

IMPROVING *METHANOSARCINA* GENOME-SCALE MODELS FOR STRAIN DESIGN
BY INCORPORATING COFACTOR SPECIFICITY AND FREE ENERGY CONSTRAINTS

BY
MATTHEW AARON RICHARDS

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Chemical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Adviser:

Associate Professor Nathan D. Price

ABSTRACT

Genome-scale metabolic models have the potential to revolutionize synthetic biology by informing the development of modified organism strains with enhanced production of desired compounds. A major caveat to this dogma is the lack of extensive experimental data for validating the models of less-studied organisms, relegating the confidence in the capabilities of these models to a much lower level than could potentially be attained with more comprehensive knowledge of organism metabolism. The process of enhancing the predictive capabilities of such models requires an iterative process of ongoing improvement to better reflect established knowledge of organism-specific biology. This work reports modifications to models of the metabolic networks of two methane-producing *Methanosarcina*, the iMB745 model of *M.acetivorans*, and the iMG746 model of *M.barkeri*. With these modifications, I integrated new experimental data and resolved cofactor specificity for reactions that utilize NAD and NADP. I also developed a new method for adding additional to these models by including free energy data for exchange reactions, thus allowing use of experimental data to restrict model predictions to flux distributions that satisfy the second law of thermodynamics.

The updated models are each a more extensively curated body of data than the original models that more accurately reflect wet lab observations than previous iterations. I tested the updated models for their utility as metabolic engineering tools in two ways: (1) by creating models of mutants predicted to consume ethanol and pyruvate, two substrates with immense potential as carbon sources; and (2) by searching for potential knockout targets to enhance methane production *in silico*. These efforts to improve the models of the *M.acetivorans* and *M.barkeri* metabolic networks can serve as a blueprint for understanding cofactor specificity in a range of organisms, and the novel approach for integrating thermodynamics-based constraints by adding free energy data serves as a general tool for improving the constraint based approach to simulating the function of metabolic networks.

Table of Contents

BACKGROUND	1
MODIFICATIONS TO THE RECONSTRUCTIONS	6
Eliminating Cofactor Redundancy.....	6
Updates to General Annotations	9
THERMODYNAMIC CONSIDERATIONS	13
Incorporating Free Energy Data	13
Effects of Thermodynamic Constraints.....	16
APPLICATION TO METABOLIC ENGINEERING	21
Predicted Growth on Novel Substrates	21
Optimization of Methane Production.....	24
MATERIALS AND METHODS	26
CONCLUSIONS	29
REFERENCES	31
APPENDIX A – FIGURES	36
APPENDIX B - TABLES	44

BACKGROUND

Genome-scale metabolic network reconstructions are invaluable tools in the field of systems biology. In addition to serving as organism-specific databases to catalogue biological data, such networks have multiple applications for facilitating biological discovery, particularly by converting information from the metabolic network reconstruction into a simulatable genome-scale model (GEM). In the years since the first GEM was published[1], the process for creating a metabolic network reconstruction has been refined and described in detail [2], resulting in higher-quality models and greater efficiency in model creation. In accordance with these efforts to make these models more widespread, the number of completed GEMs has swelled to nearly 100, encompassing species from all three domains of life [3]. One of the most promising potential uses of these GEMs is to direct metabolic engineering [4-6], such that systems biology can be paired with traditional synthetic biology to create a framework for modifying microbial species. Various groups have achieved some measure of success with this type of application [5], including one pertinent example in which a GEM for *Escherichia coli* was used to guide the improvement of an L-valine-producing strain [7]. In that study, the model was simulated to predict gene knockout targets that could increase the production of L-valine while maintaining a reasonable growth rate. These *in silico* predictions were verified by wet lab experiments that showed high agreement *in vivo*, demonstrating the efficacy of this technique for driving microbial improvement. Other successful examples of using GEMs to predict modifications for existing organisms include engineering an *E.coli* strain with increased L-threonine production [8] and a *Saccharomyces cerevisiae* strain with improved vanillin production [9].

