ITERATIVE MULTI LEVEL CALIBRATION OF METABOLIC NETWORKS

MAX CONWAY [1], CLAUDIO ANGIONE [2], AND PIETRO LIÒ [1]

Abstract. Frameworks for metabolic engineering have been successfully applied in combination with pre- and post-processing algorithms on genome-wide metabolic models. However, genetic engineering methods with a particular focus on understanding results from multiple perspectives and combining automated and human design are still lacking. To this end, we adopt a multi-objective genetic design technique to find the optimal gene expression levels in genome-scale metabolic reconstructions. Then, we analyse the optimized network by introducing a new multi-omic, multi-level post-processing and visualization procedure, Metabex, which uses Cytoscape for network visualization. These two components are connected together to form a feedback loop that establishes a continual process of machine optimization and human analysis and guidance. To benchmark our framework, we optimize two species of *Geobacter* for electricity production and biomass synthesis; we achieve increases in electricity production for only a slight decrease in biomass. Many regulatory strategies contributed to this value, locally and globally. For instance, a direct, local strategy was a down-regulation of Cytochrome C Oxidase, while there was simultaneously a global reduction in cofactor and prosthetic group biosynthesis. Finally, we discuss multiple applications of our tool, including model exploration, model engineering, comparative modelling, meta-analysis and model refinement.

1. Introduction

Metabolic engineering is one of the most industrially important areas of genetic design [1]. It is commonly achieved by the combination of a heuristic search algorithm, to generate draft genomes, and Flux Balance Analysis (FBA) to evaluate them. Post-processing algorithms and toolboxes (e.g. Cobra [2] and Sybil [3]) are then required to find the most useful features of the draft genomes. This paper presents a framework that consists of a continuous, multi-objective search and evaluation algorithm, expFBA, integrated with a new post-processing procedure, which uses Cytoscape[4] for network exploration, thus providing a multi-omic view of an optimized metabolic network. A key feature of our method is that the results of the post-processing step feed back into the optimal search procedure to guide the design process.

Figure 1 shows a flowchart of our framework. The input is a metabolic model in the form of a gene-protein-reaction table, such as those downloadable from

(1) Computer Laboratory, University of Cambridge
(2) School of Computing, Teesside University

Key words and phrases. Meta-analysis, Gene expression engineering, Multi-scale model, Metabolic network exploration.

EPSRC Engineering and Physical Sciences Research Council Mission-T2D.

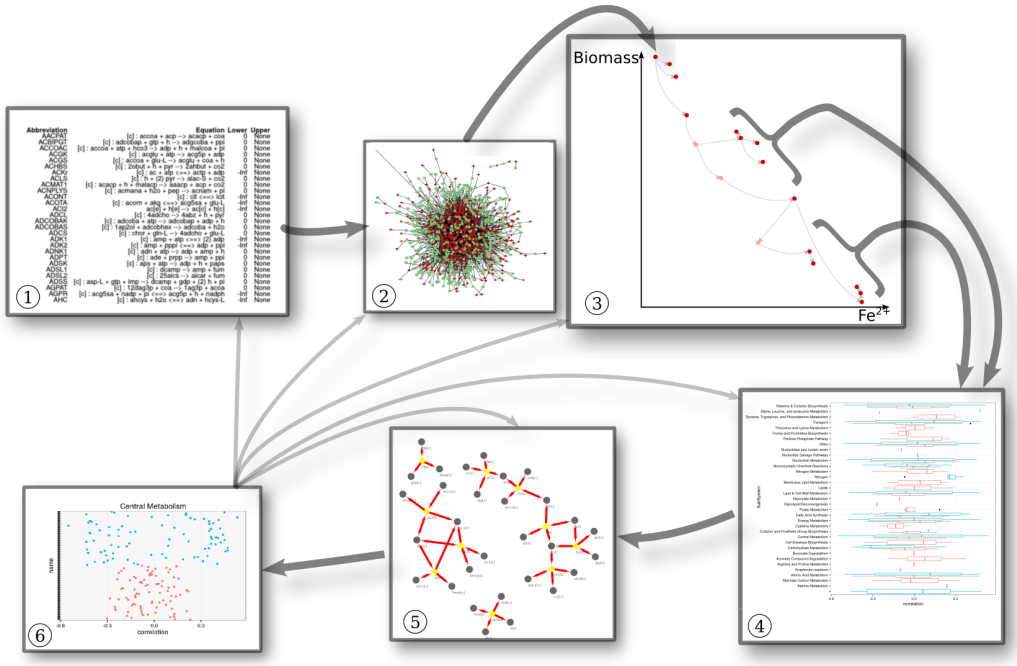


FIGURE 1. Integrative Hypothesis testing multi-scale Software Platform for evaluating metabolic multi-omic models. (1) Reaction table. (2) Whole cell metabolic model. (3) Pareto front. (4) Statistical analysis of the Pareto front. Here we illustrate a comparison of aggregates over two regions of the front, but any number of other approaches are possible. (5) A subnetwork, such as a pathway, or a set of reactions identified as important by whole network level statistics. (6) Further analysis of the subnetwork. This can be statistical or via expert knowledge based examination.

BiGG [5]. This model table forms the input to expFBA, which returns a population table with rows representing individuals, and columns representing gene expressions levels and phenotype fluxes. The metabolic model table and population table are loaded by Metabex, which can then connect to Cytoscape and render the metabolic network. Expression vectors, or aggregates thereof, can be transferred from the population table to provide an overlay in Cytoscape. From here, analyses are more free-form, and a number are discussed in the case study, but the general flow is a feedback loop of some statistical technique producing a subset or aggregate of the population table in R, then transferring the results to Cytoscape. These results can then be viewed and manipulated further, and passed back to R, completing the loop.

Our framework (Figure 1) creates several advantages over the existing methods for combined network optimization and visualization. First, expFBA gives continuous reaction regulation rather than Boolean on/off genetic strategies; this creates a smoother fitness landscape, and hence faster convergence and detailed information about the relative effects of reactions. The multi-objective optimization algorithm

produces a Pareto front, i.e. a set of solutions with varying properties; this helps untangle the exact effects of different reactions and genes, when combined with our analytic procedure [6]. Furthermore, we facilitate viewing the metabolic network and its underlying regulatory structures from different omic perspectives and at different scales, which allows a comprehensive understanding to be built that is difficult to develop if the different layers are not considered together. We use Cytoscape for network visualization, which provides a familiar graphical environment. Finally, built in feedback and integration between automated optimization and human expert knowledge gives faster, more useful results.

