**Supplementary Methods**

Table of Contents

[Improved kcat matching algorithm 1](#_Toc67068956)

[Estimation of phylogenetic distance between pairs of organisms 2](#_Toc67068957)

[Iterative curation of limiting kcat numbers based on parameter sensitivity coefficients 3](#_Toc67068958)

[Incorporation of proteomics constraints 3](#_Toc67068959)

[Comparative flux variability analysis 5](#_Toc67068960)

[Absolute protein quantification 6](#_Toc67068961)

[Total protein extraction 6](#_Toc67068962)

[Proteins digestion 7](#_Toc67068963)

[Mass spectrometry analysis 7](#_Toc67068964)

[Estimation of absolute protein abundance values 8](#_Toc67068965)

[References 8](#_Toc67068966)

# **Improved *k*cat matching algorithm**

The *k*cat matching algorithm in theGECKO toolbox queries kinetic parameters from BRENDA, the largest database available on enzymatic information1. However, one of the most important limitations to consider is that such parameters are only available for <10% of the known biochemical reactions2. The turnover number assignment to each of the enzymatic reactions present in a GEM is based on a flexible algorithm that allows the incorporation of kinetic parameters even when values for the specific organism and natural substrate of the enzyme are not available. However, as overestimation of microbial growth rates under environmental and genetic perturbations remains one of the main challenges for GEM development, biological relevance of the imposed kinetic constraints plays a crucial role for improving predictive accuracy3. In this regard, a global analysis for the reported *k*cat values on BRENDA (**Supp. file 1**) pointed out the following potential issues.

1. The availability of kinetic parameters is highly heterogeneous, i.e. not all organisms have been studied to the same extent.
2. *k*cat value distributions showed to be significantly different amongst kingdoms of life, therefore the catalytic activity of enzymes might be phylogenetically constrained.
3. *k*cat value distributions are highly dependent on the metabolic context. For all kingdoms of life, there are important differences on the distributions for enzymes belonging to different metabolic pathways groups, being central carbon and energy metabolism enzymes the fastest group (on average) when compared to those involved in amino acid, fatty acid and nucleotide metabolism and secondary and intermediate metabolism.

In order to address the aforementioned limitations, the GECKO *k*catmatching algorithm was modified aiming to provide a more accurate parameterization of models. A comparison between the introduced and previous hierarchical algorithms is shown in **Table x.**

**Table S2.1.- kcat matching algorithms comparison.**

|  |  |
| --- | --- |
| **Original *k*cat matching criteria** | **New criteria** |
| 1. As a first option, it will try to match the E.C. number, the organism and the corresponding substrate to some *k*cat annotation in the BRENDA database.
 | 1. Same as original.
 |
| 1. If no match is found, it will try to match the E.C. number and the substrate, but with any organism available.
 | 1. If no match is found, it will try to match the E.C. number and the substrate, but for the **phylogenetically closest** organism with available values.
 |
| 1. If no match is found, it will try to match the E.C. number and the organism, but with any substrate available.
 | 1. Same as original.
 |
| 1. If still no match is found, it will try to match the E.C. number for any organism, and any substrate available.
 | 1. If no match is found, it will try to match the E.C. number and the organism but looking in **specific activity** values instead of *k*cat (S.A.\*Mweight = *k*cat).
 |
| 1. If still no match is found, then it will introduce one wildcard to the E.C. number and attempt all previous 4 steps again.
 | 1. If still no match is found, it will try to match a *k*cat value for the E.C. number, any substrate but for the **phylogenetically closest** organism with available values.
 |
|  | 1. If still no match is found, it will try to match a **specific activity** value for the E.C. number, any substrate but for the **phylogenetically closest** organism available.
 |
|  | 1. Finally, if still no match is found, then it will introduce one wildcard (WC) to the E.C. number and attempt all previous 6 steps again.
 |

# **Estimation of phylogenetic distance between pairs of organisms**

The phylogenetic distance between organisms is measured as the number of nodes of separation between two organisms in the KEGG taxonomical tree (incorporated as a MATLAB workspace file into the toolbox), this new feature follows from the assumption that kinetic parameters on enzymes have been finely tuned by evolution and are phylogenetically related4. The incorporation of specific activity values increases the parameter coverage and avoids the assignment of a high number of wild cards, making the assignments as close as possible to the original metabolic function of the specific enzyme class.

