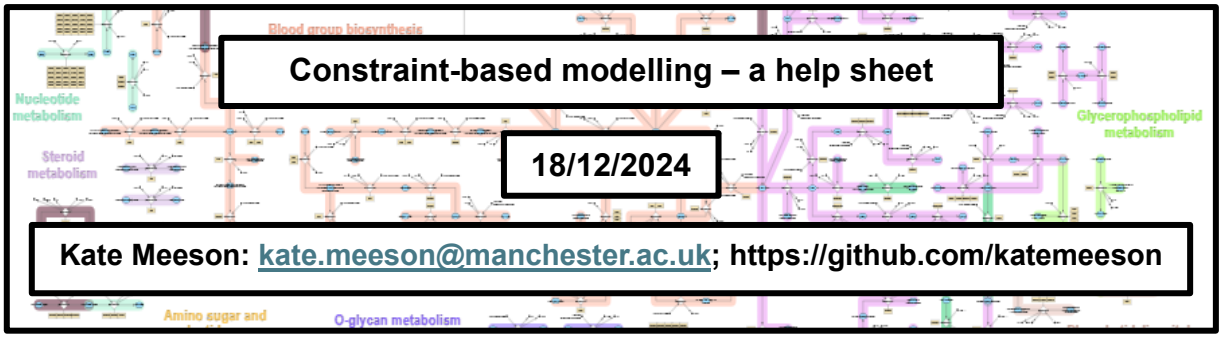

Constraint-based modelling – a help sheet

18/12/2024

Kate Meeson: kate.meeson@manchester.ac.uk; https://github.com/katemeeson

Definitions

Genome-scale metabolic model (GEM): a computational reconstruction of the metabolic network of an organism, containing metabolic reactions connected by their metabolites, and overlaid with regulatory enzyme annotations and Gene IDs

Objective function: The single or multiple reactions which have been selected to be optimised during flux balance analysis or flux variability analysis. An objective function can be maximised, minimised or sub-optimal

Constraint: Gene or protein expression value or other 'omics measurement which has been converted to reaction boundaries within a GEM to reduce the solution space; constraints can be integrated using pre-existing algorithms

Flux: the rate of a metabolic reaction; measured in mmol/gDW/hour for all reactions except growth rate, which is g/gDW/hour, however, always double-check source code for proper units

Gene-protein-reaction rules (GPRs): a linear formula containing 'and', 'or' or 'andor', informing the enzyme(s) and respective subunits required to catalyse a reaction

Flux balance analysis: a method of solving a GEM, which uses an objective function, works off the steady-state assumption and outputs a single solution

Flux variability analysis: a method of solving a GEM, similar assumptions to FBA, but outputs a minimum and maximum instead of a single solution; more computationally intensive

Flux sampling: a method of solving a constrained GEM, does not assume an objective function and outputs a probability distribution across a range of solutions; considered the more reliable flux analysis method but is computationally intensive requiring a computational shared facility

1. What is it?

Constraint-based modelling is a method we use to predict metabolic fluxes in a computational model cell, to a scale which is not achievable experimentally (see ***Figure 1***). This method is a hypothesis-generating bioinformatics approach, which can suggest metabolic mechanisms underlying *in vitro* phenotypes, or which can predict targets for experimental work. Constraint-based modelling is used across a variety of projects, spanning industry and academia, such as chemical engineering to optimise antibody production in mammalian cells, find new drug targets in a cancer cell line or improve the taste of beer! Therefore, this modelling can contribute to very interesting interdisciplinary projects, suiting your personal research interests. This method has been around since the early 2000's, but every year there are new approaches and software emerging.

