Identification of metabolic network models from incomplete high-throughput datasets: Supplementary information

Sara Berthoumieux [1,*], Matteo Brilli [2,1], Hidde de Jong [1],
Daniel Kahn [2] and Eugenio Cinquemani [1*]

[1]INRIA Grenoble - Rhône-Alpes, Montbonnot, France
[2]Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR 5558,
Université Lyon 1, INRA, Villeurbanne, France

S1 Identifiability analysis

First of all, we note that in model (7) of the main text, for each reaction i, only a subset of metabolites is involved. That is, only a subset of the entries of vector $c_{\cdot i}$ need to be identified, while the values of the remaining entries are fixed to 0. Let us call n_{c_i} with $0 < n_{c_i} < (n+p)$ the effective number of parameters to identify. A straightforward reformulation of the regression model is the following:

$$w_{\cdot i} = Y^{'} \cdot c' + \varepsilon_{\cdot i} \tag{S1}$$

with $c' \in \mathbb{R}^{n_{c_i}}$ a vector collecting the nonzero values of $c_{\cdot i}$ and $Y^{'} \in \mathbb{R}^{q \times n_{c_i}}$ a matrix composed of the corresponding columns of Y,*i.e.,* the metabolites involved in reaction i. Bearing this in mind, for simplicity, we will drop index i in the sequel writing n_c in place of n_{c_i} and sticking to the usual notation $w = Y \cdot c + \varepsilon$.

To deal with identifiability issues, we use Principal Component Analysis (PCA) on the model (S1) [Jolliffe, 1986, Nikerel et al., 2009]. To detect nonidentifiable parameters, we decompose the data matrix Y using Singular Value Decomposition (SVD):

$$Y = U \cdot diag(s_1, s_2, ..., s_{n_c}) \cdot V^T \tag{S2}$$

with $U \in \mathbb{R}^{q \times q}$ and $V \in \mathbb{R}^{n_c \times n_c}$ orthonormal matrices and $s_1 \geq \cdots \geq s_{n_c} \geq 0$ the singular values of Y. Note that the sum of squared singular values is equal to the square of the Frobenius norm of Y, $\|Y\|^2 = \sum_k \sum_j Y_{j,k}^2$, that is, in an equivalent statistical interpretation, to q times the sum of the variances of all metabolite concentrations over q experiments (recall that each column of Y has zero mean by definition).

In presence of dependencies among data, there exists an index r with $1 \leq r < n_c$ such that $s_{r+1} = \cdots = s_{n_c} = 0$. Then Y is of rank r. As a consequence, for any two vectors w

*to whom correspondence should be addressed

and c such that $w = Y \cdot c$, there exists an $(n_c - r)$-dimensional vector space $K_Y \subseteq \mathbb{R}^{n-r}$ (the kernel of Y) such that $w = Y \cdot (c + k_Y)$ also holds for any $k_Y \in K_Y$. For the purpose of identification, this implies that c cannot be uniquely reconstructed from the data. (The case where $(s_{r+1}, \cdots, s_{n_c})$ are only approximately 0 will be discussed later in this section.)

We rely on the observation that $K_Y = \text{range}(V_{r+1:n_c})$, the vector space generated by the last $n_c - r$ columns of V. In order to formulate a regression problem with a well-defined solution, we rewrite model (S1) in terms of a reduced data matrix $Y^* \in \mathbb{R}^{q \times r}$ and a reduced parameter vector $c^* \in \mathbb{R}^r$ as follows:

$$\left\{ \begin{array}{l} w = Y^* \cdot c^* + \varepsilon \\ Y^* = Y \cdot V_{1:r} \end{array} \right. \tag{S3}$$

where $V_{1:r} \in \mathbb{R}^{n_c \times r}$ is the matrix obtained by extracting the first r columns of V. Since Y^* is full-column rank, $Y^* \cdot c^*$ is a linear combination in c^* of independent data vectors. This ensures that the solution to the regression (S3) is unique, hence we call c^* 'identifiable'.

Given a unique solution c^* to the regression (S3), the space of undistinguishable solutions for the original parameter vector c in (S1) can then be defined as follows:

$$c \in \{V_{1:r} \cdot c^* + k_Y \ , \ k_Y \in K_Y\} \tag{S4}$$

Depending on the structure of the orthonormal matrix V, we may be able to isolate some entries of c that can be uniquely determined from the reduced model, that is, from the estimates of c^*. This happens when all elements of at least one row of $V_{r+1:n_c}$ are equal to 0. Indeed, this is the criterion we used in Sec. 5 to isolate identifiable parameters in nonidentifiable reactions, such as reactions 10 and 11 in Table 1.