# **Iterative curation of limiting *k*cat numbers based on parameter sensitivity coefficients**

Once kinetic parameters and protein pool bounds have been incorporated into the ecModel it is very likely that overconstraining arises due to the intrinsic uncertainty of the incorporated *k*cat values. For such cases, the module ***kcat\_sensitivity\_analysis*** flexibilizes the coefficients with a higher control on the simulated objective function value based on parameter sensitivity coefficients given by

$$C\_{ij}=\frac{k\_{cat}^{ij}}{v\_{obj}}\frac{∆v\_{obj}}{∆k\_{cat}^{ij}}$$

in which $k\_{cat}^{ij}$ represents the *k*catparameter of the enzyme *i* in reaction *j*; $v\_{obj}$ is the original value in the objective function; $∆k\_{cat}^{ij}$ is an induced perturbation in the *k*cat equivalent to 10-fold increase of its initial value; $∆v\_{obj}$ is the change in the objective function after perturbing $k\_{cat}^{ij}$. The magnitude of the induced perturbations was chosen based on the fact that *k*catvalues for a given enzyme class may span several orders of magnitude, as shown by main **Figure 1A.**

The ECCs are ranked in a decreasing way and the enzyme with the coefficient is then selected for a 10-fold *k*cat increase, based on the assumption that *k*catparameters may span orders of magnitude across organisms and substrates even for the same enzyme class. This procedure iterates until the ecModel is able to reach to provided experimental growth rate in the ***getModelParameters.m*** function. Information regarding the flexibilized *k*cat values, their respective proteins and reactions, ECCs, flexibilized and original *k*cat

s is saved as a text file in the ***GECKO/model*** folder of the toolbox under the name ***kcat\_modifications.txt***.

# **Incorporation of proteomics constraints**

The ***integrate\_proteomics*** module in GECKO enables the generation of condition dependent models with proteomics constraints for any given dataset of absolute protein abundances [mmol/ gDw] with **m** replicates for **n** conditions. The data incorporation algorithm consists of the following steps.

1.- For each experimental condition, the abundance values are filtered, excluding proteins that are not present in at least 2/3 of the total number of condition replicates and also noisy measurements (proteins with relative standard deviation higher than 1 across replicates). Median abundance values ($\left[E\_{i}\right]$) and their standard deviations across replicates ($δ$) are calculated for each protein.

2.- Upper bounds are imposed on the corresponding enzyme usage pseudo-reactions as follows:

$$ub\_{e\_{i}}= \left[E\_{i}\right]+1.96\*δ$$

The addition of $1.96\*σ$ accounts for a confidence interval of 0.95 in the protein abundance measurement.

3.- The experimental value for cellular growth rate at which the proteomics samples were obtained is then fixed as lower bound for the biomass pseudo-reaction and measured fluxes on glucose uptake rate and, optionally, byproducts secretion rates are set as upper bounds for their respective exchange reactions (adding a numerical tolerance of 5%).

4.- The remaining total protein pool is constrained by

$$ub\_{P\_{pool}}= \left(P\_{total}-\sum\_{i}^{}M\_{w\_{i}}\*\left[E\_{i}\right]\right)\*σ\*f$$

where $P\_{total}$ is the measured total protein content in the cell in gprot/gDw; $M\_{w\_{i}}$is the molecular weight of the measured protein *i*; $σ$ represents an average saturation factor for the unmeasured enzymes, assumed as 0.55,6; $f$ accounts for the fraction that the unmeasured protein sector represents out of the total proteome in the cell, this value is calculated by using a paxDB proteome abundance file for the organism of interest as a reference, if no paxDB file is provided then a value of 0.5 is assumed.

5.- Protein abundances are corrected for the oxidative phosphorylation complexes, trying to avoid overconstraining of potentially erroneously measured subunits that might limit the whole pathway. This correction is limited just to this pathway as it is desirable to modify the original dataset the least possible and abundance changes in OxPhos subunits are key to meet the phenotype energy requirements. Medium constraints are set by allowing free uptake of all compounds available in the culture medium and closing the rest of the uptake reactions. Additionally, all the upper bounds for production reactions (secretion of metabolites) are set to 1000 mmol/gDw h.

6.- Using an ***ecModel\_batch*** with the same constraints setup, minimal enzyme requirements for the proteins present in the filtered dataset are retrieved from a parsimonious FBA solution vector. Enzyme abundances that are lower than the minimum requirements calculated by the FBA solution are corrected in the dataset.

7.- A proteomics constrained ***ecModel\_prot*** is obtained by the function ***constrainEnzymes.m***. If the model is overconstrained after imposing all the afore mentioned constraints, then the function ***flexibilizeProteins.m*** flexibilizes the top-limiting abundances (based on shadow prices for the measured proteins, given by: $shadow prices= \frac{∂v\_{bio}}{∂ub\_{e\_{i}}}$) until the model is able to grow at the provided experimental growth rate. After this, an optimal enzyme usage profile compatible with the provided constraints is obtained and optimal levels are set as upper bounds for the flexibilized protein usages.

