of enzyme kinetic parameters based on statistical learning. *Genome Informatics Series*, 17(1), 2006.

[17] R.R. Gabdoulline, M. Stein, and R.C. Wade. qPIPSA: Relating enzymatic kinetic parameters and interaction fields. *BMC Bioinformatics*, 8:373, 2007.

[18] W. Wiechert. The thermodynamic meaning of metabolic exchange fluxes. *Biophysical Journal*, 93(6):2255–2264, 2007.

[19] A. Kümmel, S. Panke, and M. Heinemann. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Molecular Systems Biology*, page 2006.0034, 2006.

[20] N. Zamboni, A. Kümmel, and M. Heinemann. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics*, 9(1):199, 2008.

[21] D.A. Beard, S. Liang, and H. Qian. Energy balance for analysis of complex metabolic networks. *Biophysical Journal*, 83(1):79–86, 2002.

[22] A. Hoppe, S. Hoffmann, and H.G. Holzhütter. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Systems Biology*, 1(1):23, 2007.

[23] D.A. van Dyk and X. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

[24] W. Liebermeister and E. Klipp. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theoretical Biology and Medical Modelling*, 3:41, 2006.

[25] N.D. Price, J. Schellenberger, and B.Ø. Palsson. Uniform sampling of steady-state flux spaces: Means to design experiments and to interpret enzymopathies. *Biophysical Journal*, 87:2172–2186, 2004.