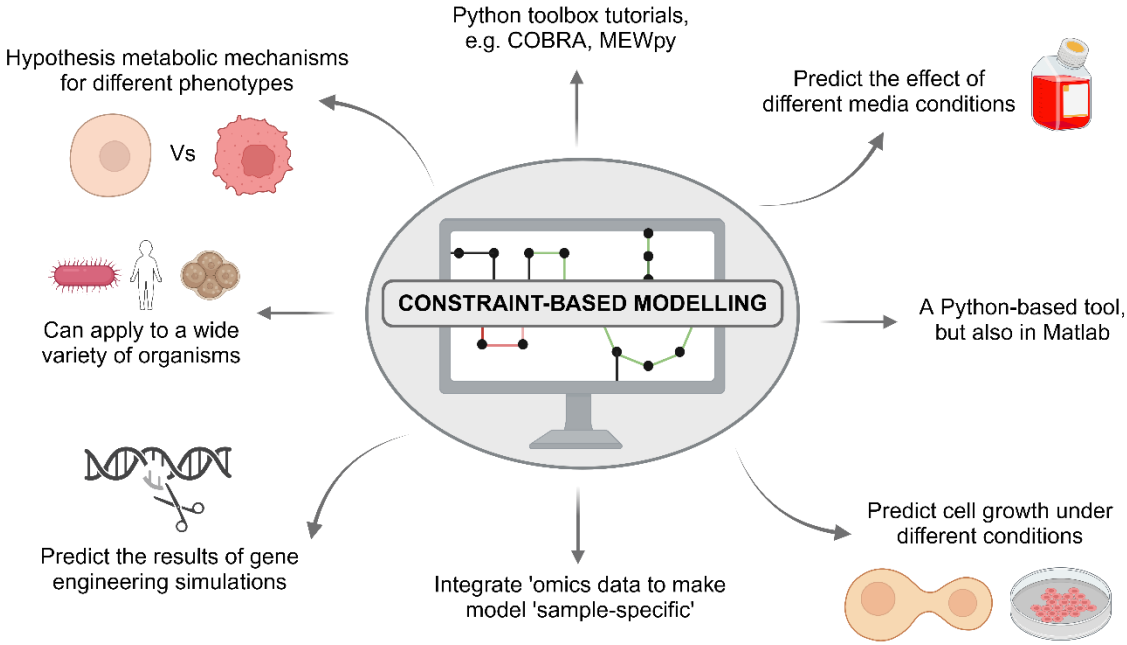

Figure 1. An overview of constraint-based modelling.

Constraint-based modelling is applied to genome-scale metabolic models (GEMs), which are *in silico* reconstructions of the metabolic network of a cell with a few different layers. GEMs connect individual metabolic reactions to one another in a multi-dimensional network, and layer information regarding the enzymes which regulate these pathways and the genes which encode these enzymes, so you can imagine how this method is attractive to groups with lots of 'omics data to work with. See ***Figure 2*** for an example of a two-dimensional representation of a GEM and its different elements.

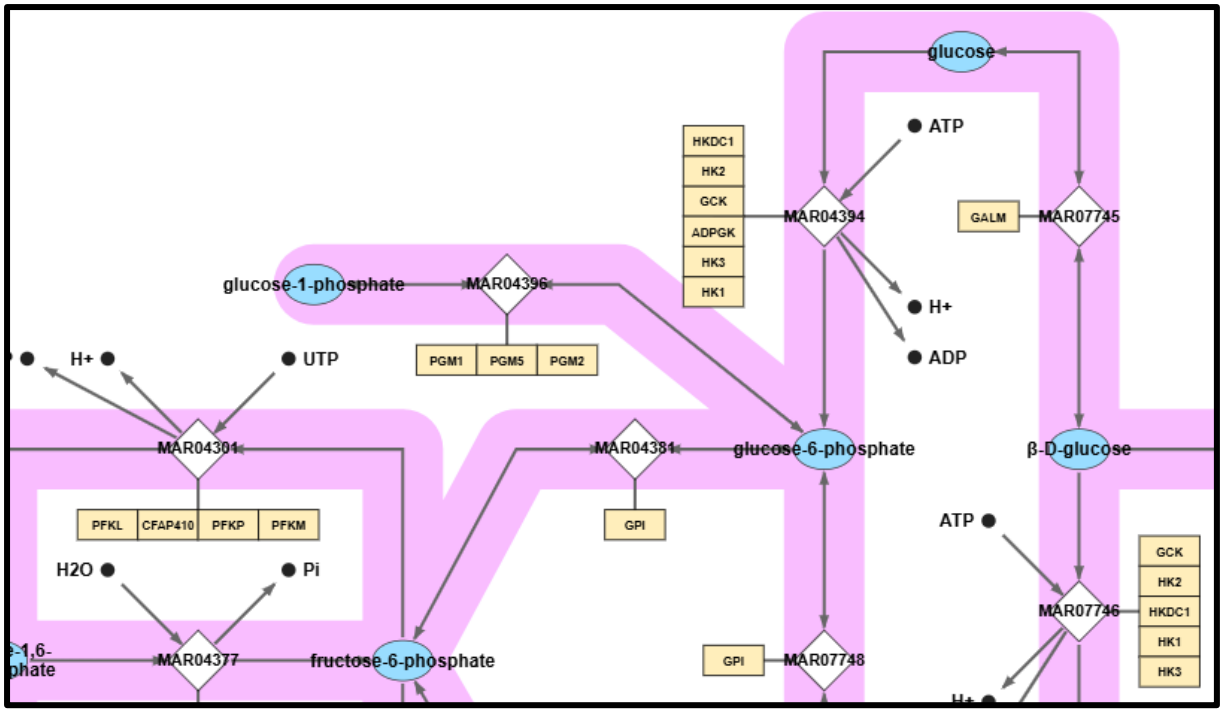

Figure 2. A 2D snapshot of a human GEM.

Here is a snapshot of the 'Glycolysis/Gluconeogenesis' subsystem from the Human1 genome-scale model (GEM) (Robinson et al., 2020). Here you can see metabolites (e.g. 'glucose') and reactions (e.g. MAR07745) – each of which have unique reaction IDs and regulatory enzymes (e.g. 'HK2': hexokinase 2). Visualisation from the Metabolic Atlas portal (https://metabolicatlas.org/).

There are base GEMs for a variety of organisms, which may be personalised using the integration of 'omics data, such as gene or protein expression values, media conditions and metabolite uptake or secretion rates. In this way, you have a sample-specific metabolic model which will act as the framework for your constraint-based modelling simulations. The integration of sample data and further details will be discussed below.