As a case study for our framework, we compare two species of *Geobacter* (Section 3). Specifically, we design and optimize the metabolic network towards biomass production and Fe^{2+} excretion, which is a proxy for electricity production. As a result, we find a variety of optimal solutions, with electricity production increases of up to 206 %, while sacrificing only 4 % of biomass generation capacity. Our visualization tools give insight into how this result is achieved, and allows the comparison between the different Pareto fronts regarding the electricity production of the two species.

We finally describe several applications of our framework (Section 4), from straightforward model exploration to meta-analysis and model refinement. These applications include comparison between models, in order to better understand the individual models. This step enables the generation of new models that better approximate a wild-type, or present some property not seen in the wild-type. This is important because, as more metabolic models are created, cross-comparison becomes key to understand how they relate to each other, as well as their properties in isolation.

2. Components of the Metabolic Engineering Framework

Here we discuss in detail the three separate components of this metabolic design framework:

(1) expFBA, a heuristic search and optimization procedure. This performs multi-objective optimization of metabolic models using a continuous interpretation of the gene–protein–reaction mapping;
(2) a post-processing procedure, which provides a multi-level view of the Pareto front;
(3) the expert knowledge integration feature, whereby expFBA can be guided towards searching areas that are known, or have been found, to be particularly relevant.

2.1. Gene Expression Based Search and Evaluation Algorithm.

Genetic modifications in metabolic engineering are often simulated by reaction knockouts. This approach has been successful, but has a number of limitations: firstly, it provides no information about the effects of essential reactions, since there will be no individuals missing these; secondly, the fluxes possible are quantized, leaving areas of the Pareto front unavoidably empty; and thirdly, optimization of a knockout vector is difficult, since the majority of knockouts will either be lethal or have no effect on the phenotype.

To overcome the limitations of Boolean knock-out approaches, we link gene expression level to an FBA model (background box 1), thus enabling a multi-objective

BACKGROUND 1. Flux Balance Analysis

Flux balance analysis (FBA) is a technique for simulation of metabolic networks. It has three prerequisites:

- the assumption that the metabolic network is in a steady state,
- the assumption that the cell will always regulation the metabolic network to produce the maximum biomass, and
- a list of reactions in the metabolome, together with their reversibility and bounds on their rates.

Using the reaction list, we can establish the topography of the metabolic network. Then, using the steady state assumption, we can introduce constraints on the rates of reactions so that consumption and production of every metabolite is equal. We can further constrain the resulting set of simultaneous equations by the explicit reaction rate bounds. Finally, we can use linear programming to solve the simultaneous equations, optimizing for the solution with the highest biomass output. (For more details, see [7].)

search for the optimal gene expression values to maximize or minimize chemicals of interest in a constraint-based model. We term this approach 'expFBA'.

Each gene expression profile provided by the compendium (normalized as fold change from wild-type), is mapped to bounds for the flux rates in the FBA model. This allows us to avoid expression thresholds to determine if a reaction is active or not, and is supported by the increasing evidence that the metabolic reaction activity correlates with the mRNA levels [8]. The *Geobacter* model is run with the new bounds, thus yielding two output values for the two objectives chosen. Flux balance analysis is performed in a bilevel fashion, giving preference to the inner maximisation of the biomass (natural objective), and then to the outer maximisation or minimisation of the second objective (synthetic) [9].

In order to map the gene expression profiles provided by the multi-objective algorithm onto the Geobacter models, we associate every profile with a specific configuration of reaction bounds in the model. Our model is based on the assumption that the lower and upper bounds of the ith flux v_i depend on the expression of the gene set controlling the ith reaction. A gene set is made up of one or multiple genes, and therefore a map from gene expression to gene set expression has to take into account how each gene set is defined, and specifically the AND/OR Boolean relation between its genes. If x_1 and x_2 are the gene expression levels of the two genes s_1 and s_2, and if $s_1 \wedge s_2$ and $s_1 \vee s_2$ are two basic gene sets, we define the *gene set expression level* using the map τ defined as:

$$s_1 \wedge s_2 \xrightarrow{\tau} \min\{x_1, x_2\} \quad \text{(enzyme complex)}, \tag{1}$$

$$s_1 \vee s_2 \xrightarrow{\tau} \max\{x_1, x_2\} \quad \text{(isozymes)}. \tag{2}$$

The bounds of a reaction catalyzed by an enzyme complex s_1 AND s_2 will be a function of $\min\{x_1, x_2\}$, while the bounds of a reaction catalyzed by s_3 OR s_4 will be a function of $\max\{x_3, x_4\}$. The gene set expression levels of a complex gene set is defined with the same approach, i.e. applying 1 and 2 recursively.

Finally, after choosing a bidimensinal objective space (e.g. Fe^{2+}–biomass), we solve the following bilevel problem:

$$
\begin{array}{lll}
\max & g^{\intercal}v & \\
\text{such that} & \max & f^{\intercal}v \\
 & \text{such that} & Sv = 0 \\
 & & V^{min}\phi(y) \leq v \leq V^{max}\phi(y)
\end{array}
\tag{3}
$$

where f and g are the arrays used to set the weights associated with the objectives, while y is the gene expression profile. In a two-objective problemwe set f_{j_1} and g_{j_2} to 1 if we want to maximise the flux rate of the natural objective v_{j_1} and the synthetic objective v_{j_2}, thus running the optimization in the v_{j_1}/v_{j_2} objective space. The function ϕ is defined as:

$$\phi(y_i) = [1 + |log(y_i)|]^d, \tag{4}$$

where $d = (y_i - 1)/\left|y_i - 1\right|$, and $\phi(1) = 1$.

The use of gene expression data in metabolic models may be particularlyuseful when comparing a normal and a pathogenic strain optimized towards two or more objectives. More specifically, our approach would allow one to understand how the extra pathogenic genes are behaving in different conditions.

Compared to traditional approach of conducting boolean knockouts before flux balance analysis, expFBA's continuous approach is capable of reaching a wider range of phenotypes (specifically, uncountably infinite phenotypes for expFBA, rather than a finite, if huge, range, for boolean knockouts). There is no significant computational performance difference between boolean knockouts and expFBA, since knockouts or knockdowns are conducted in $\mathcal{O}(n)$ time, versus the polynomial average time complexity of the simplex algorithm used in the actual FBA procedure itself [10].

Previous gene expression level integration approaches have been attempted, such as in [11, 12, 13], but expFBA is uses fully continuous expression values, unlike the thresholding and discretization in [13] and [11] respectively, and moreover continuous gene expression data provides unique advantages when combined with multi-objective optimization.

2.2. **Exploration Tools Utilizing Cytoscape plugin: Metabex.** expFBA produces a Pareto front. i.e. a population containing the individuals that represent the best possible trade-off between two or more objectives. Each individual consists of an equally complex, but slightly different, metabolic network. This huge amount of information is beyond direct human interpretation, and sophisticated tools are required. We approach this by viewing the whole dataset at different scales to build up an understanding of the metabolome in the context of its evolutionary landscape.