In practice, singular values are rarely exactly 0, even in presence of data dependencies. This can be due to several causes, including measurement noise, numerical roundoffs, etc. Still, for some $r < n_c$, values of $(s_{r+1}, \cdots, s_{n_c})$ sufficiently close to 0 can make the estimates of c solving regression poorly determined. Thus a criterion to discard those singular values needs to be defined.

In this paper, we approximate by zero all singular values whose total contribution to the variance of Y, *i.e.,* to the 'informativity' of the metabolite data, is under some suitable threshold λ, that is, we define

$$r = \min \left\{ t \in [1..(n_c - 1)] \ , \ \sqrt{\frac{\sum_{k=1}^{t} s_k^2}{\sum_{k=1}^{n_c} s_k^2}} \geq \lambda \right\} \tag{S5}$$

and set $s_{r+1} = \cdots = s_{n_c} = 0$. By this approximation, from (S2) we obtain a data matrix Y of rank $r < n_c$. The PCA method described before then applies. The results discussed in Sec. 4 and Sec. 5 were obtained with $\lambda = 0.99$.

S2 Likelihood-based identification of linlog models

We rely on the notation of Sec. 3 of the main section, *i.e.,* we focus on a single reaction and drop index i from the notation. The loglikelihood of the model is:

$$\log \mathscr{L}(c) = \log \int f_{W|\check{y},\tilde{y},c}(w) f_{\widetilde{Y}|\check{y},c}(\tilde{y}) d\tilde{y}. \tag{S6}$$

For convenience, we rewrite (S6) in terms of the random variable $Z = \widetilde{Y} \cdot c$ introduced in Sec. 3 so that it becomes:

$$\log \mathscr{L}(c) = \log \int f_{W|\breve{y},z,c}(w) f_{Z|\breve{y},c}(z) dz. \tag{S7}$$

Here $f_{W|\breve{y},z,c}(\cdot)$ is the Gaussian likelihood function of model (7), equivalently rewritten as $W = \breve{Y} \cdot c + Z + \varepsilon$, given $\breve{Y} = \breve{y}$ and $Z = z$, with z varying over all possible values of Z, and $f_{Z|\breve{y},c}$ is the Gaussian prior of $Z = \widetilde{Y} \cdot c$ following from (10). The expressions of $f_{W|\breve{y},z,c}$ and $f_{Z|\breve{y},c}$ are thus

$$\begin{cases} f_{W|\breve{y},z,c}(w) = \frac{1}{\sqrt{\det(2\pi\Sigma_\varepsilon)}} \exp(-\frac{1}{2}[w - \breve{Y} \cdot c - z]^T \Sigma_\varepsilon^{-1} [w - \breve{Y} \cdot c - z]), \\ f_{Z|\breve{y},c}(z) = \frac{1}{\sqrt{\det(2\pi\Sigma_{\breve{y},c})}} \exp(-\frac{1}{2}[z - \mu_{\breve{y},c}]^T \Sigma_{\breve{y},c}^{-1} [z - \mu_{\breve{y},c}]), \end{cases} \tag{S8}$$

with $\mu_{\breve{y},c} = M \cdot c$, where the entry $M_{j,k}$ of matrix M is the mean $\mu_{j,k}$ of the distribution of $\widetilde{Y}_{j,k}$, and $\Sigma_{\breve{y},c}$ is the variance matrix of the random variable Z. By the independence assumptions on $\widetilde{Y}$, it turns out that

$$\Sigma_{\breve{y},c} = diag\left(\sum_{k=1}^{n_c} c_k^2 \cdot [\sigma_{1,k}^2 \cdots \sigma_{q,k}^2]^T \right) \tag{S9}$$

where $\sigma_{j,k}$ is defined in (10).