2. How do we do it?

Constraint-based modelling does involve coding, which is mainly Python-based, however some algorithms use Matlab. At the University of Manchester, one module which teaches this method is 'BIOL66021: Computational Approaches to Biology'. If you have not formally learnt constraint-based modelling and would like to, there are various tutorial documentations available online for free, with the main being COBRApy (Ebrahim et al., 2013), where there are tutorials for different simulations, such as changing media conditions and analysing fluxes or practising gene engineering simulations (https://cobrapy.readthedocs.io/en/latest/getting_started.html; https://opencobra.github.io/cobratoolbox/stable/tutorials/).

a. Choosing your metabolic model

Having been originally developed for bacteria, there are GEMs available for a wide variety of organisms, including human and yeast. A useful database to purvey these models is BioModels (https://www.ebi.ac.uk/biomodels/), which includes over 100 organisms! It might also be interesting to you to look through the patient-derived metabolic models, if you work on

cancer research (https://www.ebi.ac.uk/biomodels/pdgsmm/index). To begin with, it might be helpful to choose a model which has been well studied and worked with, so there are examples of its application for you to follow. Alternatively, more recent models with less associated literature might have new aspects and features which make working with them easier, so it's your choice! You may work with a comprehensive GEM, which includes all known metabolic reactions of an organism, or you can work with a simpler, reduced model containing only central metabolic pathways. The latter option might be more appropriate with more complicated flux analysis methods, which are more computationally intensive. Always go to their original publication to find examples of the model's use and to see how the model has been experimentally validated before its release.

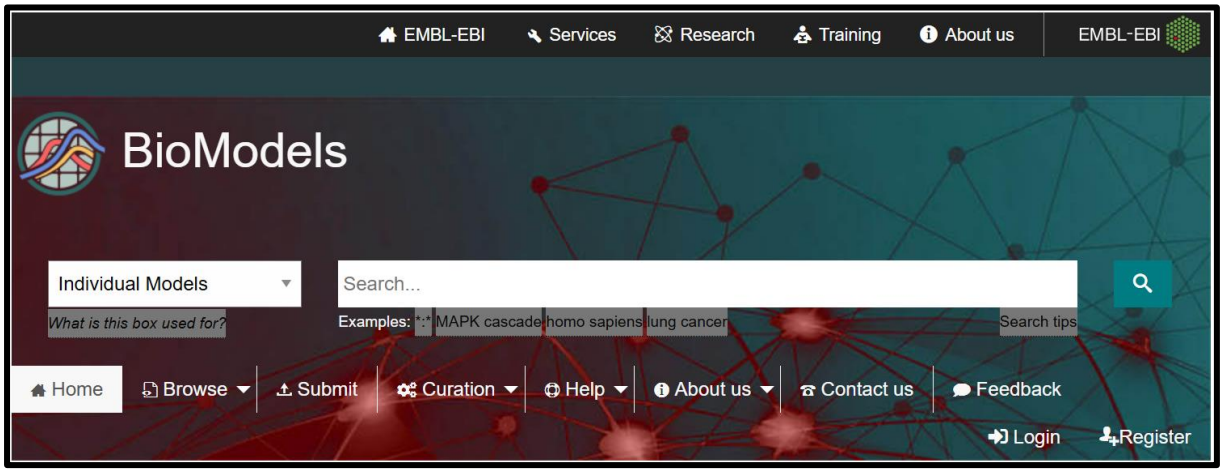

Figure 3. BioModels database. (https://www.ebi.ac.uk/biomodels/)

b. Personalising your metabolic model

The way in which a metabolic model is personalised towards your sample is by setting the lower and upper reaction bounds, which instruct the model in which direction (if the reaction is reversible) and to what magnitude (in the form of a 'flux' value) this specific reaction can be allowed to flow. Setting a reaction bound is referred to as adding a model 'constraint' and if this constraint is relatively tight, then this reaction can become a bottleneck in the model if there are many downstream reactions. On the other hand, if this reaction is left unconstrained, then there could be a predicted metabolic flux anywhere from 0 to the maximum upper reaction boundary (mmol/gDW/hour). An illustration for this data integration process has been shown in ***Figure 4***.