These successes are an encouraging demonstration of the utility of GEMs for bioengineering, especially for improving the production of amino acids and other nutrients in well-studied (“model”) organisms. However, despite these successful cases, endeavors to use GEMs to guide synthetic biology are still in their infancy, particularly for organisms that have not been studied as thoroughly as have *E.coli* and

S.cerevisiae [6]. In general, construction of a genome-scale metabolic network is firmly based not only on the genome sequence of an organism, but also on the accumulation of various empirical data gathered from growth in the wet lab[2]. Hence, GEMs for well-studied organisms such as *E.coli* and *S.cerevisiae* tend to be more extensively curated than those for other organisms, encompassing a much larger body of knowledge from wet lab experiments than any available for less-understood organisms.

In addition to the disparity of existing knowledge available for well-studied organisms compared to less-studied ones, it is also important to consider the degree to which organism GEMs are updated as new information becomes available. The development of new technologies such as next-generation sequencing [10,11] and RNA-Seq [12] have contributed to a massive expansion of available biological data. As more of this information becomes available, it is essential that a GEM be periodically updated in order to ensure that the model is the most comprehensive reflection of known biology. The drive to keep models updated with the latest annotations has spurred the completion of multiple GEM iterations of several different organisms, most notably *E.coli* [13-16] and *S.cerevisiae* [17-23]. Thus, in addition to the already-present knowledge gap resulting from the relative abundance of data for model organisms, multiple efforts to update models for these organisms have further widened the gap. This gap in level of curation between model organisms and less extensively characterized microbes is a primary reason why the majority of successful *in silico* metabolic engineering experiments have been conducted in *E.coli* and *S.cerevisiae* and why there has been much less success with non-model organisms[6].

One particular area of interest for advancement of *in silico* metabolic engineering techniques is in the field of biofuel production, where the development of microbial strains with improved production rates could dramatically impact future energy sources [24]. There are existing models for a number of candidate organisms that naturally produce a variety of valuable fuel stocks, including ethanol, butanol, methane, and hydrogen [6]. For this study, I chose to focus on two methane-producing organisms

(methanogens) from the *Methanosarcina* genus, *M.acetivorans* and *M.barkeri*. Methanogens are archaea that are able to grow on low-energy carbon substrates by undergoing methanogenesis, whereby the carbon substrate is reduced to methane in order to produce the ATP necessary to achieve growth. *Methanosarcina* are particularly notable in that unlike other methanogens, they are known to utilize all the different carbon substrates possible for methanogenesis, including carbon dioxide, acetate, and a number of one-carbon (C1) compounds [25]. In addition to possessing a substrate robustness that is highly desirable in metabolic engineering targets, the *Methanosarcina* also benefit from having an established set of genetic engineering tools [26], making them an excellent candidate for pairing *in silico* engineering efforts with wet lab experiments. Moreover, the methane produced by these organisms not only holds enormous potential as a biofuel, but also contributes to global warming by way of ozone degradation. Thus, by studying the metabolisms of these archaea, we may gain insight into ways we can affect both the global carbon cycle and the direction of the burgeoning biofuels industry. However, as with other non-model organisms, the *Methanosarcina* GEMs have not been as extensively curated as GEMs of model organisms. Therefore, additional model curation is a prerequisite to applying GEMs of these non-model organisms to guide strain development for improved production of methane.

Along these lines, one important opportunity for improving GEMs is the careful curation of enzyme substrate, product, and cofactor specificity. This aim is particularly salient in the case of reactions that transfer reducing equivalents using the nicotinamide electron carriers, NAD and NADP. These cofactors are often dually attributed to the same primary reaction, but based on values found in the BRENDA database [27], they display different levels of enzyme activity. In addition to displaying differing levels of activity, these dually-associated cofactor pairs can greatly affect predictions derived from using a GEM due to their ability to alter cofactor balancing in model simulations [28]. Thus, updating general annotations and distinguishing between potential electron carriers in dually-associated cofactor pairs are both critical steps towards improving the accuracy of a genome-scale model.