In the visualization of the metabolic network, nodes and edges inevitably cover each other up in 2-dimensional projections of all but the smallest networks. Because of this, a high quality visualization tool is required to provide smooth interactivity, so that network data can be properly understood. The Cytoscape network visualization package was used as a base for the network viewing facility in this framework, since it is freely available for all major platforms and it has a number of

useful plugins. Cytoscape was connected to R via the RCytoscape R package [14], XML-RPC, and the Cytoscape plugin CytoscapeRPC [15].

We created an R package, named Metabex, to wrap up and integrate interaction with Cytoscape, with Pareto fronts created in Matlab using expFBA and with models from Sybil. This enables us to reduce the complexity of interacting with these various systems. For instance, displaying a Sybil model as a metabolic network in Cytoscape becomes just one command, as does using the R.Matlab package to import a Matlab Pareto front in a format suitable for display on a Cytoscape network diagram. This package is available at `github.com/maxconway/metabex`.

2.2.1. *Species Level Comparative Metabolomics.* The highest level analysis conducted in this framework is to compare related species, as illustrated in figure 2.This can enable us to investigate the aspects where their metabolomes differ, and predict the reason according to the type of difference observed. Two related species will have sets of reactions S and T; from these we can derive three reaction sets:

$$\begin{aligned} A &:= S \cap T, \\ B &:= S \setminus T, \\ C &:= T \setminus S. \end{aligned} \tag{5}$$

B and C are typically small, and easily interpreted by simple examination of reaction lists, especially since these auxiliary reactions often serve narrower purposes than the core reactions.

More interesting is to understand the effects of the presence of B or C on A. In order to see real differences in A, rather than just random variation, we must average over multiple individuals; however we only use a small section of the Pareto front, rather than the whole population, so that we are aggregating similar phenotypes and do not remove the signal along with the noise.

2.2.2. *Population Scale Aggregate Genomics.* The shape of the Pareto front can provide valuable information about what phenotypes are likely to be well adapted under particular circumstances. For instance, knee points indicate phenotypes that are likely to be successful across a wide range of conditions, while sparser areas of the front show conditions that the organism has difficulty adapting to. High level features can provide very useful information about what phenotypes are possible, but understanding the genetic and proteomic causes of these features is a difficult endeavour, for two main reasons. First, a large amount of noise is likely to hide features. Second, confounding vestigial traits can exist, where a particular genotype is correlated with a successful phenotype due simply to being present in a successful common ancestor, without having actually contributed to the success of that ancestor.

Our idea to evaluate a phenotype adaptation is to establish a measure of phenotypic similarity to the wild-type, and examine the correlation between this measure and gene expression level. In the two dimensional case, establishing a measure of phenotypic similarity amounts to fitting a curve to the Pareto front, and then using distance along this curve as the similarity measure. In our *Geobacter* case study, the front was approximately straight (see fig. 6), so we were able to use principle components analysis to fit a straight line, but in other cases more sophisticated dimensional reduction techniques may be appropriate.

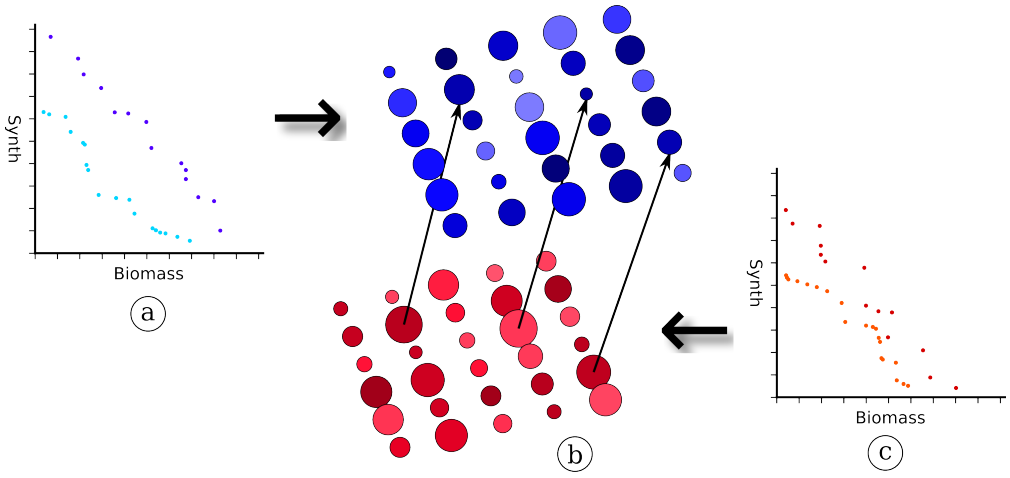


FIGURE 2. An illustration of model comparison. We can compare the pathways in *G. Sulfurreducens* (blue, ⓐ) and *G. Metallireducens* (red, ⓒ) pairwise, across the whole Pareto front. In ⓑ, circles represent pathways, and their sizes represent pathway sensitivity. In the graphs, the lighter coloured dots indicate metabolic positions accessible via boolean knockouts, versus via a transcriptomic approach. We can compare models at different levels of granularity depending upon what is most appropriate. For instance, for we can look at pathway sensitivity to identify what pathways are similar between models, and compare them at the more detailed reaction level to find subtle differences.

The Pearson product-moment correlation coefficients between the activities of each gene and the position along the line approximating the Pareto front were measured. They were found to be approximately normally distributed due to random genetic drift, but crucially there were some outlying genes, with anomalously high or low correlations with phenotype. Furthermore, these correlations could be aggregated across subsystems, to find the areas or cellular activity that had been strongly up or down-regulated. Figure 7 shows a subnetwork resulting from this procedure, while section 3.2 discusses how this figure could imply that these regulatory strategies work by creating an oxygen deficit, or by downregulating Fe^{2+} oxidation.

2.2.3. *Individual Scale Genomics and Metabolomics.* Some Pareto fronts are smooth, while others are rough and contain discontinuities. Where these discontinuities exist, they present an interesting comment on the evolvability of the trade-off in question, but can also provide information on the pathways involved, since a discontinuity will be likely to be the site of a qualitative change in the reactions involved.