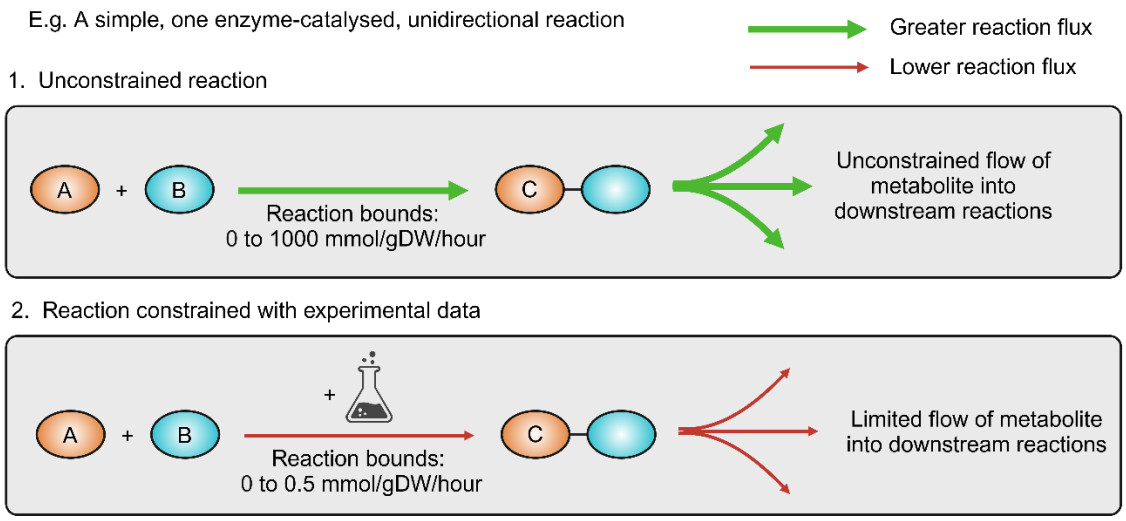

Figure 4. Illustration of setting reaction boundaries.

i. Defining media conditions within the GEM

There are multiple layers to personalising a generic GEM to be specific to your sample. The first step here, provided it has been properly defined, is to specify the concentration of extracellular metabolites which are available to your cell. In other words, you need to set the *in silico* media conditions. This is important as the extracellular concentration of a metabolite restricts what is available intracellularly, once this compound has been imported into the cell. This in turn affects the intracellular metabolic flux distribution. In order to define the *in silico* media conditions, you must alter the lower and upper bounds of the corresponding exchange reaction, which refers to the reactions taking place at the perimeter of the model. This is actually an artificial reaction to maintain some of the mathematical assumptions of the GEM ('the steady state') and to balance reactions secreting this same metabolite to the extracellular space. For example, if there is 10mM glucose in your cell culture, you might set the lower and upper boundaries of the corresponding exchange reaction to 0 and 10 mmol/gDW/hour, which translates to a maximum glucose uptake by the cell of 10 mM (equivalent to 10 g/L). There is guidance code and tutorials on the COBRA documentation: https://cobrapy.readthedocs.io/en/latest/media.html.

ii. Integrating 'omics data

Similarly to defining the media conditions, 'omics constraints are added at the GPRs in the GEM. However, here, instead of constraining metabolite exchange reactions, we will be defining the intracellular reaction boundaries. Enzyme expression values obtained from 'omics data, such as transcriptomics or proteomics, are translated to the lower and upper reaction boundaries, depending on the reversibility of the reaction. As shown in **Figure 4** this is fairly straightforward for unidirectional simple reactions. For example, the glycolytic reaction isomerising glucose-6-phosphate to fructose-6-phosphate in human cells may only be catalysed by glucose-6-phosphate isomerase, so if we have available expression values for this enzyme we are able to constrain it. However, some reactions are more complicated, involving isoenzymes, or genes encoding subunits of an enzyme. For this reason, we use well-validated integration algorithms which decide how the 'omics data is translated to reaction constraints and save us the hassle! Popular integration algorithms include iMAT and GIMME

(Becker & Palsson, 2008; Zur et al., 2010), whilst some algorithms have also been developed by members of our lab (Meeson & Schwartz, 2024; Timouma et al., 2024). Alternative data types can be integrated into the GEM, potentially reducing uncertainty, and these include kinetics, thermodynamics and genomics (Lewis et al., 2021).

iii. Integrating direct flux measurements

Alternatively, you may have access to direct flux measurements, for example, the glucose uptake rate, growth rate or protein production. Some of these measurements are more commonly acquired for industry-related projects, such as antibody production using mammalian cells (Strain et al., 2023). Although it is important to reserve some experimental measurements to validate your model's predictions, these values can be used to directly constrain their corresponding model reactions and personalise your model more towards a more realistic solution space. Together, the media conditions, 'omics measurements and direct flux measurements combine to create a personalised, sample-specific GEM, ready for flux simulations and analysis (***Figure 5***).

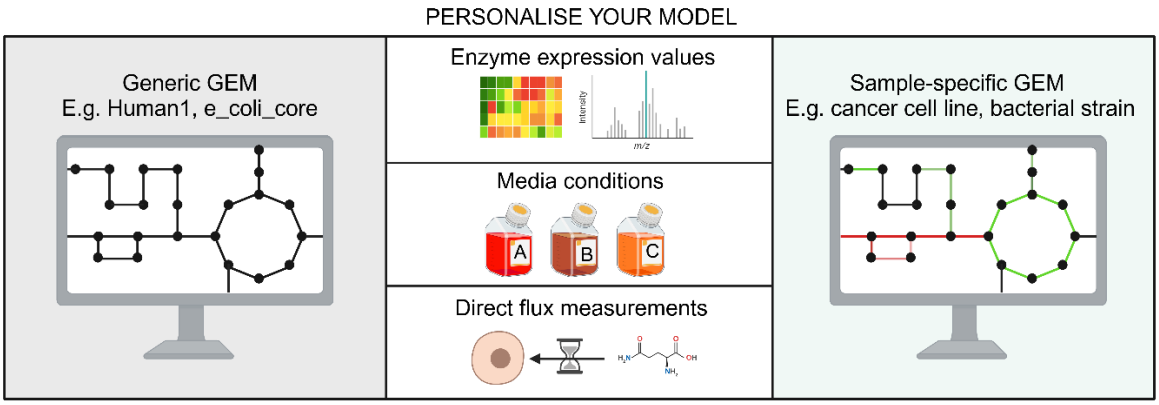

Figure 5. The features of a GEM which may be personalised.