Aside from the improving reaction and product information in the reactions themselves, there is a growing effort to incorporate thermodynamic data into metabolic networks. From an energetic standpoint, even models with the best possible stoichiometric representations of metabolism cannot be considered comprehensive without the addition of constraints on free energy. Several groups have harnessed the implications of the second law of thermodynamics to restrict metabolic models by constraining all reactions in the model to operate in the direction of favorable (negative) free energy [29-31]. More recently, groups have developed an extension of the COBRA toolbox to assign directionality to each reaction of a GEM [32], as well as a method designed to perform flux balance analysis (FBA) and flux variability analysis (FVA) with energetic constraints at greater computational speed [33]. All of these methods rely on a wealth of data for free energies of formation, most of which were calculated using the group contribution method [34] due to the dearth of measured data for common biological compounds. The reliance on generated data in place of empirical evidence has certainly been necessary to undertake the ambitious task of constraining every reaction in a genome-scale model, but this reliance introduces uncertainty in the model that can affect the viability of these methods. Notably, it has been suggested that a reaction-by-reaction approach to restricting network reversibility could give misleading information if knowledge of the network thermodynamics is incomplete [35]. In order to avoid the possibility of such a misrepresentation, the constraints on metabolic models must be based upon measured experimental data. Hence, much like updates to the reaction network, any application of thermodynamic data to constrain a GEM must reflect biologically-verified information.

In this study, the two most recently published *Methanosarcina* reconstructions, iMB745 and iMG746 [36,37] were updated to reflect the most current biological information available. They were also modified in order to remove redundant cofactor reactions and apply overall thermodynamic constraints on free energy, both of which eliminated infeasible results that appeared in simulations of the original

models. All of these modifications were part of an effort to produce the highest-quality models possible for generating *in silico* growth predictions to guide wet lab efforts for engineering these organisms. As a first step in this direction, I employed the new models to predict growth on novel carbon sources and simulated reaction knockouts that could increase methane secretion rates in these organisms.

MODIFICATIONS TO THE RECONSTRUCTIONS

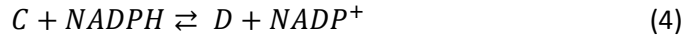
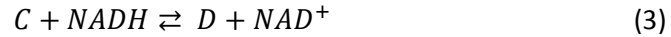
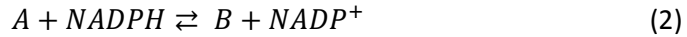
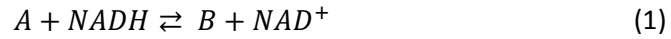
Both *Methanosarcina* models were altered to improve the accuracy of the reconstruction of established biochemical knowledge. In some cases, these modifications were made in order to incorporate novel biological information that necessitated that reactions be added, removed, or associated with new genes in the models. Aside from updating based on new findings, the *Methanosarcina* models were also modified to more accurately simulate the roles of the electron carriers NAD⁺ and NADP⁺.

Eliminating Cofactor Redundancy

The distinction between reactions using NAD(H) as an electron carrier and those that prefer NADP(H) is one that is generally difficult to make, particularly when selecting reaction associations directly from a reaction database that is not organism-specific, such as KEGG or MetaCyc [38,39]. If a reaction is likely to use one of these two molecules, it is often given two separate associations in these databases, such that a single gene is linked to both possible reactions. When a metabolic network and its corresponding GEM are constructed from the information, this double reaction association is usually incorporated into the reconstruction, resulting in a reaction pathway that can utilize either cofactor (NAD/NADP) in order to transfer reducing equivalents. However, as evidenced by experimental measurements of activities for enzymes that use these cofactors[27], there is generally a preference for one species over the other. The accuracy of reconstruction of this cofactor specificity impacts the predictive ability of a GEM: studies in *Saccharomyces cerevisiae* have suggested that switching enzyme cofactor specificity can have a major effect on how cofactors are balanced in a metabolic model as a whole, even when the specificity is changed in only two enzymes in the entire model [28].