To enable fluid exploration of the changes in gene expression across discontinuities, we built tools to enable individuals on the Pareto front to be selected graphically and their reaction network examined. With traditional FBA, this would still leave the daunting task of finding the important differences among many active

SubSystem	*G. Metallireducens*			*G. Sulfurreducens*		
	mean	σ	p-value	mean	σ	p-value
Alanine Metabolism	0.78	0.44	8.05E-128			
Alternate Carbon Metabolism	1.07	0.53	1.64E-105			
Amino Acid Metabolism	1.00	0.59	1.29E-03	1.02	0.59	3.24E-58
Anaplerotic reactions	1.59	0.34	0.00E+00			
Arginine and Proline Metabolism	1.23	0.44	3.53E-144			
Aromatic Compound degradation	1.07	0.62	1.71E-279			
Benzoate Degradation	0.85	0.31	1.17E-117			
Carbohydrate Metabolism	0.98	0.52	4.75E-25	0.88	0.66	1.14E-175
Cell Envelope Biosynthesis	1.07	0.54	1.08E-124			
Central Metabolism	1.01	0.57	2.63E-28	1.00	0.60	1.71E-01
Cofactor and Prosthetic Group Biosynthesis	1.00	0.62	7.19E-01	0.96	0.42	1.37E-06
Cysteine Metabolism	0.70	0.57	9.70E-283			
Energy Metabolism	1.04	0.53	4.52E-165	1.00	0.57	1.25E-02
Fatty Acid Synthesis	1.06	0.57	2.43E-95	1.05	0.64	4.52E-47
Folate Metabolism	0.92	0.70	2.81E-31			
Glycolysis/Gluconeogenesis	0.52	0.43	0.00E+00			
Glyoxylate Metabolism	1.52	0.39	0.00E+00			
Lipid and Cell Wall Metabolism	1.02	0.59	3.51E-67	1.03	0.60	1.40E-96
Nitrogen Metabolism	1.00	0.55	9.39E-01	1.17	0.54	0.00E+00
Nonenzymatic Chemical Reactions	1.15	0.70	1.49E-82	0.87	0.57	2.10E-66
Nucleotide Metabolism	0.99	0.60	8.59E-04	1.03	0.61	2.81E-69
Nucleotide Salvage Pathways	0.89	0.26	3.25E-99			
Nucleotides and nucleic acids	0.71	0.22	0.00E+00			
Other	1.00	0.60	4.00E-01	1.00	0.60	2.92E-01
Pentose Phosphate Pathway	0.93	0.61	1.04E-26			
Purine and Pyrimidine Biosynthesis	0.86	0.74	4.26E-64			
Threonine and Lysine Metabolism	0.89	0.70	1.30E-28			
Transport	1.05	0.62	0.00E+00	1.04	0.59	1.24E-279
Tyrosine, Tryptophan, and Phenylalanine Metabolism	0.87	0.58	9.37E-116			
Valine, Leucine, and Isoleucine Metabolism	1.21	0.37	3.09E-156	0.51	0.22	0.00E+00
Vitamins & Cofactor Biosynthesis	0.98	0.59	1.36E-69	1.02	0.61	7.73E-56

TABLE 1. Subsystem regulation table. Subsystems are those supplied in the original model files (see Section 3).Columns show the mean and standard deviation of regulation in the subsystem, and the p-value for this regulation being significantly different from Gaussian genetic drift, as returned by Student's t-test.Subsystem regulation is also shown graphically in Figure 4.

reactions, but expFBA gives a continuous value for reaction activity, so that filters can be used with an arbitrary degree of specificity to home in on the most highly selected areas of the metabolic network, such as demonstrated in Figure 7. The tools we built are general enough to allow examination of other, related properties of genes or reactions.

2.2.4. *Pathway Scale Genomics and Metabolomics.* Individual subsystems and pathways can be approached either in a bottom up manner, where reaction or gene properties are aggregated by their associated subsystem or pathway, or in a top down manner, where a subset of the metabolic network is selected for more in depth analysis. A particularly effective approach is to use these techniques in sequence—first aggregated metrics can be used to identify which pathways or subsystems deserve further analysis, and then these components can then be selected and examined. Table 1 and figure 4 show pathway level aggregation of the genomic data from the geobacter case study. We can see that for sufficiently large pathways, very small p-values are achievable, although in some cases the actual magnitude of mean regulation is small.

2.3. **Expert Knowledge Integration.** As the evolutionary algorithm progresses,it can become apparent that mutations in some areas of the metabolome produce more interesting effects than others. A typical example is that the vast majority of reactions will be involved in biomass synthesis, but have nothing to do with the

other synthesis target, so down-regulation will reduce biomass production, while up-regulation will have no effect (and in a live cell would waste resources on excess enzymes).

These sorts of effects can be very easy to see once the Pareto front has started taking shape, assuming that one has a reasonable knowledge of general biological principles. However, it can be hard to recognise these effects algorithmically, and very difficult to codify the human knowledge that can help to identify why many mutations are not interesting for a particular optimization problem [16]. For this reason, we implemented a supervised exploration systemwhereby reactions or sets of reactions can be selected in the Cytoscape metabolic network, and designated for focussed exploration by the evolutionary algorithm, as it runs. This positive feedback loop allows the optimization algorithm to benefit from human knowledge and guidance, at the same time as it delivers hints to update that guidance.

In an attempt to automate the procedure of gene selection, we trained a linear model and a random forest [17, 18] to predict phenotype values from genotype values. The prediction was good, with a mean of squared residuals of 4×10^{-5}, 99 % of variance explained, prediction vs truth correlation of 1. This performance was not surprising given the derivation of the dataset, however much more interestingly the 'importance' (specifically, mean decrease in Gini index) statistics from the random forest had a heavy positive skew of 12, suggesting that very few genes supplied the majority of the predictive power. We tested this and found that only 14 could indeed provide a prediction that was almost as good as the whole set, however when we looked at the reactions they controlled, no pattern was obvious. We have not yet pursued this avenue further, since it is out of scope for this paper.

3. Geobacter Comparison Case Study

Here we discuss in detail the application of the framework to comparing metabolic models of two species of *Geobacter*, namely *G. Sulfurreducens* and *G. Metallireducens*. *Geobacter* [19] is a genus of anaerobic proteobacteria with a number of possible industrial applications. Many of these applications stem from its ability to utilize insoluble materials as electron acceptors, via conductive surface pilli [20]. This ability makes *Geobacter*interesting as a candidate for use in bacterial fuel cells, and the pilli used to transport electrons have potential applications in themselves, as nanowires. Here we optimize for biomass and Fe^{2+} synthesis, as a proxy for electricity generation.

Both metabolic models were constructed with import and export constraints such that their growth was limited by the availability of acetate as an energy source. The models for *G. Sulfurreducens* and *G. Metallireducens* are described in [21] and [22], respectively, and the model files are available in the supplementary materials of these papers.expFBA was able to increase Fe^{2+} excretion in *Geobacter Sulfurreducens* by a maximum of 206 %, while keeping biomass production at 96 % of the wild-type value. *Geobacter Metallireducens* was able to achieve a more modest result, with excretion increased by a maximum of 35 % of the wild-type value, with a biomass production of 98 % of wild-type. These results are compared to results from Sybil in Figure 5.