c. Beginning your flux analysis

Once the model has been constrained with experimental data, you may begin flux analysis. This process involves 'solving' the GEM with all the appropriate constraints and obtaining a flux distribution of the predicted metabolic reaction rates (**Error! Reference source not found.**). The simplest way to do this is using flux balance analysis (FBA), where you must specify an 'objective function' (Orth et al., 2010). An objective function is a single (or multiple) metabolic reactions within the GEM, which are to be optimised to a maximum or minimum, whilst all other fluxes combine to work towards this optimum. For example, when modelling a bacterial strain, a popular objective function is the maximisation of cell growth because adaptive evolution suggests that this is what bacteria have evolved towards (Ss et al., 2003) . On the other hand, setting the objective function becomes more complicated for higher organisms, such as humans, which have cells specialised for a variety of metabolic functions, such as managing redox potential, secreting specific proteins or maximising ATP production, for example. If multiple metabolic reactions are specified within the objective function, a weighting can be applied to these, for example 50% growth maximisation and 50% ATP production. The objective function is represented by **Formula 1**, where 'c' refers to the weighting of the individual metabolic reaction 'v'.

Formula 1. Formulating the linear objective function

$$z = c_1 * v_1 + c_2 * v_2 \dots c_n * v_n$$

Some useful references for exploring possible objective functions, per organism:

- **E. coli:** (Schuetz et al., 2007), (Knorr et al., 2007)
- **Yeast:** (Schnitzer et al., 2022)
- **Chinese hamster ovary cells (used in industry for antibody production):** (Chen et al., 2019)**,** (Jiménez Del Val et al., 2023)
- **Human:** (Nagrath et al., 2007)

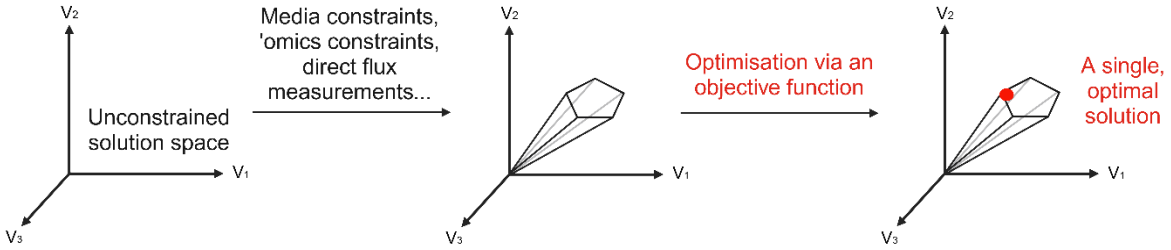

Figure 6. Setting the objective function using FBA identifies a single, optimal flux distribution. Figure adapted from (Orth et al., 2010).

More often than not, objective functions are linear, however they may also be quadratic. The reason they might be quadratic is that individual reactions in the multi-objective function might need maximising whilst others need minimising, and there might be a more complex relationship between the flux and activity of these reactions which cannot be represented linearly. This has been explored in the context of human cells to study cancer metabolism, where ATP production, lactate production and ATP yield were all specified in the objective function (Zhang & Boley, 2022). There are useful explanations of playing with the objective function provided by COBRA: https://cobrapy.readthedocs.io/en/latest/simulating.html

d. Alternative flux analysis methods

Sometimes, you might want to use a different flux analysis method than FBA. One of the reasons for this is that with each iteration, FBA can give different solutions, which could contribute to uncertainty. Or, you might want to run model simulations across different experimental conditions, and detect differences between two models, but FBA might not be sensitive enough to show any differences. In this case, you can use flux variability analysis (FVA), which outputs a range of possible solutions for a constrained model, in the form of a maximum and minimum predicted flux per reaction. FVA is more computationally intensive than FBA, but works off similar principles, whereby you specify an objective function.

A step up from FVA is flux sampling, which provides even more detailed solutions. As well as giving the user a range of possible predicted fluxes per reactions, flux sampling outputs a probability distribution across this range, which is produced through millions of iterations (Herrmann et al., 2019). This flux analysis method is suitable for users who want to simulate a population of solutions to find the difference between one pool and another, for example the flux distributions of the low and high growth solutions. Furthermore, flux sampling does not require the user to specify an objective function, so avoids this assumption. Due to its computational demand, flux sampling can be run on a high-performance computing node,

such as the computational shared facility at the University of Manchester. These different flux analysis methods have been compared in ***Figure 7***.