Furthermore, the presence of dual associations can greatly expand the space of feasible solutions achieved by solving a metabolic model. A large number of reactions in any given GEM tend to depend on the availability of a sufficient number of cofactors, particularly in their reduced forms. In biological systems, this need places a limitation on metabolism because there must be enough reducing

equivalents generated somewhere in the system to appease the demands set by NAD(P)H-requiring reactions. As an additional limitation, the ratios of oxidized to reduced cofactors (NAD/NADH and NADP/NADPH) are kept very different from one another in cellular environments, making the energetic contributions of these compounds noticeably different and putting a biological constraint on the relationship between these two ratios that requires that they be properly balanced [28]. However, the dually-associated cofactor pairs are able to avoid these constraints for the most part due to their ability to cycle cofactors with one another. As an example, suppose that the reversible reaction of compound A to compound B requires reducing equivalents and the reversible reaction of compound C to compound D does as well. If both reactions are dually-associated, then there are 4 reactions in the model:



Suppose that the overall reaction of primary metabolites (A,B,C,D) requires that $2A + C \rightarrow 2B + D$; there are an infinite number of ways to solve this problem, all of which affect the number of each cofactor being produced. One possible solution to the GEM might be $3(1) - (2) + (3)$, which also results in $4 NADH + NADP^+ \rightarrow 4 NAD^+ + NADPH$. A solver could choose this solution when the model is simulated, but perhaps there is another reaction in the model that requires more NADPH be produced. Rather than imposing this requirement as a constraint on another reaction of the model, this set of reactions can simply adjust to “cycle” cofactors by effectively reducing NADPH using electrons from NADH. A solution in this case could be $4(1) - 2(2) + 3(3) - 2(4)$, which also gives the correct primary solution, but results in $7 NADH + 4 NADP^+ \rightarrow 7 NAD^+ + 4 NADPH$. The ability to cycle in this manner and use 3 NADH to produce 3 NADPH gives the models extra flexibility that artificially expands the realm of

possible solutions achievable. Thus, it is imperative that such solutions be eliminated by removing redundant cofactor pairs from the models.

I investigated both *Methanosarcina* models for reaction pairs involving either NAD(H) or NADP(H) where all compounds and their stoichiometry were the same, save for a switch from NAD(H) to NADP(H). In the iMB745 model of *M.acetivorans*, I found 16 such pairs; in the iMG746 model of *M.barkeri*, I found 12 such pairs. Eliminating these redundant pairs required that one or both of the reactions in each pair be removed from the metabolic model. In order to investigate model dependence on specific cofactors, I performed a swap experiment in which one member of each cofactor pair was randomly removed, and the model was simulated using FBA to determine whether growth was possible. I reasoned that it was possible that the models were dependent on certain reactions to use specific cofactors, thus the knockout of these reactions would prove lethal to the models and would demonstrate that these reactions should be kept in the models and the corresponding members of their cofactor pairs should be eliminated. This swapping experiment proved to be unfruitful for determining reaction essentiality based solely on model performance because I was unable to find specific cofactor preferences. Instead, the confirmation of model independence with respect to cofactor specificity allowed me to explore other ways of choosing between reactions in redundant pairs.

In a second effort to determine cofactor specificities for these reactions, I turned to literature evidence contained in the BRENDA database [27]. All cofactor pairs were evaluated using the best available information on enzyme activities to distinguish which cofactor was most likely preferred in each scenario. These literature sources [40-51] provided an experimental basis for my inferences and increased my confidence that my reaction choices had at least some biological evidence for their inclusion in the models. Through this literature review, I eliminated 11 reactions from iMG746 and 15 from iMB745 that had incorrect cofactor specificity. However, each model still had a remaining pair of

reactions (succinate semialdehyde dehydrogenase reactions) with insufficient evidence of cofactor specificity. Further, these reactions had insufficient literature or bioinformatics evidence to justify their inclusion in the metabolic reconstructions, and so I removed them both from each model. Thus, I removed 13 and 17 reactions, respectively, from iMG746 and iMB745.