3.1. **Global View on Synthesis.**

Figure 6 shows the Pareto fronts of both species on the same axes. We can see that while *G. Sulfurreducens* is capable of much higher Fe^{2+} excretion, *G. Metallireducens* has a much higher biomass synthesis rate. In

fact, the strain of *G. Metallireducens* with the highest Fe^{2+} excretion rate still produces more biomass than *G. Sulfurreducens* wild-type.

The whole Pareto front view of Figure 6 tells us about what the two species of *Geobacter* are capable of, but it tells us little about how they achieve this, or why their Pareto fronts differ. To this end, Figure 4 represents a view of the expression levels of genes, aggregated by subsystem. The box and whisker plots show us those pathways that are most important for core biomass production, which are zero-centered,; and some of those which are not involved, and hence have been downregulated to provide more energy for Fe^{2+} excretion, such as Cysteine Metabolism.

3.2. **Local View on Synthesis.** Applying the outlying gene detection approach outlined in Figure 3 to *G. Metallireducens* allows producing a detailed multiplex visualization of the network (Figure 7). We can see that of the three reactions that are up-regulated, two produce cytoplasmic H^- ions, and two consume cytoplasmic water (one does both). Of the four reactions that are down-regulated, two consume cytoplasmic H^- ions and produce water, while one performs the opposite task. One possible conclusion is that this works to create an overall oxygen deficit, so that more Fe^{3+} must be reduced. However, it is important to keep in mind that in a complex network, any changes normally have many effects. For instance, while pumping hydrogen ions, Cytochrome C Oxidase (CYOO2), which is controlled by four genes (*Gmet_0249*, *Gmet_0250*, *Gmet_0251* and *Gmet_0252*), converts focytcc to ficytcc. These represent cytochrome bound Fe^{2+} and Fe^{3+} ions respectively, so down-regulating this reaction is also directly preventing Fe^{2+} loss by oxidation.

A local view can also be used for validation of expFBA, specifically in terms of whether the procedure has run long enough to converge, and which pathways are constrained by the objectives chosen. If we select a specific region of the Pareto front, we know that the fluxes within this region are all similar, and we can therefore expect to find that important reactions are regulated similarly. If they are not, we must conclude that either expFBA has not converged, or those reactions were not as important as was believed. Reactions with wildly varying regulation inside a small portion of the Pareto front are likely to be unimportant for the chosen objectives, at least in the area of the Pareto front chosen.

3.3. **Network Analysis.** While manual examination is the best technique for fully understanding small subnetworks, full metabolic models are too large to be visualized in full. At this point, network statistics can become useful to allow some degree of interpretation of larger networks. We explored which network statistics might be good indicators of how reactions would be regulated, be that direction or degree of regulation. We found that betweenness and closeness [24, 25], when aggregated by subsystem, may give an indication of the degree of regulation possible while remaining viable. However, with only two models and a handful of subsystems any effect was too small to be statistically significant, but we expect the included network measures to be useful for larger genome-wide models. Table 2 shows the betweenness and closeness values for *G. Sulfurreducens*, averaged by subsystem, with the expression values. For this particular aggregation, indegree and outdegree were found to be on the edge of significance as predictors of expression level.

subsystem	betweenness	closeness	indegree	outdegree
Amino Acid Metabolism	3379.40	9.14E-05	2.49	2.14
Carbohydrate Metabolism	3233.34	9.48E-05	2.57	2.29
Central Metabolism	4972.80	8.96E-05	2.39	2.12
Cofactor and Prosthetic Group Biosynthesis	3942.96	9.00E-05	2.00	2.00
Energy Metabolism	9022.62	9.54E-05	2.86	2.71
Fatty Acid Synthesis	4656.61	9.37E-05	2.62	2.62
Lipid & Cell Wall Metabolism	2910.96	9.08E-05	2.51	2.24
Nitrogen	8818.28	9.80E-05	5.00	5.00
Nucleotide Metabolism	2867.23	9.03E-05	2.55	2.22
Other	3045.53	8.99E-05	1.96	2.09
Putative Transporters	1995.03	9.47E-05	2.00	2.00
Transport	4308.28	9.07E-05	2.60	2.24
Valine, leucine, and isoleucine metabolism	2440.65	8.97E-05	3.25	3.00
Vitamins & Cofactor Biosynthesis	4660.62	9.26E-05	2.72	2.34
Exchange	954.22	7.86E-05	0.06	1.66
Enzymatic Chemical Reactions	7291.98	9.28E-05	2.00	1.75

TABLE 2. Network measures and expression level for *G. Sulfurreducens*, averaged by subsystem. Of particular note is the low indegree for exchange reactions, caused by their modelling as one-sided reactions.

4. Discussion

By virtue of its multi-level, iterative approach, our framework can be used in several different ways. In this section models could imply complete eukaryotic or prokaryotic cells (as in this paper), organelles ([26, 27, 28]) or even interacting groups of cells, from a multicellular organism or in interacting single celled organisms.

Model Exploration and Engineering. The simplest application is exploration of the properties of individual models. For this, we set biomass production as one objective, and choose one or more pathways of interest as secondary objectives. expFBA can then show what reactions have important effects on what objectives, even when those reactions are not part of the subsystems that they affect most heavily. Simultaneously, our visualization tool allows models to be viewed from global and local perspectives, and for human feedback to expFBA in order to provide local searches that are more thorough than would be possible with a global approach. Figure 7 is exemplar of this usage, and highlights how expression levels correlate with a synthetic objective across the Pareto front.

When the synthesis objective under study is industrially relevant, the next question is how to increase production. expFBA suggests optimized regulation vectors as part of its nature as a multi-objective optimization algorithm, but understanding their relative merits would be difficult without our visualization software. As a result, regulation strategies can be examined to understand how they work, and hence if they are biologically feasible, and to find their simplest to implement and most powerful components. Once these modifications are found, a strategy is needed to check that they will not have too many adverse effects on other areas of the network.

To do this, we can regenerate the Pareto front with the modification included, and check that is has not changed significantly.