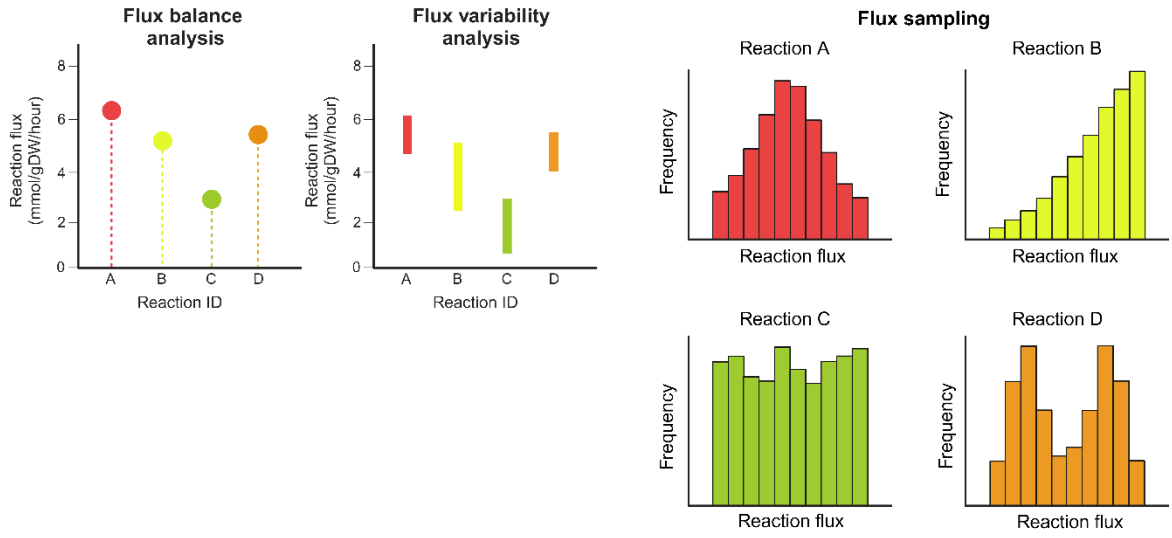

Figure 7. Visualising flux distributions with FBA, FVA and flux sampling.

e. How to simulate and predict with your model

Once you have constrained your GEM to be sample-specific, and chosen an appropriate flux analysis method, you may wonder what to do with your model! This, in my opinion, is the most fun bit of the project. This is where you can simulate phenotypic and environmental changes to your model, to predict or explain experimental results. Your model may have been built based on a cell culture experiment, with specific media conditions and gene expression. And for your project, you might like to predict targets for an siRNA experiment to reduce cellular of a cancer cell line or generate a new media formulation to improve growth and immortality. Your sample-specific GEM is a framework for these simulations, a lot of which are described in COBRA tutorials: https://cobrapy.readthedocs.io/en/latest/deletions.html. Here, you can simulate the deletion of a single gene, or reaction, or double gene/reaction if you are interested in synthetic lethality, for example. Gene deletion is simulated by restricting the lower and upper bounds of reactions the enzyme encoded catalyses. You can save deletion results and compare to a wild-type control to predict phenotypic characteristics of the knockout cell.

As well as gene engineering, you can perform media optimisation tasks with your model. You may make changes to the initial media you specified, be this glucose concentrations or the oxygen availability of complexity of the media, and predict how this may change the cell growth or stress response. This approach is interesting for both industry-partnered projects who might want to develop a new, marketable cell culture medium, or an academic lab who are interested in metabolic pathways which might be perturbed in response to a new feed.

f. How to validate and understand your results

The result of flux analysis on your constraint-based model is a large dataframe, with flux predictions for each reaction given those specific constraints. This dataframe could contain one column of single flux solutions for FBA, or two columns of a minimum and maximum for FVA. When I first learnt modelling, I thought getting to this dataframe was the hard part, but actually for me it was more challenging to interpret the results! But with some experimental

validation to give a sanity check, and some statistical interpretation, you should be able to make some really interesting predictions from your modelling.

i. Experimental validation

Something that's worth mentioning and remembering is that it is very important to experimentally validate model predictions. This doesn't mean you have to go into the lab yourself, it might be that you can find a matching dataset for the expression data you have used to constrain the model, which describes metabolite uptake rates, secretions or cell growth (see ***Table 1*** for resources). This is important because it means that before you start simulating changes to the cell using your model, you know that it is properly predicting the wild-type behaviour. It also means experimentalists are more likely to be on-board with your modelling! Experimental validation could be as simple as comparing model predicted cell growth, in the form of doubling times, to experimental measurements. Or you could predict the effect of a gene knockout if you have the data for this. Other validations can include direct flux measurements if you have access to these, such as lactate or acetate production, or glucose uptake, for example. It is worth reading the literature to see the variety of ways in which people validate their results and decide what approach will work best for you.

Table 1. Resources for multi-omics datasets and experimental validation.