Following the removal of the reactions from the two models, I wished to be sure that the changes undertaken did not prove detrimental to the models' predictions. The swapping experiment had already shown that my modifications would not be considered lethal alterations, but I was unsure what the quantitative effects would be, particularly on predicted growth yields and on predicted flux of methane and carbon dioxide. To these ends, I simulated both models on a variety of substrates for which experimental data was available [52-62] and compared the results of my simulations to both the measured experimental values and the values from the original models [36,37]. As shown by these comparisons (Figure 1, Figure 2), the deviations from the previous model predictions were minimal, with the largest change constituting only an 18% difference from the predicted value in the original model (growth rate of *M.barkeri* on CO_2+H_2). Most importantly, my confidence in the new models was higher than in the original models because the absence of any redundant cofactor pairs lent more credibility to my models' depiction of actual biology. Therefore, the new models were deemed an improvement on the previous iterations in their handling of the electron carriers NADP(H) and NAD(H).

Updates to General Annotations

Even though both GEMs being used in this study were published within the last 2 years [36,37], sufficient additional data was available to make a number of changes to both models. The majority of these updates came from comparative genomics evidence with other *Methanosarcina*, though some were also inferred from other sources [40,63]. For the *M.barkeri* and *M.acetivorans* models, respectively, 8 and 9 reactions were added, 10 and 13 were removed, and 4 and 8 had their gene-association changed. A comparison of the resulting models with the original models can be found in

Table 1. In addition to revising the existing annotations to include new information, these updates provided several new suggested changes to refine the reconstructions of *Methanosarcina* metabolic networks that were also incorporated into the models. Although these modifications were not part of my originally-planned model updates, their inclusion lent more confidence to my updated models and served as an example of how curating genome-scale models can be valuable as tools to generate biological hypotheses.

A common step of model curation is to perform gene or reaction essentiality tests, both as a qualitative comparison to experimental results and as a check for model consistency throughout the curation process. In an essentiality simulation, all genes or reactions for a model growing on a certain substrate are iteratively removed in single knockouts and the results indicate which genes/reactions are necessary for model growth to be predicted. During one such test, it was predicted that *M.barkeri* required the enzyme threonine aldolase (THRAr) in order to grow, a puzzling result considering that I expected the model to rely on threonine synthase (THRS) for this function. Furthermore, my experimental knowledge of the organisms led me to believe that THRAr was an incorrect annotation and did not belong in either model.

Using the models as a guide, the root of the essentiality prediction for THRAr was traced back to one particular metabolite, acetaldehyde, which was predicted to be produced elsewhere in the model and was used as a reactant in THRAr. Because there was no other sink for acetaldehyde in the models, the knockout of this reaction caused an acetaldehyde buildup and violated the steady-state assumption. In order to fix this problem and correctly remove THRAr from the models, I added an exchange reaction for acetaldehyde, such that the metabolite could freely leave the model as an excreted product. This addition represented a possible biological pathway for which I had no empirical evidence, complicated by the fact that acetaldehyde has not traditionally been a compound of much experimental interest and

therefore, to my knowledge, acetaldehyde transporters have not been investigated in *Methanosarcina*, in an experimental setting. However, despite the lack of biological evidence for the inclusion of this reaction, the exchange of acetaldehyde freely through the cell membrane during growth is a reasonable possibility because the membrane is permeable to acetaldehyde and, more importantly, allowed me to remove the incorrectly-annotated THRAr reaction from the models. Moving forward, this new exchange reaction can serve as a possible area for further model improvement, particularly if future genome annotations can suggest another possible sink for acetaldehyde in *Methanosarcina* metabolism.

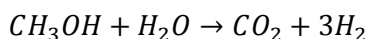
During the process of eliminating cofactor redundancy, as described previously, I encountered another issue with the models that necessitated an unintended modification. In the NADH/NADPH cofactor pair for glyceraldehyde-3-phosphate dehydrogenase (GAPD), there was substantial evidence suggesting that NADPH was the preferred cofactor in *Methanosarcina* [40]. However, once the NADH-dependent GAPD was removed from the models, predicted growth rate fell to zero as a result of the models' inability to deal with the now-present cofactor imbalance, chiefly due to the cofactor preferences selected for several other enzymes. Restoration of the predictive capabilities of the models depended on finding a way to rebalance the cofactors by allowing them to exchange with one another.

