Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Metabolic network reduction based on Extreme Pathway sets

Wannes Mores[a], Satyajeet S. Bhonsale[a], Filip Logist[a], and Jan F.M. Van Impe[a]*

[a] BioTeC+ KU Leuven, Department of Chemical Engineering, Gent, Belgium
* Corresponding Author: jan.vanimpe@kuleuven.be.

ABSTRACT

The use of metabolic networks is extremely valuable for design and optimisation of bioprocesses as they provide great insight into cellular metabolism. Within bioprocess optimisation, they have enabled better (economic) objective performance through more accurate network-based models. However, one of the drawbacks of using metabolic networks is their underdeterminacy, leading to non-unique flux distributions. Flux Balance Analysis (FBA) reduces this issue by making assumptions on the behaviour of the cell. However, for metabolic networks of higher complexity, can still struggle with underdeterminacy. Metabolic network reduction can remove or greatly reduce this effect but can be difficult, especially when data is limited. Structural analysis of the metabolic network through Elementary Flux Modes (EFM) or Extreme Pathways (EP) can help locate the relevant information within the network. This work presents a metabolic network reduction approach based on the EPs that best explain a small set of available measurements. Many of the reactions will not be active during the process and a significantly smaller network can therefore be constructed. A case study of oxygen-limited *Escherichia coli* is presented which showcases this approach, enabling accurate prediction of the process with much smaller network-based models. This leads to much lower complexity bioprocess models while keeping the necessary information on cellular metabolism for the given process.

Keywords: Model Reduction, Multiscale Modelling, Biosystems

INTRODUCTION

Model-based optimisation of bioprocesses has already shown great potential, enabling higher return on economic objectives for a given process. In many optimisation studies, a macroscopic model is used which relies on kinetic expressions such as Haldane or Monod to quantify the dynamic evolution of (by-)product and substrate concentrations within the reactor. However, these types of models do not consider the complex interactions on the intracellular level, which has been shown to lead to poor predictive performance in a dynamic context [1].

The incorporation of intracellular information is possible through metabolic networks, which has led to significant improvements in objectives such as yield, productivity, and final concentration of product [2]. However, these studies rely on smaller, less complex metabolic networks as they require a high computational effort for the optimization study. Currently, many strains used for bioproduction have a genome-scale metabolic network (GEM) available, easily surpassing 1000 reactions. These GEMs are not suitable for the existing optimisation approaches, highlighting the need for reducing the complexity of the metabolic network, while ensuring minimal loss of key information regarding the bioprocess.

Constraint-based modelling

The metabolic network can be represented through its stoichiometric matrix $\boldsymbol{S} \in \mathbb{R}^{m \times r}$, which links metabolites m to the reactions r they take part in. Analysis of metabolic networks usually relies on the constraint defined by the Pseudo-Steady-State-Assumption (PSSA):

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{1}$$

This implies that there is no accumulation of metabolites within the network. The reaction rates or fluxes are then captured in vector $\boldsymbol{v}$. Constraint-based modelling techniques such as Flux Balance Analysis (FBA) are built

on this assumption. As $\boldsymbol{S}$ is not a full column rank, there is more than one possible solution for the network, making it an underdetermined system. Moreover, the solution space of the PSSA is usually high-dimensional, especially for large networks. Even with measurements to further constrain the solution space, many possible solutions will exist. This issue of underdeterminacy can be dealt with in many ways [3]. Ideally, the underdeterminacy should be eliminated or reduced significantly.

One way of dealing with the underdeterminacy is by assuming that the cell follows a specific metabolic target (e.g., maximising growth). Flux Balance Analysis or FBA solves an optimsation problem with the assumed objective. In many cases, this leads to good prediction of metabolic behaviour, but can still lead to non-unique solutions. Lexicographic optimization can ensure a unique solution but needs a priority list of metabolic objectives [4].

Metabolic Flux Analysis or MFA aims to estimate the fluxes based on only measurement data. For a dynamic MFA-based model, kinetic equations are used to describe some of the key fluxes. Nimmegeers et al. [5] used the concept of free fluxes, which are directly related to the degrees of freedom of the PSSA. For large metabolic networks, this leads to too many parameters to be estimated, especially when data is limited. Reducing the degrees of freedom while keeping the relevant information of the network is therefore key in this approach.