Resource	Information
Cancer Cell Line Encyclopaedia https://sites.broadinstitute.org/ccle/datasets	Human Gene and protein expression data and matched datasets such as CRISPR-Cas9 results and growth measurements for a large collection of cancer cell lines
Human Protein Atlas https://www.proteinatlas.org/about/download	Human Gene and protein expression and prognostic datasets, such as for cancer or blood disorders
Human Microbiome Project https://portal.hmpdacc.org/	Bacteria Multi-omics and metadata for human microbiota
NCBI Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geoprofiles/	Multiple organisms Public repository of genomic and gene expression measurements for a variety of organisms
Other databases…	Figshare https://figshare.com/category is a common database for depositing data upon publication to promote FAIR data You can also find datasets in the supplementary of papers

ii. Interpreting model results

Once you have the model output, you want to be able to interpret the flux predictions and identify key differences between models. This can be according to the portion of model predicted to be active (carry a non-zero flux) between two conditions, the activity through subsystems, i.e. central carbon metabolism, fatty acid metabolism, or it could be through specific reactions, such as the transport of particular amino acids. A difference in model predictions may come in the form of a reaction being switched off or on, an increase or decrease in flux or a change in direction of a reversible reaction.

But, how do we find the most important differences between two models, when we can have thousands of reactions in a GEM? This information has been summarised in ***Figure 8***. Although you can simply report a difference between two predicted fluxes, this can be quite arbitrary as a flux of 0.0001 mmol/gDW/hour and 10 mmol/gDW/hour may be as meaningful as one another between two reactions, depending on the reaction itself. Given the predicted fluxes of reactions can be Log scales different, one approach for interpretation is to report the percentage change in flux from the original between two models. Alternatively, you could transform the flux values to something similar to a Z-score, or you could square-root the flux values, whilst keeping the negative or positive symbol so information on reversibility is not lost. In this way, when the fluxes have been normalised, they can be visualised alongside one another on a heatmap or similar plot.

If multiple models have been constrained, it could be helpful to use a dimensionality reduction technique, such as Principal Components Analysis (PCA), and observe the distribution of models according to predicted reaction fluxes. Here, if you then apply a clustering algorithm such as K-means, you could identify similarities between the overall flux distributions of multiple models and draw meaning.

If you are fortunate enough to have a rich dataset including multiple replicates of a particular condition compared to a control sample, you could use a statistical test such as a t-test or the Mann-Whitney U test. Guidance for selecting an appropriate statistical test can be found here: https://www.graphpad.com/support/faqid/1790/ and in the GraphPad Prism software itself.

To analyse the model output visually, network visualisation softwares exist which allow you to plot a two-dimensional image of your GEM, and even overlay some flux data on top. This can make nice figures for publications or reports. Softwares for this visualisation include the Escher package (https://escher.github.io/) (Python), Cystoscape (https://cytoscape.org/) (downloadable software) and Newt Editor (https://newteditor.org/) (online interface).

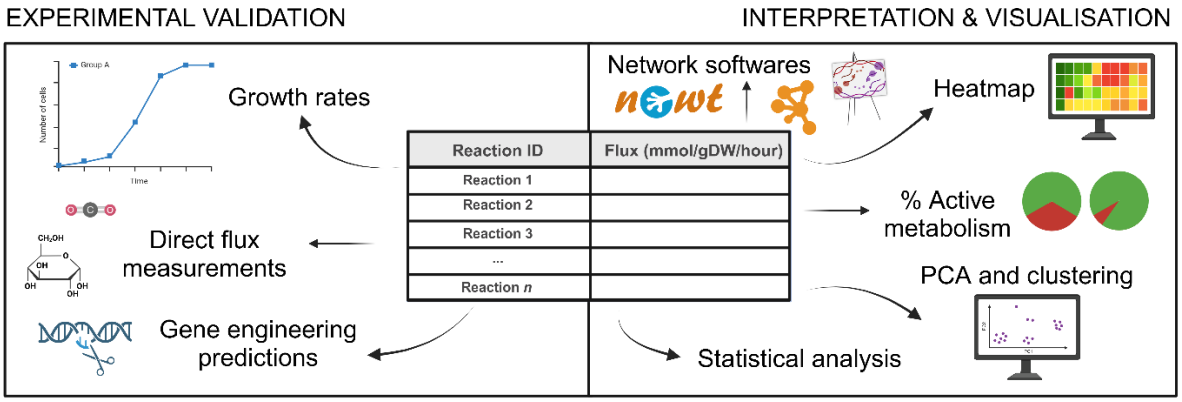

Figure 8. Validation and interpretation of metabolic models.

3. Important notes

FBA although the most efficient flux analysis method, holds various assumptions:

- **The steady state assumption:** the rate of production and consumption of a particular metabolite are equal, so that there is no net accumulation of this metabolite in the system
- **Objective function:** the correct choice of objective function is much debated and must be considered on the basis of your experimental context and the organism you are representing. Adaptive evolution influences the use of biomass optimisation for bacterial FBA, but this can be more complicated with higher organisms and should be justified
- **Cannot predict metabolite concentrations** and kinetic parameters are not routinely integrated, therefore the model will hold **uncertainty**
- **Only produces single solutions:** the flux distribution solved using FBA could change each iteration, as multiple possible distributions might achieve the same objective function

If you don't for some reason get along with the COBRA toolbox for flux analysis, there are alternative packages available, for example MEWpy (https://mewpy.readthedocs.io/en/latest/).