Assume for the moment that $\Sigma_{\breve{y},c}$ is invertible. Defining

$$f_{p_c}(z) = f_{W|\breve{y},z,c}(w) \cdot f_{Z|\breve{y},c}(z), \tag{S10}$$

after simple but tedious calculations, we obtain

$$f_{p_c}(z) = \kappa_{f_c} \cdot f_c(z) \tag{S11}$$

with f_c the density function of a Gaussian distribution $\mathscr{N}(\mu_{f_c}, \Sigma_{f_c})$ and

$$\begin{cases} \Sigma_{f_c} = [\Sigma_\varepsilon^{-1} + \Sigma_{\breve{y},c}^{-1}]^{-1}, \\ \mu_{f_c} = \Sigma_{f_c} \cdot [\Sigma_\varepsilon^{-1} \cdot (w - \breve{Y} \cdot c) + \Sigma_{\breve{y},c}^{-1} \cdot \mu_{\breve{y},c}], \\ \kappa_{f_c} = \frac{\exp(-\frac{1}{2}[w - \breve{Y} \cdot c - \mu_{\breve{y},c}]^T \cdot [\Sigma_\varepsilon + \Sigma_{\breve{y},c}]^{-1} \cdot [w - \breve{Y} \cdot c - \mu_{\breve{y},c}])}{\sqrt{\det(2\pi[\Sigma_\varepsilon + \Sigma_{\breve{y},c}])}}. \end{cases} \tag{S12}$$

The proportionality factor κ_{f_c} does not depend on the integration variable z, so it can be taken out of the integral and (S7) can be rewritten as follows:

$$\log \mathscr{L}(c) = \log(\kappa_{f_c}) + \log\left(\int f_c(z) dz \right). \tag{S13}$$

The integral of a normalized Gaussian density function being 1, we finally have an analytical expression for the loglikelihood: $\log \mathscr{L}(c) = \log(\kappa_{f_c})$.

The above results are used in the expectation step of the EM algorithm. Recall the definition

$$Q(c|\hat{c}^{\ell-1}) = \int \log(f_{Z,W|\breve{y},c}(z,w)) f_{Z|\breve{y},\hat{c}^{\ell-1},w}(z) dz. \tag{S14}$$

The Bayes theorem allows us to rewrite (S14) as follows:

$$Q(c|\hat{c}^{\ell-1}) = \int \log(f_{W|\breve{y},z,c}(w) f_{Z|\breve{y},c}(z)) \frac{f_{W|\breve{y},z,\hat{c}^{\ell-1}}(w) f_{Z|\breve{y},\hat{c}^{\ell-1}}(z)}{f_{W|\breve{y},\hat{c}^{\ell-1}}(w)} dz. \tag{S15}$$

Function $f_{W|\breve{y},\hat{c}^{\ell-1}}(w)$ does not depend on z so it can be taken out of the integral. Moreover, this function does not depend on c so it will have no impact on the maximization step of EM. Thus, we can ignore this function from the computation of the expectation function above.

Using definitions (S10) and (S11), we can rewrite (S15) in the following way:

$$\begin{aligned} Q(c|\hat{c}^{\ell-1}) &\propto \int \kappa_{f_{\hat{c}^{\ell-1}}} f_{\hat{c}^{\ell-1}}(z) \log(\kappa_{f_c} f_c(z)) dz \\ &\propto \int f_{\hat{c}^{\ell-1}}(z) \log(\kappa_{f_c} f_c(z)) dz. \end{aligned} \tag{S16}$$

We have dropped the constant factor $\kappa_{f_{\hat{c}^{\ell-1}}}$ as it does not depend on c and thus does not influence the maximization step of EM. By replacing $\log(\kappa_{f_c} f_c(z))$ by $\log(\kappa_{f_c} f_c(z) f_{\hat{c}^{\ell-1}}(z)/ f_{\hat{c}^{\ell-1}}(z))$ and separating the integrand in a sum of terms, we can rewrite (S16) as

$$-Q(c|\hat{c}^{\ell-1}) \propto \int f_{\hat{c}^{\ell-1}}(z) \log\left(\frac{f_{\hat{c}^{\ell-1}}(z)}{f_c(z)}\right) dz - \int f_{\hat{c}^{\ell-1}}(z) \log(f_{\hat{c}^{\ell-1}}(z)) dz - \log(\kappa_{f_c}). \tag{S17}$$