Model Comparison and Meta-analysis. When multiple similar models are available, comparative analysis can be an effective exploration tool. Models can be directly compared in terms of their available phenotypes, but model comparison can also help to understand the individual models, since each can act as a baseline for the other. For instance, both models may be capable of exhibiting a similar phenotype, but facilitated by different reactions. expFBA has a particular advantage over knock-out based approaches when comparing closely related models, such as comparing normal and pathogenic strains. This is because by modelling threshold-free expression regulation, we can see how components which are present in both strains are differently in the pathogenic strain. Furthermore, multi-objective optimization allows us to see potential evolutionary paths to and from pathogenic strains, which is useful information in the fight against antibiotic resistance. This advantage of understanding trade-offs is one of the key benefits of multi-objective optimization when compared to multi-modal optimization, as discussed in [29].

Use of multiple models can enhance the engineering of new strains in the same way that it can enhance the understanding of existing strains. For instance, reactions can be substituted in from a related species to shore up biomass production after aggressive modifications. These sorts of changes can be achieved manually, but they can be achieved more quickly by use of expFBA: new reactions can be added to the model with initial rate bounds of zero, and then gradually introduced by the algorithm. Figure 2 shows an example of how we could compare the new strains generated, with added reactions, to the wild-type, by comparing pathways individually to identify areas where the modifications have significant effects.

Model Refinement and Multi-level Validation. Our framework can also be used in model refinement. This is achieved in much the same way as meta-analysis, but optimizing synthesis of biomass: hypothetical reactions can be inserted with rate bounds of zero, and then gradually introduced by expFBA. If reactions are highly selected for in high biomass strains, they are likely to be present in the wild-type; if they are highly selected against, they probably are not. This model refinement is particularly effective when combined with model comparison, since by understanding several models, and where appropriate transferring elements between them, we can improve them all more effectively than we could if we took into account each model in isolation.

To validate in-silico models against natural data, one option is to perform a multi-level comparison of sensitivity and robustness data [16], at pathway and Pareto front levels. Figure 2 illustrates how one can observe pathway levels sensitivity data, and then aggregate it up to the whole Pareto front. This can provide a whole-organism view, with the precision of a pathway-level analysis, and allow a pathway by pathway validation by sensitivities of each.

5. Conclusion

The combination of techniques described here have been framed in an iterative pipeline. Expression based genetic design appears able to provide more nuanced designs than pure knock-out procedures, and also has the advantage that it is more amenable to evolutionary optimization, since it is a more gradual process with a smoother fitness landscape. Combining this with analyses that aggregate over

multiple phenotypes and take a multi-omic view provides a broad understanding of the complexities of the metabolic network [30]. Our iterative approach also allows us to use an experimental feedback loop: once parts of the network are understood, hypotheses about the knock on affects of changes can be quickly tested. This approach can be conducted across regions, or at differing scales.

Our software is a powerful tool for supporting hypothesis generation, mode of action understanding for candidate drugs, as well as supporting the construction of disease pathway interactions and modelling for drug development projects. In our case study, this framework is able to deliver quantitative improvements in synthesis, and elucidate how these results are generated. The use of a Pareto front allows an understanding of the engineering possibilities that would be more difficult to achieve through other approaches.

References

[1] Benjamin M Woolston, Steven Edgar, and Gregory Stephanopoulos. Metabolic engineering: past and future. *Annual review of chemical and biomolecular engineering*, 4:259–288, 2013.

[2] J Schellenberger, R Que, RMT Fleming, I Thiele, JD Orth, AM Feist, DC Zielinski, A Bordbar, NE Lewis, S Rahmanian, J Kang, DR Hyduke, and B Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0. *Nature Protocols*, 6:1290–1307, 2011.

[3] Gabriel Gelius-Dietrich, Claus Jonathan Fritzemeier, Abdelmoneim Amer Desouki, and Martin J Lercher. sybil – efficient constraint-based modelling in r. *BMC Systems Biology*, 7(1):125, 2013.

[4] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[5] Jan Schellenberger, Junyoung O Park, Tom M Conrad, and Bernhard Ø Palsson. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11(1):213, 2010.

[6] Claudio Angione, Jole Costanza, Giovanni Carapezza, Pietro Lió, and Giuseppe Nicosia. A design automation framework for computational bioenergetics in biological networks. *Molecular BioSystems*, 9(10):2554–2564, 2013.

[7] Jeffrey D Orth, Ines Thiele, and Bernhard ØPalsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–8, March 2010.

[8] Joel F Moxley, Michael C Jewett, Maciek R Antoniewicz, Silas G Villas-Boas, Hal Alper, Robert T Wheeler, Lily Tong, Alan G Hinnebusch, Trey Ideker, Jens Nielsen, et al. Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator gcn4p. *Proceedings of the National Academy of Sciences*, 106(16):6477–6482, 2009.

[9] Anthony P Burgard, Priti Pharkya, and Costas D Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657, 2003.

[10] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

[11] Hadas Zur, Eytan Ruppin, and Tomer Shlomi. imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24):3140–3142, 2010.

[12] Yuliang Wang, James A Eddy, and Nathan D Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre. *BMC systems biology*, 6(1):153, 2012.

[13] Scott A Becker and Bernhard O Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology*, 4(5):e1000082, 2008.

[14] Paul Shannon. *RCytoscape: Display and manipulate graphs in Cytoscape*, 2013. R package version 1.10.0.

[15] Jan J Bot and Marcel JT Reinders. Cytoscaperpc: a plugin to create, modify and query cytoscape networks from scripting languages. *Bioinformatics*, 27(17):2451–2452, 2011 Sep 1 2011.

[16] Jole Costanza, Giovanni Carapezza, Claudio Angione, Pietro Lió, and Giuseppe Nicosia. Robust design of microbial strains. *Bioinformatics*, 28(23):3097–3104, 2012.

[17] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[18] Ka-Chun Wong and Zhaolei Zhang. Snpdryad: predicting deleterious non-synonymous human snps using only orthologous protein sequences. *Bioinformatics*, 30(8):1112–1119, 2014.

[19] Radhakrishnan Mahadevan, Bernhard ØPalsson, and Derek R Lovley. In situ to in silico and back: elucidating the physiology and ecology of Geobacter spp. using genome-scale modelling. *Nature reviews. Microbiology*, 9(1):39–50, January 2011.

[20] Daniel R Bond and Derek R Lovley. Electricity Production by Geobacter sulfurreducens Attached to Electrodes. *Applied and Environmental Microbiology*, 69(3):1548–1555, 2003.

[21] Radhakrishnan Mahadevan, Daniel R Bond, Jessica E Butler, Abraham Esteve-Nuñez, Maddalena V Coppi, Bernhard O Palsson, Christopher H Schilling, and DR Lovley. Characterization of metabolism in the fe (iii)-reducing organism geobacter sulfurreducens by constraint-based modeling. *Applied and environmental microbiology*, 72(2):1558–1568, 2006.