Model outputs predicted fluxes as mmol/gDW/hour for all reactions except growth rate, which is g/gDW/hour. Growth rate can be converted to doubling time by taking the inverse (1/growth rate). Flux values can be converted to experimental units if the dry cell weight is known.

I made this help sheet in December of 2024, so there could have been new softwares and versions released since then! Always check Github and most up-to-date documentation.

Different solvers can give different results, I use Gurobi, which is a linear solver, but there are others, e.g. CPLEX, GLPK… - these can all give different results!

4. References

Becker, S. A., & Palsson, B. O. (2008). Context-Specific Metabolic Networks Are Consistent with Experiments. *PLoS Computational Biology*, *4*(5), e1000082. https://doi.org/10.1371/journal.pcbi.1000082

Chen, Y., McConnell, B. O., Gayatri Dhara, V., Mukesh Naik, H., Li, C.-T., Antoniewicz, M. R., & Betenbaugh, M. J. (2019). An unconventional uptake rate objective function approach enhances applicability of genome-scale models for mammalian cells. *NPJ Systems Biology and Applications*, *5*, 25. https://doi.org/10.1038/s41540-019-0103-6

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, *7*(1), 74. https://doi.org/10.1186/1752-0509-7-74

Herrmann, H. A., Dyson, B. C., Vass, L., Johnson, G. N., & Schwartz, J.-M. (2019). Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *Npj Systems Biology and Applications*, *5*(1), 32. https://doi.org/10.1038/s41540-019-0109-0

Jiménez Del Val, I., Kyriakopoulos, S., Albrecht, S., Stockmann, H., Rudd, P. M., Polizzi, K. M., & Kontoravdi, C. (2023). CHOmpact: A reduced metabolic model of Chinese hamster ovary cells with enhanced interpretability. *Biotechnology and Bioengineering*, *120*(9), 2479–2493. https://doi.org/10.1002/bit.28459

Knorr, A. L., Jain, R., & Srivastava, R. (2007). Bayesian-based selection of metabolic objective functions. *Bioinformatics*, *23*(3), 351–357. https://doi.org/10.1093/bioinformatics/btl619

Lewis, J. E., Forshaw, T. E., Boothman, D. A., Furdui, C. M., & Kemp, M. L. (2021). Personalized Genome-Scale Metabolic Models Identify Targets of Redox Metabolism in Radiation-Resistant Tumors. *Cell Systems*, *12*(1), 68-81.e11. https://doi.org/10.1016/j.cels.2020.12.001

Meeson, K. E., & Schwartz, J.-M. (2024). Constraint-based modelling predicts metabolic signatures of low and high-grade serous ovarian cancer. *NPJ Systems Biology and Applications*, *10*(1), 96. https://doi.org/10.1038/s41540-024-00418-5

Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A., & Yarmush, M. L. (2007). Integrated Energy and Flux Balance Based Multiobjective Framework for Large-Scale Metabolic Networks. *Annals of Biomedical Engineering*, *35*(6), 863–885. https://doi.org/10.1007/s10439-007-9283-0

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., Anton, M., Ferreira, R., Domenzain, I., Billa, V., Limeta, A., Hedin, A., Gustafsson, J., Kerkhoven, E. J., Svensson, L. T., Palsson, B. O., Mardinoglu, A., Hansson, L., Uhlén, M., & Nielsen, J. (2020). An atlas of human metabolism. *Science Signaling*, *13*(624), eaaz1482. https://doi.org/10.1126/scisignal.aaz1482

Schnitzer, B., Österberg, L., & Cvijovic, M. (2022). The choice of the objective function in flux balance analysis is crucial for predicting replicative lifespans in yeast. *PLoS ONE*, *17*(10), e0276112. https://doi.org/10.1371/journal.pone.0276112

Schuetz, R., Kuepfer, L., & Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular Systems Biology*, *3*(1), 119. https://doi.org/10.1038/msb4100162

Ss, F., Jy, M., & Bø, P. (2003). Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. *Journal of Bacteriology*, *185*(21). https://doi.org/10.1128/JB.185.21.6400-6408.2003

Strain, B., Morrissey, J., Antonakoudis, A., & Kontoravdi, C. (2023). How reliable are Chinese hamster ovary (CHO) cell genome-scale metabolic models? *Biotechnology and Bioengineering*, *120*(9), 2460–2478. https://doi.org/10.1002/bit.28366

Timouma, S., Balarezo-Cisneros, L. N., Schwartz, J.-M., & Delneri, D. (2024). Development of a genome-scale metabolic model for the lager hybrid yeast S. pastorianus to

understand the evolution of metabolic pathways in industrial settings. *mSystems*, *9*(6), e0042924. https://doi.org/10.1128/msystems.00429-24

Zhang, Y., & Boley, D. (2022). Nonlinear multi-objective flux balance analysis of the Warburg Effect. *Journal of Theoretical Biology*, *550*, 111223. https://doi.org/10.1016/j.jtbi.2022.111223

Zur, H., Ruppin, E., & Shlomi, T. (2010). iMAT: An integrative metabolic analysis tool. *Bioinformatics (Oxford, England)*, *26*(24), 3140–3142. https://doi.org/10.1093/bioinformatics/btq602