This problem was solved by incorporating an electron bifurcating ferredoxin reduction reaction that has been described in the metabolisms of *Clostridium kluyveri* [64] and *Moorella thermoacetica* [65] into the models. In this reaction, electrons from reduced ferredoxin were used to drive the transfer of electrons from NADH to NADPH, effectively carrying out a cofactor switch that transferred reducing equivalents from 2 reduced ferredoxins and 1 NADH into 2 NADPH molecules. Electron bifurcation is a common phenomenon that has been previously described as being present in *Methanosarcina*, as well as several other organisms [66,67]. The presence of this particular reaction was strongly supported by sequence homology to the corresponding genes in *C.kluyveri*, as verified by running BLASTP[68] on the associated

protein products. Its inclusion in the models helped incorporate an important feature of *Methanosarcina* biochemistry that had been previously unaccounted for while simultaneously allowing me to eliminate cofactor redundancy caused by the presence of two GAPD enzymes. This reaction addition should serve as a catalyst for further investigation into other possible sites of electron bifurcation in the metabolic network and explore their effects on the *Methanosarcina* electron transport chain

THERMODYNAMIC CONSIDERATIONS

The need for thermodynamic constraints was made evident by a deficiency discovered in the *M.barkeri* model. During simulations to compare reaction essentiality in the models, I found that the model of *M.barkeri* predicted growth on methanol without the reactions necessary for methanogenesis. Specifically, it was predicted that the final 3 steps in the methanogenesis pathway (see Figure 3) were not essential for growth. As discussed previously, methanogens must necessarily perform methanogenesis in order to achieve growth, and so this curious result presented a clear violation of biological knowledge. In whole, the model predicted that instead of reducing methanol to methane, it could use an alternate pathway (Figure 3) to produce hydrogen instead with the overall reaction:



From purely a stoichiometric perspective, there was nothing inherently wrong with this overall reaction. In fact, had it not been for my knowledge of the biology of this organism, I may have accepted this prediction as a possible alternate pathway in *M.barkeri* growing on methanol. This example illustrated not only the inaccuracy of my model, but also underscored the importance of using measured values to constantly evaluate the predictions. In this case, the evidence I turned to was thermodynamics, which suggested that if the reaction were to occur under standard conditions (25 C, 1 M, 1 bar, pH = 7) the free energy of the reaction would be +79.3 kJ/mol [69]. Thus, my thermodynamic analysis showed that this alternate pathway was infeasible under these conditions and suggested that in order to eliminate the possibility of generating thermodynamically-infeasible flux solutions in my simulations, I should incorporate additional constraints based on free energy data into the models.

Incorporating Free Energy Data

Imposing thermodynamic constraints on GEMs has long been an area of interest in the systems biology community. Generally, previous efforts have focused on restricting reaction reversibility in accordance with data on the Gibbs free energy for every reaction in the model [29-31]. These groups have used

their resulting constraints to predict limitations on possible flux distributions achievable in FBA solutions, as well as to explore the allowable concentrations of different metabolites in the model [29,30]. However, due to the relative shortage of thermodynamic information for the majority of metabolites within a model, most of these free energies have been estimated using the group contribution method (GCM) [34]. Despite the advances in the implementation of GCM, predicted values for Gibbs free energy of formation (and thus of reactions) have an inherent amount of uncertainty associated with them. Indeed, applying reversibility constraints in a model of *E.coli* without considering uncertainty in GCM values resulted in a model that could not predict growth when simulated, whereas incorporating uncertainties into the model allowed for *in silico* growth, but forfeited the original goal of restricting reaction reversibilities [29]. Additionally, internal concentrations (those for metabolites present only inside the cell) are difficult to measure, restricting the applicability of concentration measurements on a genome-wide basis. Thus, I sought to formulate a new method for incorporating free energy constraints into my metabolic models that did not rely on GCM values or on internal concentration measurements.