We recognize in the first term the definition of the Kullback-Leibler divergence $KL(f_c||$ $f_{\hat{c}^{\ell-1}})$ between the two probability distributions f_c and $f_{\hat{c}^{\ell-1}}$ and in the second term the entropy $H(f_{\hat{c}^{\ell-1}})$ of $f_{\hat{c}^{\ell-1}}$ [Cover and Thomas, 2006, Stoorvogel and van Schuppen, 1996]. For Gaussian distributions, these can be written explicitely as

$$\begin{aligned} KL(f_c||f_{\hat{c}^{\ell-1}}) = \frac{1}{2}(\log\left(\frac{\det(\Sigma_{f_c})}{\det(\Sigma_{f_{\hat{c}^{\ell-1}}})}\right) + Tr(\Sigma_{f_c}^{-1}\Sigma_{f_{\hat{c}^{\ell-1}}}) \\ + [\mu_{f_c} - \mu_{f_{\hat{c}^{\ell-1}}}]^T \Sigma_{f_c}^{-1} [\mu_{f_c} - \mu_{f_{\hat{c}^{\ell-1}}}]), \end{aligned} \tag{S18}$$

where $Tr(\ldots)$ stands for trace and

$$H(f_{\hat{c}^{\ell-1}}) = \log\left(\sqrt{\det(2\pi e \Sigma_{f_{\hat{c}^{\ell-1}}})}\right). \tag{S19}$$

To summarize, together with (S12), this gives us the explicit formula

$$Q(c|\hat{c}^{\ell-1}) \propto -KL(f_c||f_{\hat{c}^{\ell-1}}) - H(f_{\hat{c}^{\ell-1}}) + \log(\kappa_{f_c}), \tag{S20}$$

which we employ in our implementation of EM.

In more generality, for some values of c, $\Sigma_{\breve{y},c}$ may be singular or poorly conditioned. To avoid this circumstance, we can adapt our procedure as follows. We consider a decomposition

$$W = \breve{Y} \cdot c + Z + (\varepsilon' + \varepsilon'') = \breve{Y} \cdot c + (Z + \varepsilon') + \varepsilon'' \tag{S21}$$

where ε' and ε'' are independent zero-mean Gaussian random vectors such that $\Sigma_{\varepsilon'} \triangleq Var(\varepsilon') = \alpha\Sigma_\varepsilon$ and $\Sigma_{\varepsilon''} \triangleq Var(\varepsilon'') = (1-\alpha)\Sigma_\varepsilon$, with $\alpha \in (0,1)$ a tunable parameter. Since $\Sigma_\varepsilon > 0$ by assumption, it follows that $\Sigma_{\varepsilon'} > 0$ and $\Sigma_{\varepsilon''} > 0$. Moreover, $\Sigma_\varepsilon = \Sigma_{\varepsilon'} + \Sigma_{\varepsilon''}$, *i.e.,* the statistics of ε and of $\varepsilon' + \varepsilon''$ are identical. Since $Var(Z + \varepsilon') = \Sigma_{\breve{y},c} + \Sigma_{\varepsilon'} > 0$, if we interpret $Z + \varepsilon'$ as the unknown observations (in place of Z) and ε'' as the model noise (in

place of ε), we ensure that the variance of the 'missing data' is invertible. Thus, in practice, we apply all formulas developed above with $\Sigma_{\breve{y},c} + \Sigma_{\varepsilon'}$ in place of $\Sigma_{\breve{y},c}$ and $\Sigma_{\varepsilon''}$ in place of Σ_{ε}.

The effect of the specific choice of α is under investigation. In this work, we took $\alpha = 0.2$, a value that leads to good results in practice.

S3 Validation on synthetic data

The model used for comparing peformance of the identification algorithms is a reduced synthetic linlog model of the *E. coli* central carbon metabolism network (Fig. S1 of the main text). This network contains 17 variables, describing internal and external metabolites, and 25 reactions, summarized in Table S1 and Table S2, respectively. The linlog model has the form of Eq. (1)-(2) of the main text.

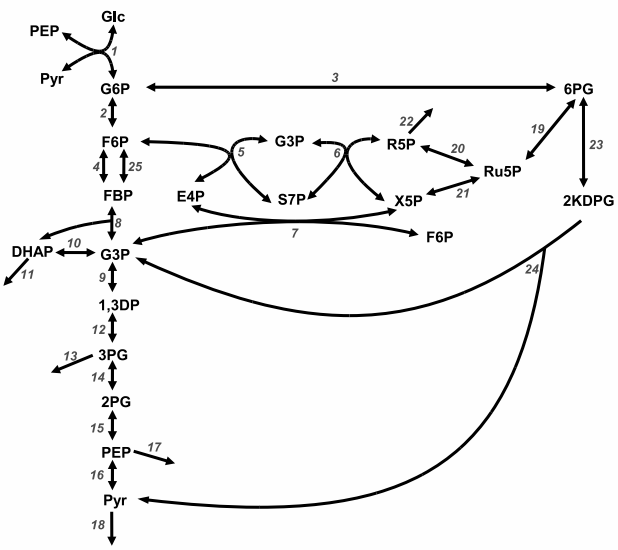

Figure S1: Network for the synthetic model, a reduced version of the *E. coli* central carbon metabolism network.