Metabolic network reduction

Many metabolic network reduction approaches exist in literature, mainly focusing on retaining as much information as possible. Most of them can be described as a top-down or bottom-up approach. With top-down approaches [6,7], the full network is the starting point. With techniques such as Mixed-Integer Linear Programming (MILP), reactions are then iteratively removed. With bottom-up approaches [8], smaller subnetworks are the starting point, which are recombined to mimic the behaviour of the original network while maintaining a certain limit on the size of the network.

Alternatively, small macroscopic models can be constructed based on vectors characterising the solution space such as EPs and EFMs. Maton et al. [9] enumerated EFMs for a medium-scale metabolic network and selected the most informative ones. This leads to good reproduction of the experimental data and therefore a promising approach to analyse and model certain aspects of the metabolic network. Moreover, by selecting an upper bound on the number of EFMs, the degrees of freedom or free (macro-)fluxes can be directly controlled.

In this work, the potential of using characterising vectors is further investigated, evaluating how they can help reduce and extract relevant information from large metabolic networks. A relevant case study showcases the applicability of this method to model bioprocesses.

MATERIALS AND METHODS

Case study

A dFBA-based multi-scale model is used to simulate the oxygen-limited growth of *E. coli* on glucose in a batch reactor. The e_coli_core model [10] was used to estimate the fluxes in CobraPy throughout the process, applying bounds on the uptake fluxes through kinetic expressions. The dynamic model will consider 6 metabolites of interest: biomass, glucose, oxygen, acetate, formate, and ethanol. A term for the oxygen transfer from the gas phase to the liquid is incorporated, using a mass transfer coefficient of $k_L a = 7.5\ hr^{-1}$ [11]. The model can then be described as:

$$\frac{dC_X}{dt} = v_X \cdot X \tag{2}$$

$$\frac{dC_{glc}}{dt} = v_{glc} \cdot X \tag{3}$$

$$\frac{dC_{O_2}}{dt} = v_{O_2} \cdot X + k_L a(0.21 - C_{O_2}) \tag{4}$$

$$\frac{dC_{ac}}{dt} = v_{ac} \cdot X \tag{5}$$

$$\frac{dC_{for}}{dt} = v_{for} \cdot X \tag{6}$$

$$\frac{dC_{eth}}{dt} = v_{eth} \cdot X \tag{7}$$

The extracellular fluxes corresponding to the metabolites of interest are obtained from FBA at each timestep. The bound on the glucose uptake rate v_{glc} was defined by Monod kinetics based on the glucose concentration C_{glc} as follows [12]:

$$0 \geq v_{glc} \geq -10\frac{C_{glc}}{\mathbf{5.55}+C_{glc}} \tag{8}$$

The system of ODEs is then passed to SciPy's solve_ivp function, together with the initial conditions $\boldsymbol{C}_0 = [0.01\frac{g}{L}, \frac{20}{180}mM,\ 0.21\ mM, 0.4\ mM, 0\ mM, 0\ mM]$. Simulation was done on a virtual machine with a Intel(R) Core(TM) i5-1235U processor and 8GB of RAM available. The dFBA simulation ran in 133 seconds. For the e_coli_core [10] network, all the EPs are enumerated using an accelerated version of Canonical Basis Approach [13] developed in Mores et al. [14].

Selection procedure

Given that the number of EPs is usually quite large, efficient procedures have to be developed to select the most informative EPs for the process. In this work, a two-step approach is presented. First, a preselection step based on yield is carried out to get a more manageable number of EPs. This is followed by an iterative routine based on optimization problems to find the final set of EPs to be used to create smaller biological models.

Preselection based on yield analysis

Since not all EPs carry relevant information regarding the case study, a preselection is carried out to reduce the large number of EPs to a more manageable size. With yield analysis [15], the EPs are projected into yield space. The metabolites are split into products and substrates, which is used to define the yields of each EP. From the metabolites of interest in the case study, glucose and oxygen are the only substrates. As specified in Ramkrishna, a reference substrate is chosen (glucose) and the other substrates are given a minus sign. Within this yield space, a convex hull is constructed for the set of EPs. The points that define this convex hull encompass and therefore be able to describe the experimental data.