In contrast to the metabolites and reactions inside the cell, there is a relative wealth of measured data available for common biological compounds observable outside the cell environment [70,71]. This is because it is much easier and more accurate to measure the free energies and concentrations for media and byproducts than for intermediates that occur only inside the cellular environment. Thus, we can usually find better data for metabolites that are consumed or secreted by the model than for internal metabolites. Hence, rather than restricting the reversibility of every reaction in a model, I can instead use thermodynamic data to determine the feasibility of my overall model “reaction”. In the case discussed where hydrogen was being produced from methanol, imposing a thermodynamic constraint for the overall reaction of the model would restrict the model from allowing a reaction with an unfavorable (positive) free energy and prevent the problem of generating thermodynamically infeasible

solutions. Additionally, using measured data would increase confidence in the free energy constraints being imposed on the model because they would not be associated with the uncertainty attached to values calculated by GCM.

As a test of the proposed approach of applying thermodynamic constraints on metabolites external to the cell, the *Methanosarcina* GEMs were simulated under substrate conditions used in experimental settings to determine non-zero exchange fluxes in the models. This generated a list of metabolites for each organism model that could be considered as participating in the overall “reaction” stoichiometry. For each list, I found the free energy of formation (ΔG_f) value for each compound, using measured data in nearly all cases [71], though GCM values were used for several cases where I could not find measured data. The inclusion of GCM values was not ideal, particularly because this was in contrast with the aim of using measured data, but using GCM values in these cases did not impact the model calculations because reactions using these compounds carried only small fluxes. The calculated values were kept in simulations primarily for consistency, such that all metabolites entering or exiting the models had an attached value for free energy of formation. All ΔG_f data for these compounds were gathered using the eQuilibrator tool [69] and standard temperature of 25°C, concentration of 1M, and pH of 7. These data encompassed all chemically-defined compounds exchanged by the model, but neglected the free energy contribution made by the formation of biomass. To find a reasonable estimate of free energy required for biomass formation in *Methanosarcina*, I began with a literature review of approaches used in other organisms. In general, “biomass” is an organism-specific mixture of amino acids, lipids, carbohydrates, and various other components, and as such there is not a measured free energy for biomass formation in *Methanosarcina*. However, based on a statistical thermodynamics approach to calculate the entropy of biomass [72], Roels calculated the free energy of combustion for an average biomass composition of $\text{CH}_{1.8}\text{O}_{0.5}\text{N}_{0.2}$ [73]. In addition to having a good agreement with several methods used to estimate free energy of biomass formation in *E.coli* and *S.cerevisiae* [74], the biomass formulation matches that of the

Methanosarcina models fairly closely ($\text{CH}_{1.84}\text{O}_{0.66}\text{N}_{0.28}$) and was therefore determined to be a suitable estimate for use in model simulations. In accordance with unit convention typically used in GEMs, the free energy of combustion was converted to ΔG_f on a per-GDW basis, yielding a value of $-0.176 \frac{\text{kJ}}{\text{GDW}\cdot\text{h}}$ for biomass formation in *Methanosarcina*.

The standard model format employed by the COBRA Toolbox in MATLAB was appended to include this free energy of formation data. As described in Materials and Methods, the ΔG_f values for a particular organism GEM were linked to the exchange reactions themselves, such that any metabolite that was allowed to exit the model caused the “production” of its associated ΔG_f . As a convention, these exchanges reactions were standardized so that a metabolite leaving the system (i.e. an overall product) was given a positive flux, whereas a metabolite entering the system (i.e. an overall reactant) was given a negative flux. Hence, by simply making ΔG_f for a compound the product of its exchange reaction, the contribution of that compound to overall reaction free energy (ΔG_r) followed the standard convention:

$$\Delta G_r = \sum_{\text{products}} \Delta G_f - \sum_{\text{reactants}} \Delta G_f$$

In this manner, the overall free energy of an organism model was calculated by adding a “reaction” to the model that added the ΔG_f contributions generated in all of the exchange reactions and output this sum as the model ΔG_r . As a final step in implementation, FBA solutions to a model were restricted such that model $\Delta G_r < 0$; thus, only thermodynamically-feasible flux distributions were allowed.