A dataset was generated from the synthetic linlog model by setting all enzyme concentrations to 1 and choosing plausible values for the parameter vector a and matrices B^x, B^u, that is, values consistent with existing kinetic models of carbon metabolism in *E. coli* [Bettenbrock et al., 2005]. Then $q = 30$ different experimental conditions were simulated by randomly changing enzyme concentrations. For each condition $j \in \{1, ..., q\}$, vectors $\ln(x^{(j)})$, $\ln(u^{(j)})$ and $v^{(j)}$ were determined by the equations resulting from the formulation of the linlog model and the (quasi-)steady-state equation $N \cdot v = 0$:

Index	Name	Symbol
1	Pyruvate	Pyr
2	Phosphoenol-pyruvate	PEP
3	Glyceraldehyde-3-phosphate	G3P
4	Fructose-6-phosphate	F6P
5	Glucose-6-phosphate	G6P
6	3-phosphoglycerate	3PG
7	Dihydroxyacetonephosphate	DHAP
8	Ribulose-5-phosphate	Ru5P
9	Ribose-5-phosphate	R5P
10	6-phosphogluconate	6PG
11	Erythrose-4-phosphate	E4P
12	Xylulose-5-phosphate	X5P
13	2-phosphoglycerate	2PG
14	1,3-diphosphosphoglycerate	1,3DP
15	Fructose-1,6-bisphosphate	FBP
16	2-keto-3-deoxy-6-phosphogluconate	2KDPG
17	Sedoheptulose-7-phosphate	S7P

Table S1: Metabolites included in the synthetic linlog model.

$$\begin{cases} \begin{bmatrix} \ln(x^{(j)}) \\ \ln(u^{(j)}) \end{bmatrix} = -\left[N \cdot \mathrm{diag}(e^{(j)}) \cdot [B^x\, B^u]\right]^{-1} N \cdot \mathrm{diag}(e^{(j)}) \cdot a, \\ v^{(j)} = \mathrm{diag}(e^{(j)}) \cdot \left(a + B^x \cdot \ln(x^{(j)}) + B^u \cdot \ln(u^{(j)})\right). \end{cases} \tag{S22}$$

For this dataset, four scenarios were considered, corresponding to more or less favorable conditions for identification: 40 % and 75 % missing entries and 10% and 20% noise. For each column of Y, *i.e.,* each metabolite of the model, the 40% or 75% missing data were distributed randomly over the q measurements. Randomly generated noise was added to the same incomplete dataset in each of 100 Monte-Carlo repetitions.

Identifiability analysis was performed following the approach described in Sec S1, with $\lambda = 0.99$. 10 reactions were found to be nonidentifiable (reactions 2, 5, 6, 7, 8, 12, 14, 15, 20 and 21). Among these reactions only 3 identifiable parameters could be isolated (one in reaction 2, one in reaction 7 and one in reaction 12).

Results from all identification methods on identifiable reactions are summarized in Fig. S2 for the most favorable scenario with 40% missing data and 10% noise, and in Fig. S3 for the least favorable scenario with 75% missing data and 20% error. The results for the other scenarios fall between those shown in Fig. S2 and Fig. S3, and are not shown here.

S4 Application to central metabolism in *E. coli*

Fig. 3 shows a (simplified) representation of central carbon metabolism in *E. coli*. The network could not be directly transformed into a linlog model of the form (1)-(2), since metabolites G3P, E4P, X5P, 2KDPG, OAA, IsoCit, SuccoA, Acp and Glyox were not measured by Ishii et al. [2007]. This prevents the estimation of elasticities for the above metabolites

Index	Name
1	Phosphotransferase system
2	Glucose-6-phosphate isomerase
3	Glucose-6-phosphate dehydrogenase
4	Phosphofructokinase
5	Transaldolase
6	Transketolase a
7	Transketolase b
8	Aldolase
9	Glyceraldehyde-3-phosphate dehydrogenase
10	Triosephosphate isomerase
11	Glycerol-3-phosphate dehydrogenase
12	Phosphoglycerate kinase
13	Serine synthesis
14	Phosphoglycerate mutase
15	Enolase
16	Pyruvate kinase
17	PEP carboxylase
18	Pyruvate synthesis
19	6-Phosphogluconate dehydrogenase
20	Ribose-phosphate isomerase
21	Ribulose-phosphate epimerase
22	Ribose-phosphate pyrophosphokinase
23	Phosphogluconate dehydratase
24	KDPG aldolase
25	Fructose bisphosphatase

Table S2: Reactions included in the synthetic linlog model.