Final EP set selection

After a significant reduction of candidate EPs through yield analysis preselection, different optimization problems will be defined to aid in selection the most informative EPs regarding a dataset. These optimization problems are based on minimization of error to the measurement, defined as follows [9]:

$$SSE = \sum_{k=1}^{met_i} (\boldsymbol{EP}_k \cdot \varphi - \boldsymbol{v}_{obs,k}) \boldsymbol{W}^{-1} (\boldsymbol{EP}_k \cdot \varphi - \boldsymbol{v}_{obs,k})^T, \quad (9)$$

where φ corresponds to the macro flux vector of the extreme pathways, $\boldsymbol{v}_{obs}$ corresponds to the measured extracellular flux, and $\boldsymbol{W}$ corresponds to a diagonal matrix of weights defined by $(v_{obs,k})^2$, the maximum flux for each measured exchange reaction.

First, an initial sweep is done over the set of EPs where one EP is has its macro flux set to zero. If disabling an EP this way leads to an insignificant change to the SSE resulting from the optimization, the EP is considered to not carry relevant information regarding the process and therefore is removed. Important to note is that the order of this initial sweep does have an influence on the result. Here we sweep stochastically.

To get the final selection of EPs, the worst performing EP will be removed iteratively until a desired set size is achieved. Every iteration, each remaining EP is set to inactive separately and the SSE is re-evaluated. The EP whose inactivation leads to the best score is then selected for removal as it is the current worst performer.

Model reduction approaches

To reduce the network-based model, two approaches will be used. Based on the results from EP selection, models are created by either considering the EPs separately or by reconstructing a network based on the active reactions in the selected EPs.

Reduced network model

Since the EPs selected are the most informative pathways for the given process, they contain the most valuable, informative reactions. By reconstructing a network based on the reactions active in the set of EPs, a small network is created which focuses on relevant aspects of the metabolism. Direction of the reactions in the EPs is also considered, making reversible reactions with only one relevant direction irreversible. One of the benefits of using EPs here is that the pathways remain feasible since the EPs are minimal steady-state solutions of the system. This means that the reduced network still allows for techniques such as FBA to be performed.

Macro-reaction model

Alternatively, the EPs themselves can be used as a model substituting for the metabolic network. This approach is similar to Maton et al. [9], where the macro-fluxes for the most informative EPs is used to model the process. Simulating with this model is done by maximization of the biomass output, similar to FBA. However, here the macro-fluxes are manipulated to achieve the objective instead of the individual fluxes of each reaction.

RESULTS

Using the model described by Equations 2-8, a dataset is obtained for oxygen-limited *E. coli*. The dataset consists of 74 datapoints over a 15 hour time period. To account for ATP when selecting EPs, the minimal ATP requirement is added to each timepoint as an additional state. From this concentration data, the fluxes per individual cell were isolated by dividing the concentrations by the biomass.

Extreme Pathway generation

Based on the e_coli_core network, the full set of EPs is generated using the CBA method as described previously. The final set consists of 100,235 EPs, which is too large to start applying the LP-based selection. Hence, a preselection step is done based on yield analysis. Afterwards, LP-based selection is done to try and reduce the EP set below the number of measured metabolites [9], in this case less than 7 since ATP is included as an additional metabolite.

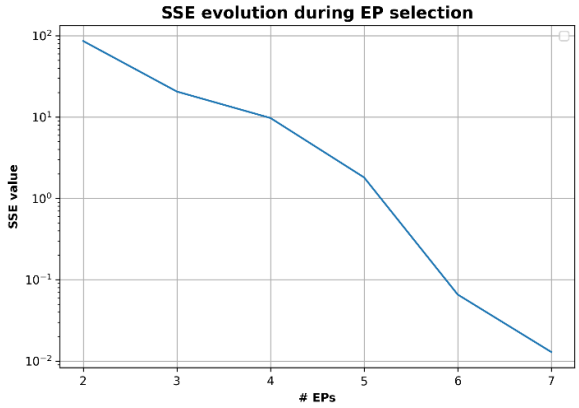

Figure 1. SSE scores during the final selection procedure

Preselection

Preselection of EPs is done based on yield analysis, with glucose selected as the reference substrate for the process and all other metabolites from the process model as the products. In the yield space, a convex hull is then constructed, whose points should be able to fully represent the dataset generated from FBA. The EPs kept after preselection are then those who correspond to the vertices of the convex hull in yield space. This leads to a significant reduction of EPs, with only 113 EPs remaining.