8.- The total flexibilized mass of protein is drawn from the remaining protein pool (upper bound for protein\_pool\_exchange pseudo-reaction) for consistency with mass conservation. Non-growth associated ATP maintenance is fitted according to condition specific experimental data if available (measurements on exchange fluxes of oxygen and CO2 from the same samples as the proteomics dataset). In the case of chemostat samples such conditions are set by first fixing the growth rate to the experimental value, minimizing the carbon source uptake, fixing its optimal value and then setting the total unmeasured enzymes usage as a new objective to minimize. Each condition-specific model is saved in ***GECKO/models/prot\_constrained*.**

# **Comparative flux variability analysis**

The function ***comparativeFVA.m*** in the FVA utitilities module provides a fair comparison of flux variability range distributions between a given GEM and its ecModel pair for glucose limited conditions (low dilution rates) and protein limiting regime (batch growth). The procedure for performing the flux variability analysis in this study is summarized by the following steps.

1.- For the chemostat case, a dilution rate of 0.1 h-1 was set as both lower and upper bound for the biomass pseudo-reaction (+/- a tolerance value of 0.01%).

2.- The glucose uptake rate is set as an objective to minimize and its optimal value is then also fixed, using the same tolerance.

3.- Additional culture medium constraints are imposed (upper bound for exchange reactions of mineral minimal medium components were set to 1000 mmol/gDw h).

4.- All applied constraints are also applied to the original GEM.

5.- For every reaction that is able to carry a non-zero flux in the original GEM (assessed by the RAVEN toolbox function ***haveFlux.m***) both minimization and maximization are performed for the original GEM.

6.- For the ecModel, such optimizations are performed on the governing pseudo-reaction representing the same original reaction flux (i.e. arm reactions when isoenzymes are present), this is done for both the forward reaction and its reversible counterpart (if present). In order to avoid the introduction of artificial variability, the forward reaction is blocked when the backwards is optimized, and the same is applied to the opposite direction.

7.- For each reaction a flux variability range is given by

$$FV\_{i}=v\_{i}^{max}-v\_{i}^{min}$$

8.- For the ecModel these ranges are given by

$$FV\_{i}=\left(v\_{i}^{max}-v\_{i}^{min}\right)-\left(v\_{i,REV}^{max}-v\_{i,REV}^{min}\right)$$

For the protein-limiting case, steps 1 and 2 are substituted by the following. Biomass production is maximized with the ecModel and then the optimal value is used to set a lower bound on the same reaction. In order to compare fairly with the original GEM, the same optimal growth rate is fixed as both lower and upper bounds for the biomass pseudo-reaction. A parsimonious flux distribution in which the total protein usage is minimized in the ecModel subject to all of the previous constraints is then obtained, an optimal glucose uptake rate is taken from this flux distribution and fixed for both ecModel and GEM.

# **Absolute protein quantification**

## **Total protein extraction**

 Cells samples for *S. cerevisiae, Y. lipolytica* and *K. marxianus* were washed and suspended in 3 ml of lysis buffer containing: 6 M urea (Sigma, U5378), 2M thiourea (Sigma, T8656), 5mM dithiothreitol (DTT, Sigma, D0632) and 0.1 M TRIS-HCl pH=8. Cell lysis occurred in a cell disruptor (Constant systems Ltd, One shot model) at 2.4 Kbars and supernatant was recovered by centrifugation (15 min at 4000 g, 4° C). Total protein concentrations were measured according to the 2D Quant kit protocol (GE Healthcare Life Sciences, 80-6483-56).

## **Proteins digestion**

Digestion was performed on excised bands from SDS-PAGE short-migration (1x1 cm lanes, Invitrogen, NP321BOX) gel gradient. Reduction was carried in DTT solution at 10 mM for 30 minutes at 56 °C. The extracts were alkylated with iodoacetamide solution at 55 mM for 45 min in darkness at room temperature (RT). Samples were first digested for 3 h at 37 °C by adding 300 ng of Lysyl-Endopeptidase (Wako, 125-05061). Then, a second digestion was performed with 300 ng of sequencing-grade modified trypsin (Promega) overnight at 37 °C. Supernatants were recovered and the peptides were extracted with a mixture of 0.5% trifluoroacetic acid (TFA) and 50% acetonitrile (ACN) in water. Extracted tryptic peptides were vacuum dried and resuspended in 75 μl of loading buffer containing 0.08% (v/v) of TFA and 2% (v/v) of ACN in water for mass spectrometry (MS) analysis. Simultaneously, the Universal Protein Standard 2 (UPS2, Sigma) was digested in a similar fashion to the experimental samples. The extracted tryptic peptides from UPS2 were vacuum dried and resuspended in 25 μl of loading buffer. We took 1.5-μl samples from the mixture of UPS2 digested peptides (424 ng/μl) and spiked them into 7.5 μl of each of the bulk samples (200 ng yeast peptides/μl) at a ratio of 1:2.35 (UPS2:yeast).