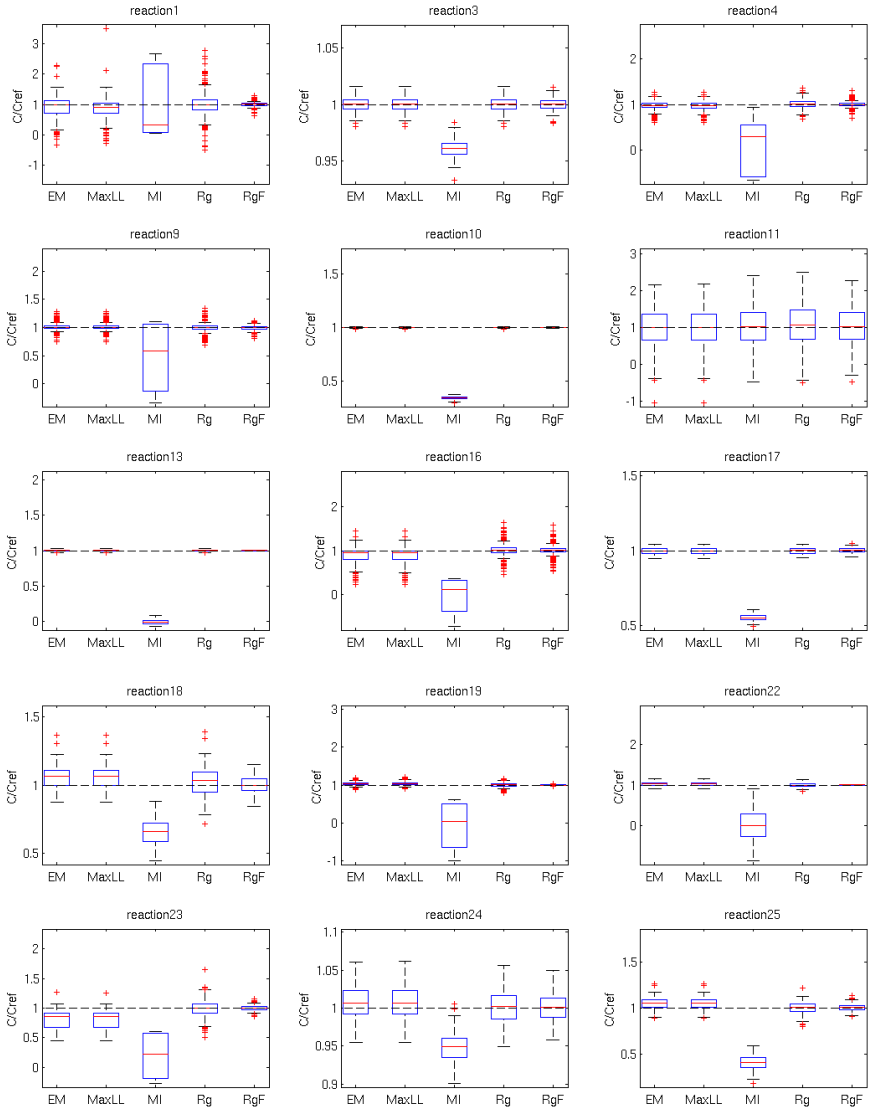

Figure S2: Statistics of estimated parameter values in identifiable reactions for datasets with 40% of missing data and 10% noise. The graphical notations are the same as for Fig. 1.