LP-based selection

A first sweep aims to eliminate the least informative EPs. The order of EPs is first randomized as the order does influence the selection results to try and reduce the bias towards some EPs. The threshold for significance is set at an SSE difference of 0.001. When an EP's removal leads to this level of decline in the SSE score, it is kept. This first sweep reduces the EP set to just 8 significant EPs. However, the goal is to have less than 7 EPs. Therefore, EPs are iteratively removed based on the worst-performer as described previously. SSE values are shown throughout this iterative reduction of EPs in Figure 1. Based on the trade-off between complexity and accuracy, a decision then has to be made on how many EPs to keep. In this work, 5 EPs are selected as it is below the threshold of 7 while keeping relatively good SSE values.

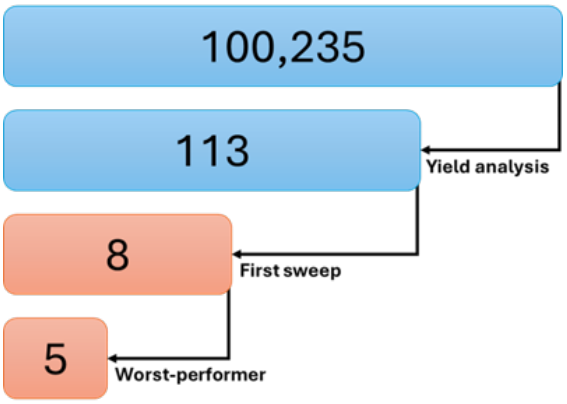

Figure 2. Evolution of EP set sizes with each reduction step. Blue indicates the yield analysis method, red indicates the LP-based reduction.

Reduced Network Model

Important reactions for the process can be found by analysing the remaining EPs. Each reaction that does not take part in any of the remaining EPs is not needed to reproduce the observed metabolic behaviour. Therefore, these reactions can be seen as less important and can be removed from the network. This already reduces the complexity of the network significantly, but considering reversibility enhances this even more. If only one direction of a reversible reaction is present in the EPs, it can be converted within the network-based model to an irreversible reaction in the relevant direction for the process.

When this principle is applied to the remaining EPs, the network is reduced from 95 reactions to 60. From these remaining reactions, 25 are converted from reversible to irreversible in the relevant direction. When purely considering the number of reactions, this does not seem like a significant reduction. However, as discussed previously, the target for the reduction in this work is more aimed towards a reduction in free fluxes. The complexity of this reduced network is significantly smaller, needing only 6 free fluxes to be estimated instead of the original 28.

To validate if this reduced network is still capable of representing the same metabolic behaviour as the full network, dFBA is also applied to the reduced network with the same initial conditions. Even though the complexity of the network is reduced significantly, dFBA on the reduced network lead to a perfect reconstruction of the original dataset. This highlights the potential of this targeted network reduction approach with limited process information requirements, which can help reduce computational effort and data requirements for estimation of a network-based bioprocess model.

Macro-reaction model

By using the EPs remaining as macro-reactions, another type of reduced model can be derived. Since there are 5 EPs as was selected, 5 macro-fluxes have to be estimated. The complexity of this network is thus even smaller than the reduced network approach since only 5 macro-fluxes must estimated instead of the previously discussed 6 free fluxes.

By using the same principle of biomass maximization using the 5 EPs as individual macro-reactions with the available uptakes of glucose and oxygen, metabolic behaviour can be approximated. Using the same initial conditions and kinetic equations as the original dataset, very similar metabolic behaviour can be observed throughout the process. Only very minor differences can be observed between the results of the macro-reaction model and the original network-based model as can be seen in Figure 3. Compared to the network-based models, the simulation time is reduced significantly, from 133 seconds to just 16 seconds.

CONCLUSION

In this work, new approaches were developed to reduce metabolic networks when a limited amount of information is available. Using extracellular measurements, relevant aspects of the network can be identified and kept in the reduced network by employing static analysis of the network through Extreme Pathways. A relevant case study was defined, where extracellular measure-