[22] Jun Sun, Bahareh Sayyar, Jessica E Butler, Priti Pharkya, Tom R Fahland, Iman Famili, Christophe H Schilling, Derek R Lovley, and Radhakrishnan Mahadevan. Genome-scale constraint-based modeling of Geobacter metallireducens. *BMC systems biology*, 3(1):15, 2009.

[23] S Boccaletti, G Bianconi, R Criado, CI Del Genio, J Gómez-Gardeñes, M Romance, I Sendiña-Nadal, Z Wang, and M Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 2014.

[24] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[25] Ulrik Brandes. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

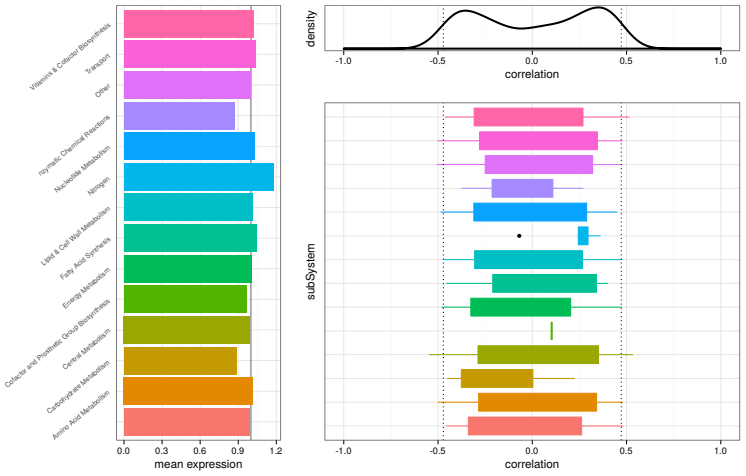
[26] Giovanni Carapezza, Renato Umeton, Jole Costanza, Claudio Angione, Giovanni Stracquadanio, Alessio Papini, Pietro Lio, and Giuseppe Nicosia. Efficient behavior of photosynthetic organelles via pareto optimality, identifiability, and sensitivity analysis. *ACS synthetic biology*, 2(5):274–288, 2013.

[27] Claudio Angione, Giovanni Carapezza, Jole Costanza, Pietro Lió, and Giuseppe Nicosia. Pareto optimality in organelle energy metabolism analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(4):1032–1044, 2013.

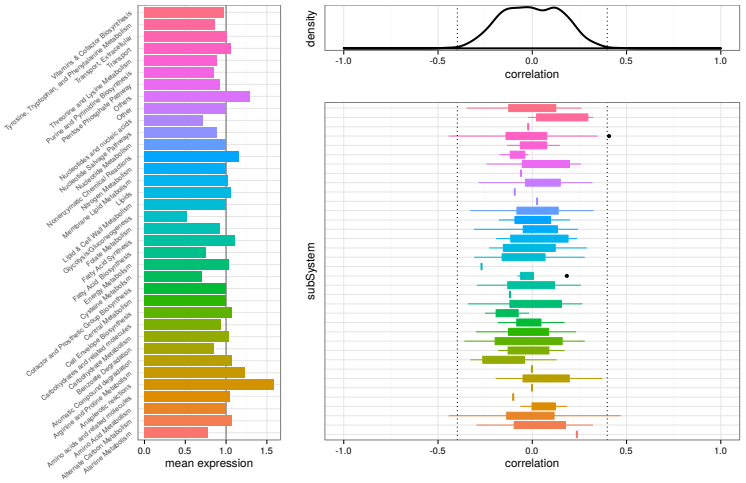
[28] Claudio Angione, Giovanni Carapezza, Jole Costanza, Pietro Lió, and Giuseppe Nicosia. Rational design of organelle compartments in cells. *EMBnet. journal*, 18(B):pp–20, 2012.

[29] Ka-Chun Wong, Chun-Ho Wu, Ricky KP Mok, Chengbin Peng, and Zhaolei Zhang. Evolutionary multimodal optimization using the principle of locality. *Information Sciences*, 194:138–170, 2012.

[30] Daniel R Hyduke, Nathan E Lewis, and Bernhard Ø Palsson. Analysis of omics data with genome-scale models of metabolism. *Mol. BioSyst.*, 9(2):167–174, 2013.



(A) *Geobacter Sulfurreducens*



(B) *Geobacter Metallireducens*

FIGURE 3. The density plot at the top of each figure shows the distribution of correlation between expression level and position in the Pareto front. The Gaussian component to the correlation distributions shows the effect of genetic drift, while in Subfigure 3a, we see marked selection away from 0—this indicates that most of the genes have either a positive or a negative effect on the phenotype, and so most face selection pressure in one direction or the other.The box and whisker plots show the genes, aggregated by subsystem, on the same x-axes as the density plot. The bar charts to the left act as a key, and show the mean expression levels for each subsystem. This figure differs from Figure 4 in that this figure shows correlation between gene activity and phenotype, whereas Figure 4 shows raw expression.

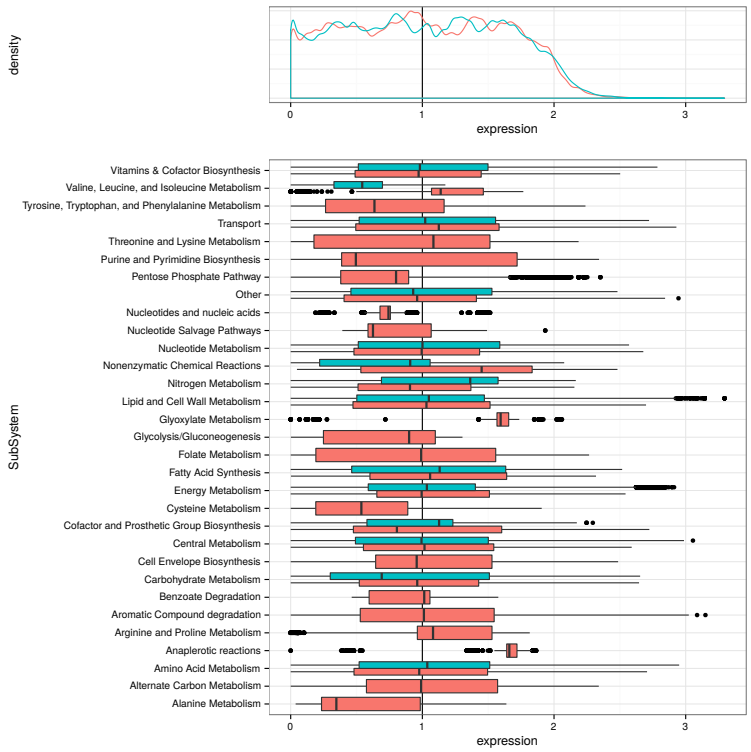


FIGURE 4. The colours show species; *G. Sulfurreducens* is blue and *G. Metallireducens* is red. The density plot at the top shows the distribution of expression levels. Expression levels expressed in fold change from wild-type, so zero means no expression, one means unchanged level from wild-type, and 2 means expression at double the wild-type level.. The box and whisker plot shows the genes, aggregated by subsystem, on the same x-axes as the density plot. All subsystems are up or down regulated significantly ($p < 0.01$, see Table 1). We can see that where subsystems are labelled in both models, they are typically regulated similarly. Where larger differences exist, such as in Valine, Leucine and Isoleucine metabolism, this is due to the subsystems being quite small, so that small differences in the reaction sets included can create relatively large differences in overall regulation. This figure differs from Figure 3 in that this figure shows raw expression, whereas Figure 4 shows correlation between gene activity and phenotype.