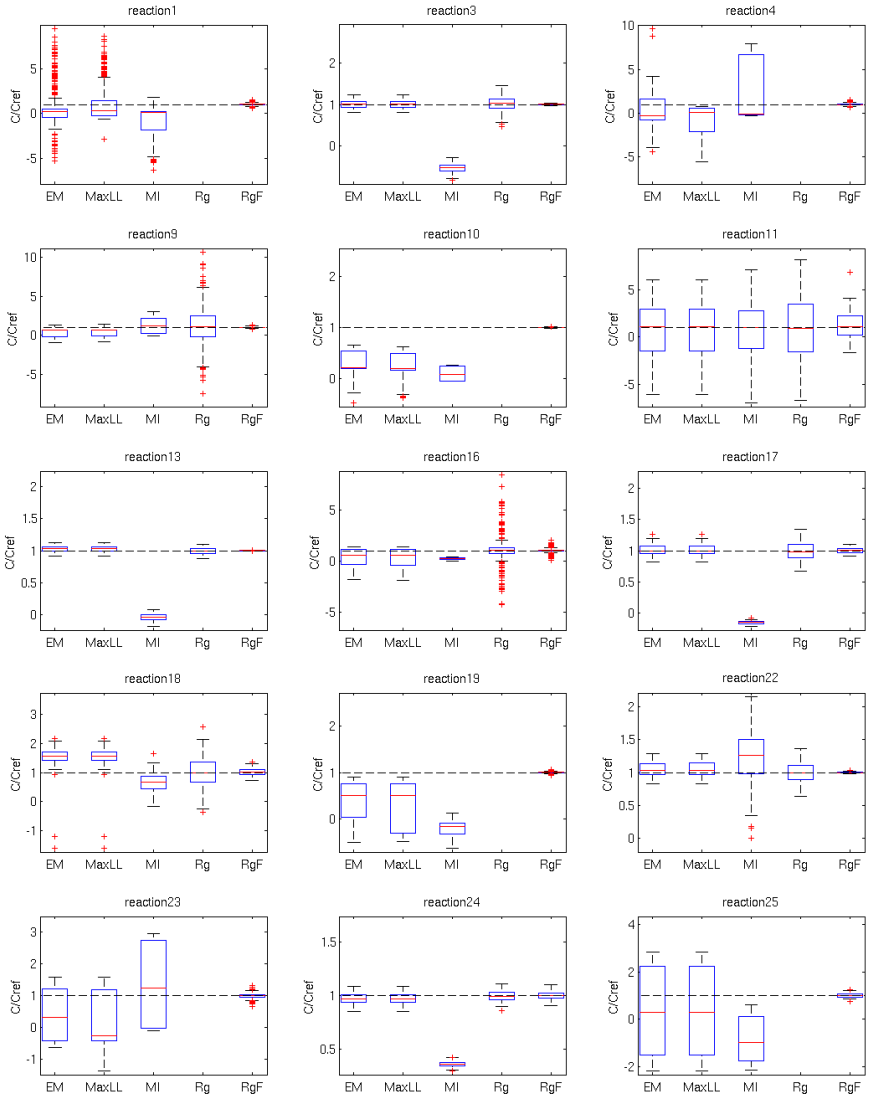

Figure S3: Statistics of estimated parameter values in identifiable reactions for datasets with 75% of missing data and 20% noise. The graphical notations are the same as for Fig. 1.

and their inclusion in the model. We overcome this limitation by lumping reactions not measured by Ishii et al. [2007].

In addition to the above model simplification imposed by the available dataset, we added a phenomenological reaction μ to model biomass production. The reaction involves 11 metabolites, the reaction flux is equal to the dilution rate under the experimental conditions of Ishii et al. [2007] and the enzyme concentration is set to 1.

The linlog model thus obtained contains 16 internal metabolites and 7 external metabolites or cofactors, listed in Table S3, as well as 31 reactions, listed in Table S4.

Internal metabolites			
Index	Symbol	Index	Symbol
1	PEP	9	Ru5P
2	G6P	10	R5P
3	Pyr	11	S7P
4	F6P	12	2KG
5	FBP	13	Suc
6	DHAP	14	Fum
7	3PG	15	Mal
8	6PG	16	Cit

Index	External metabolites or cofactors
17	Glc
18	AcoA/coA
19	ATP/ADP
20	NADPH/NADP
21	NADH/NAD
22	FAD
23	Ace

Table S3: Internal and external metabolites and cofactors of the linlog model of carbon metabolism in *E. coli*. Some of the cofactors are modeled as ratios of metabolite concentrations, *e.g.,* ATP/ADP.

References

K. Bettenbrock, S. Fischer, A. Kremling, K. Jahreis, T. Sauter, and E.D. Gilles. A quantitative approach to catabolite repression in *Escherichia coli*. *J. Biol. Chem.*, 281(5):2578–84, 2005.