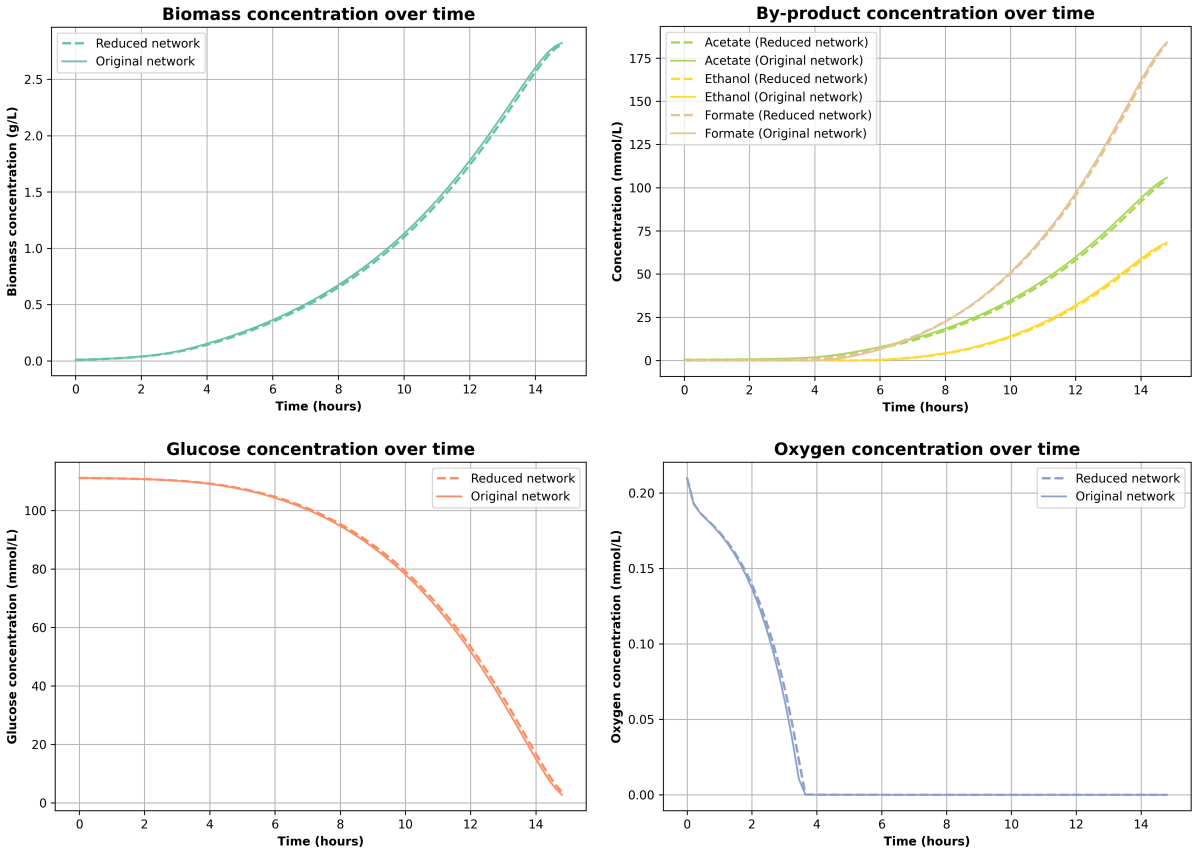

Figure 3: Reactor concentration profiles over the duration of the process. Solid lines are obtained using the full network-based model in CobraPy, dashed lines are obtained by finding the optimal macro-fluxes for the selected EPs every timestep.

ments were substituted through simulation based on a metabolic network of *E. coli*. Even though the network used is relatively small, it still suffers from the effects of underdeterminacy. In total, 26 free fluxes have to be estimated either through measurements or by making assumptions such as FBA objective functions. Many previous studies have focused on reducing the network to a minimal number of reactions, but free fluxes or degrees of freedom give a better idea of the complexity of the network. Analysis of the metabolic network is done through EPs, which are valid pathways through the network and adhere to the Pseudo-Steady-State-Assumption. The total EP set is very large, necessitating a preselection step based on the concept of yield analysis [15]. The remaining 113 EPs are then evaluated based on their ability to reproduce the measured metabolic behaviour of oxygen-limited *E. coli* and the 5 most informative EPs are finally selected to be used for model reduction.

Two model reduction approaches are defined and implemented for the case study. The first one reconstructs a network based on the active reactions in the 5 selected EPs. If a reversible reaction is only active in a single direction, it is transformed into an irreversible reaction in the relevant direction. This leads to a much less complex metabolic network, reducing the free fluxes from 28 to just 6 while keeping identical behaviour in the case study.

A second approach substitutes the network by the 5 EPs as macro-reactions themselves. Instead of estimating all intracellular fluxes, only 5 macro-fluxes then have to be estimated instead of the 28 free fluxes. Using this approach, only a small difference is found within the case study between the full network and the macro-reaction model while complexity is reduced significantly.