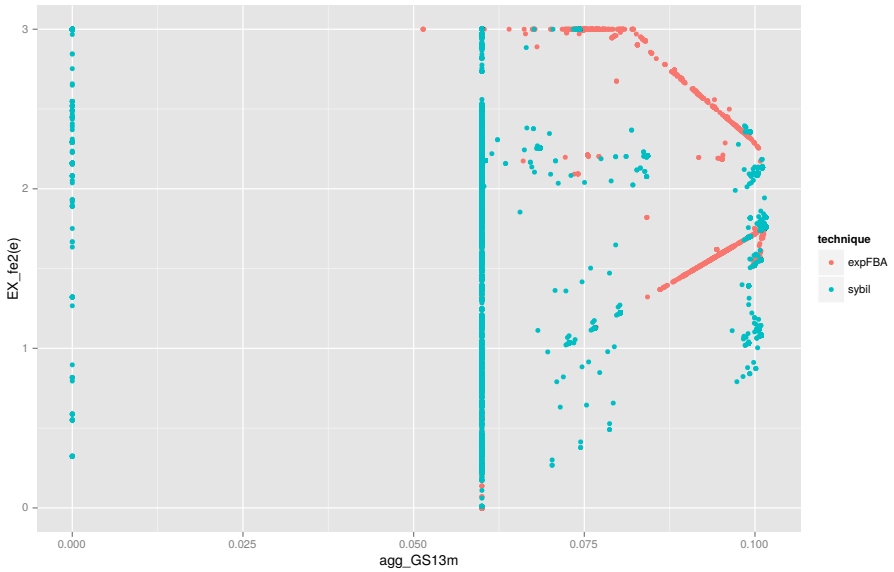


FIGURE 5. Example plot of Fe^{2+} excretion against biomass generation for results from Sybil and expFBA. We can see that the Pareto front generated by expFBA dominates the vast majority of the solutions proposed by Sybil.

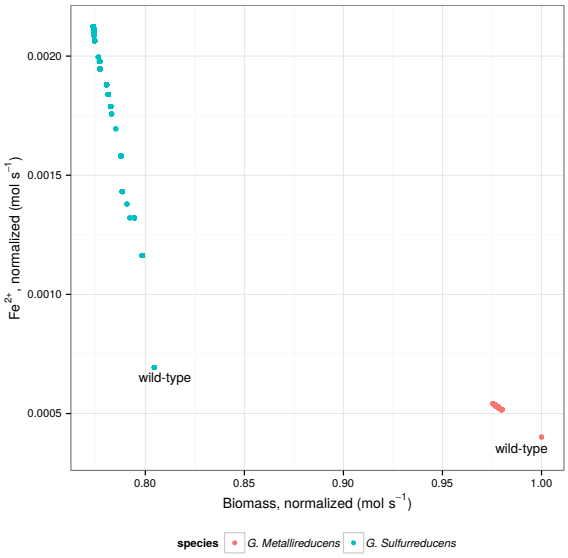
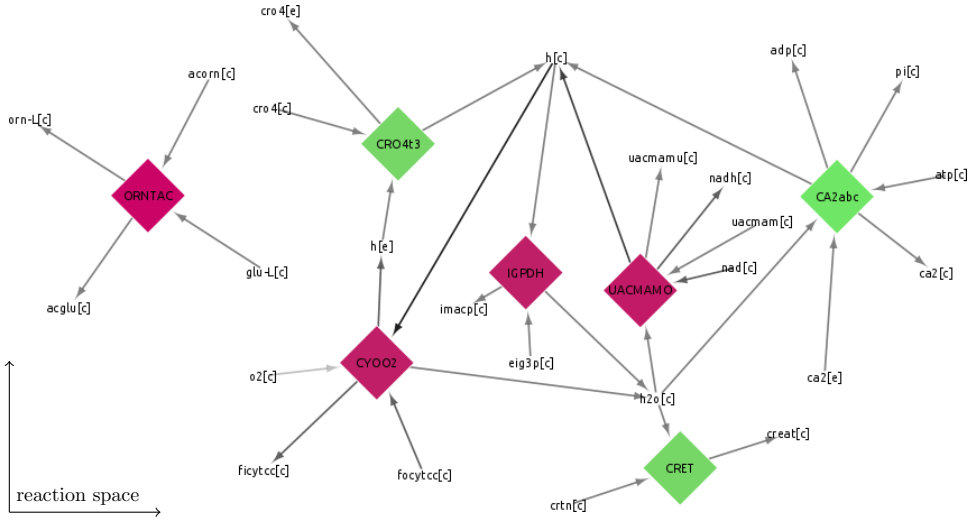


FIGURE 6. Plot of Fe^{2+} excretion against biomass generation for strains of two species of *Geobacter*, aggregated over the last 500 generations of expFBA. Note that the x scale does not start at zero. Interestingly, we can see large gaps between the wild-type and the rest of the phenotypes in both fronts, due to the wild-type dominating the local objective space.



Abbreviation	Name
CA2abc	calcium transport via ABC system
CRET	creatininase
CRO4t3	chromate transport out via proton antiport
CYOO2	cytochrome-c oxidase (2 protons translocated)
IGPDH	imidazoleglycerol-phosphate dehydratase
ORNTAC	ornithine transacetylase
UACMAMO	UDP-N-acetyl-D-mannosamine oxidoreductase

FIGURE 7. Network visualization of a subset of reactions from *G. Metallireducens* and associated key. This subset is derived via a variant on the procedure shown in Figure 3. Nodes in burgundy indicate positive correlation with Fe^{2+} synthesis, while nodes in green indicate negative correlation with Fe^{2+} synthesis. Arrow weight shows stoichiometry. The stoichiometric network shown here only represents a portion of the metabolome. Underlying this metabolic network is the genetic network of genes and their links to the reactions that they regulate, in a multiplex manner [23].