T.M. Cover and J.A. Thomas. *Elements of Information Theory, 2nd edition*. Wiley, 2006.

N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P.Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, 316 (5824):593–7, 2007.

I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

I.E. Nikerel, W.A. van Winden, P.J.T. Verheijen, and J.J. Heijnen. Model reduction and *a priori* kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics. *Metab. Eng.*, 11(1):20–30, 2009.

A. Stoorvogel and J.H. van Schuppen. System identification with information theoretic criteria. In S. Bittanti and G. Picci, editors, *Identification, Adaptation, Learning*, volume 153 of *NATO ASI*, pages 289–338. Springer Verlag, 1996.

Index	Reaction
1	Glc + PEP $\xleftarrow{ptsG}$ Pyr+G6P
2	G6P $\xleftarrow{pgi}$ F6P
3	F6P + ATP/ADP $\xleftarrow{pfkA,pfkB}$ FBP $[PEP]_{in}$
4	FBP $\xleftarrow{fbaA,fbaB}$ DHAP
5	DHAP $\xleftarrow{tpiA}$ 3PG
6	FBP + ATP/ADP $\xleftarrow{gapA;pgk}$ 3PG + NADH/NAD
7	3PG $\xleftarrow{gpmA,gpmB;eno}$ PEP
8	PEP + ATP/ADP $\xleftarrow{pykA,pykF}$ Pyr $[FBP]_{act}$
9	Pyr $\xleftarrow{aceE:aceF:lpdA}$ AcoA/coA + NADH/NAD
10	G6P $\xleftarrow{zwf;pgl}$ 6PG + NADPH/NADP
11	6PG $\xleftarrow{gnd}$ Ru5P + NADPH/NADP
12	Ru5P $\xleftarrow{rpe}$ S7P
13	Ru5P $\xleftarrow{rpiA,rpiB}$ R5P $[G6P]_{in}$
14	R5P $\xleftarrow{tktA}$ S7P
15	S7P $\xleftarrow{talA,talB}$ F6P
16	Ru5P $\xleftarrow{tktB}$ F6P
17	AcoA/coA $\xleftarrow{gltA,prpC}$ Cit $[2KG]_{in}$ $[NADH/NAD]_{act}$
18	Cit $\xleftarrow{acnA,acnB}$ 2KG
19	AcoA/coA $\xleftarrow{icdA}$ 2KG + NADPH/NADP
20	2KG $\xleftarrow{sucA:sucB:lpdA;sucC:sucD}$ Suc + NADH/NAD
21	Suc + FAD $\xleftarrow{sdhA:sdhB:sdhC:sdhD}$ Fum
22	Fum $\xleftarrow{fumA,fumB,fumC}$ Mal
23	Mal + PEP $\xleftarrow{mdh}$ Cit +NADH/NAD
24	PEP $\xleftarrow{ppc;pckA}$ Mal + Cit + ATP/ADP $[FBP]_{act}$
25	Mal $\xleftarrow{maeB,sfcA}$ Pyr + NADPH/NADP $[AcoA/coA]_{in}$ $[NADH/NAD]_{act}$
26	AcoA/coA $\xleftarrow{aceA;aceB}$ Suc + Mal
27	PEP+G6P+Pyr+F6P+3PG+AcoA/coA+R5P+2KG+ATP/ADP $\xrightarrow{\mu}$NADPH/NADP+NADH/NAD
28	6PG $\xleftarrow{edd;eda}$ Pyr
29	AcoA/coA $\xleftarrow{pta;ackA,ackB}$ Ace+ATP/ADP $[Pyr]_{act}$ $[NADPH/NADP]_{in}$ $[NADH/NAD]_{in}$
30	Pyr + NADH/NAD $\xrightarrow{ldhA}$
31	AcoA/coA $\xrightarrow{adhE}$

Table S4: Reactions of the linlog model of carbon metabolism in *E. coli*. Activators and inhibitors of the reaction are shown with $[\cdot]_{act}$ and $[\cdot]_{in}$, respectively. Reaction 27, labeled μ, is a phenomenological reaction for biomass production. The enzyme names are separated by a comma in the case of isoenzymes, by a colon for enzyme complexes, and by a semi-colon when the enzymes catalyze reactions that have been lumped together in the model. Reactions 20, 26, 28 and 29 result from the merging of reactions due to the absence of measurements of SuccoA, Glyox, 2KDPG and Acp, respectively, in the dataset of Ishii et al. [2007].