The results in this work indicate that analysis of metabolic networks through extreme rays such as EPs aids significantly in understanding and finding the relevant aspects of metabolism in a network given a limited amount of information on the process. Significantly reduced models can be created successfully using the most relevant EPs while keeping accurate metabolic behaviour. This leads to a much less complex estimation of the fluxes, potentially requiring much less data to get quality bioprocess models.

ACKNOWLEDGEMENTS

This work is funded by the European Union under grant agreement 101122224 ('ALFAFUELS'). W.M. was supported by Research Foundation Flanders (FWO) through Strategic Basic Project 1SHG124N.

REFERENCES

1. Hodgson, B. J., Taylor, C. N., Ushio, M., Leigh, J. R., Kalganova, T., & Baganz, F. Intelligent modelling of bioprocesses: a comparison of structured and unstructured approaches. *Bioprocess Biosyst. Eng.*, 26, 353-359. (2004). https://doi.org/10.1007/s00449-004-0382-0
2. Chang, L., Liu, X., & Henson, M. A. Nonlinear model predictive control of fed-batch fermentations using dynamic flux balance models. *J. Process Contr.*, 42, 137-149. (2016). https://doi.org/10.1016/j.jprocont.2016.04.012
3. Bogaerts, P., & Vande Wouwer, A. How to Tackle Underdeterminacy in Metabolic Flux Analysis? A Tutorial and Critical Review. *Processes*, 9(9), 1577. (2021). https://doi.org/10.3390/pr9091577
4. Gomez, J. A., Höffner, K., & Barton, P. I. DFBAlab: a fast and reliable MATLAB code for dynamic flux balance analysis. *BMC Bioinform.*, 15, 1-10. (2014). https://doi.org/10.1186/s12859-014-0409-8
5. Nimmegeers, P., Vercammen, D., Bhonsale, S., Logist, F., & Van Impe, J. Metabolic reaction network-based model predictive control of bioprocesses. *Appl. Sci.*, 11(20), 9532. (2021). https://doi.org/10.3390/app11209532
6. Röhl, A., & Bockmayr, A. A mixed-integer linear programming approach to the reduction of genome-scale metabolic networks. *BMC Bioinform.*, 18, 1-10. (2017). https://doi.org/10.1186/s12859-016-1412-z
7. Erdrich, P., Steuer, R., & Klamt, S. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. *BMC Syst. Biol.*, 9, 1-12. (2015). https://doi.org/10.1186/s12918-015-0191-x
8. Ataman, M., Hernandez Gardiol, D. F., Fengos, G., & Hatzimanikatis, V. redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. *PLOS Comput. Biol.*, 13(7), e1005444. (2017). https://doi.org/10.1371/journal.pcbi.1005444
9. Maton, M., Bogaerts, P., & Wouwer, A. V. A systematic elementary flux mode selection procedure for deriving macroscopic bioreaction models from metabolic networks. *J. Process Contr.*, 118, 170-184. (2022). https://doi.org/10.1016/j.jprocont.2022.09.002
10. Orth, J. D., Fleming, R. M., & Palsson, B. Ø. Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide. *EcoSal plus*, 4(1), 10-1128. (2010). https://doi.org/10.1128/ecosalplus.10.2.1
11. Mahadevan, R., Edwards, J. S., & Doyle, F. J. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys. J.,* 83(3), 1331-1340. (2002). https://doi.org/10.1234/56789.10
12. Scott, F., Wilson, P., Conejeros, R., & Vassiliadis, V. S. Simulation and optimization of dynamic flux balance analysis models using an interior point method reformulation. *Comput. Chem. Eng.*, 119, 152-170. (2018). https://doi.org/10.1016/j.compchemeng.2018.08.041
13. Schuster, S., Fell, D. A., & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, *18*(3), 326-332. (2000). https://doi.org/10.1038/73786
14. Mores, W., Bhonsale, S. S., Logist, F., & Van Impe, J. F. Accelerated enumeration of extreme rays through a positive-definite elementarity test. *Bioinformatics*, btae723. (2024). https://doi.org/10.1093/bioinformatics/btae723
15. Song, H. S., & Ramkrishna, D. Reduction of a set of elementary modes using yield analysis. *Biotechnol. Bioeng.*, 102(2), 554-568. (2009). https://doi.org/10.1002/bit.22062