## **Mass spectrometry analysis**

MS analyses were performed on a Dionex U3000 RSLC coupled to an Orbitrap Fusion™ Lumos™ Tribrid™ mass spectrometer (Thermo Fisher Scientific) using a packed column Aclaim™ PepMap™, 75 μm x 500 mm, C18, 3 μm, 100 Å, (Thermo Fisher Scientific). Buffer A consisted of 0.1% formic acid in 2% ACN and buffer B of 0.1 % formic acid in 80% ACN. The peptide separation analysis was achieved at 300 nL/min with a linear gradient from 1 to 35% buffer B for 160 min and 35% to 50% for 10 min. One run took 195 min including the regeneration step at 98 % buffer B. Ionization (1.6 kV ionization potential) and capillary transfer (275°C) were performed with a liquid junction and a capillary probe (SilicaTip™ Emitter, 10 μm, New Objective). MS/MS analysis was performed in Data Dependent Acquisition mode, with a top speed cycle of 3 s for the most intense double or multiple charged precursor ions. Ions in each MS scan over threshold 50,000 were selected for fragmentation (MS2). The mass spectrometer acquisition settings were set as follows. Full MS scan in Orbitrap (scan range [m/z] = 400–1600) with a resolution of 120,000 (AGC target = 5 x 105, max. injection time of 100 ms, data type = centroid). The dynamic exclusion within 10 ppm during 60 s and the intensity threshold was fixed at 5 x 104. And MS/MS using High Collision Dissociation (HCD) in the Orbitrap with resolution of 15,000 (30% collision energy, AGC target of 5.0 x 104 and max. injection time = 54 ms). Polysilaxolane ions m/z 445.12002, 519.13882 and 593.15761 were used for internal calibration.

## **Estimation of absolute protein abundance values**

Normalized Spectral Abundance Factor (NSAF)7 values were calculated for all detected proteins based on MS/MS data for all detected proteins. The NSAF values obtained from UPS2 proteins in bulk samples were used to determine the suitable regression curves that allowed the conversion from relative protein abundance into absolute terms. The obtained linear regression parameters are shown in **Table S2.2**. This enabled estimation of protein concentrations (*Qi*) in the analyzed bulk samples, which consisted of concentrated mixes of peptides, in units of mmol/gprotein. In order to obtain specific protein abundances, referred to the original cells dry weight, total protein content in the initial biological samples was measured by using the Lowry method for protein quantification8. Finally, absolute protein abundance values were multiplied by the total protein content in their respective samples, yielding values in units of mmol/gDw for consistency with the ecModels units for protein usages.

**Table S2.2.- Calibration parameters for protein abundance estimation.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Strain** | **Equation** | **r2** | **Detected UPS2 proteins** |
| *S. cerevisiae* CEN.PK113 | log10( *Qi* ) = 0.666 \* log10( *NSAFi* ) – 3.48 | 0.901 | 25 |
| *K. marxianus* CBS6556 | log10( *Qi* ) = 0.619 \* log10( *NSAFi* ) – 3.33 | 0.870 | 27 |
| *Y. lipolytica* W29 | log10( *Qi* ) = 0.610 \* log10( *NSAFi* ) – 3.51 | 0.874 | 27 |

# **References**

1. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: A European ELIXIR core data resource. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1048

2. Davidi, D. & Milo, R. Lessons on enzyme kinetics from quantitative proteomics. *Current Opinion in Biotechnology* **46**, 81–89 (2017).

3. Price, N. D., Reed, J. L. & Palsson, B. Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897 (2004).

4. Bar-Even, A. *et al.* The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).

5. Nilsson, A. & Nielsen, J. Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase. *Sci. Rep.* **6**, 1–11 (2016).

6. Sánchez, B. J. *et al.* Improving the phenotype predictions of a yeast genome‐scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).

7. Zybailov, B. *et al.* Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* (2006). doi:10.1021/pr060161n

8. Waterborg, J. H. & Matthews, H. R. The Lowry method for protein quantitation. *Methods in molecular biology (Clifton, N.J.)* **32**, (1994).