Effects of Thermodynamic Constraints

Both *Methanosarcina* models were augmented to incorporate the relevant free energy information, thereby restricting all flux distribution solutions to adhere to the principle of thermodynamic feasibility. I hypothesized that by including these extra constraints to restrict the solution space for the models, I had effectively restricted the models such that all possible solutions were in better agreement with

observed cellular behavior. In order to test this hypothesis, I performed reaction essentiality simulations (described previously) for the models on various substrates and compared these results to the same simulations performed on the models without free energy constraints. This comparison, shown in Table 2 along with the predicted overall free energy for each simulation, demonstrated that although adding these constraints did not change reaction essentially much for the most part, there was a noticeable change in the gene essentiality prediction for the model of *M.barkeri* growing on methanol. As expected from my earlier analysis, the addition of thermodynamic constraints made the alternate pathway in this model, whereby methanol could be converted to hydrogen, infeasible because it violates the constraint of $\Delta G_r < 0$.

Following this analysis, I wished to take a more rigorous look at the effects caused by my incorporation of free energy. I performed a sensitivity analysis aimed at quantifying the impact that varying the ΔG_f of different exchange metabolites could have on the overall ΔG_r for a model. In this analysis, models of both *Methanosarcina* were each simulated on several substrates, corresponding to the substrates simulated previously (both on methanol and acetate, *M.barkeri* on H_2/CO_2 and *M.acetivorans* on CO). For each simulation, the ΔG_f values for 5 components utilized in each model (CH_4 , CO_2 , H_2O , HCO_3^- , biomass), as well as the value for the model-specific growth substrate, were independently varied in a range of +/- 50% of the standards. Each model was simulated for this range of values and the overall ΔG_r of the model was plotted against the ΔG_f of the varied parameter.

As shown in Figures 4A-5C, perturbations of 50% in individual free energies of formation were sufficient to make 4 of the 6 models thermodynamically infeasible. For these models, four metabolites (CH_4 , CO_2 , H_2O , substrate) were able to perturb overall model free energy to the point of infeasibility (in the case of acetate as a substrate, HCO_3^- can be considered representative of CO_2 because the two are able to rapidly interconvert in nature, a fact not shown by the current models). The exceptions to this pattern

were those models growing on methanol, which had such favorable overall free energy that even variations of up to 50% in a given metabolite were not enough to render a model thermodynamically infeasible. However, the addition of thermodynamic constraints did not leave the methanol models unchanged because, as described earlier, these constraints prevented the models from predicting growth of *M.barkeri* without essential methanogenesis reactions.

Also of particular note in this analysis was that the contribution of biomass did not make sizable changes to overall model free energy in any simulation. Although this may accurately reflect free energy change in nature, I believe this contribution must be rechecked because if my assumed value for free energy of biomass formation is incorrect by an order of magnitude (which could realistically be the case), model sensitivity to perturbations of this value could greatly affect overall model feasibility, even under standard conditions. As a whole, this analysis confirmed that my implementation of free energy constraints was able to constrain the model solution space somewhat and reinforced the notion that changes to the values used for standard free energies (particularly changes caused by differences in substrate concentrations) could have a sizable impact on the scope of model predictions.

The analysis was also important for identifying the metabolites that could noticeably affect overall free energy. Going forward, it will be crucial to incorporate the effects of temperature and metabolite concentrations on overall free energy. As described above, my constraints were implemented with standard temperature of 298 K and concentrations of 1 M assumed for each metabolite, but in reality, this would be an unlikely scenario. Metabolite concentrations are related to reaction free energy by the equation

$$\Delta G_r = \sum_i \Delta G_f^0 + RT \ln Q$$

where ΔG_f^0 is the standard free energy for each component, R is the gas constant, T is temperature, and Q is the equilibrium quotient, which accounts for the concentration of each reaction species. As demonstrated by this equation, the incorporation of concentration and temperature data could have a sizable impact on the overall model free energy predictions. The sensitivity analysis was an intermediate step towards understanding the impact of these parameters on model constraints because it allowed me to explore a wide range of non-standard free energy values for my metabolites and observe their impact on ΔG_r . Additionally, this analysis also identified the metabolites that noticeably impact overall free energy predictions, allowing me to focus on these metabolites in future studies, particularly those including temperature and concentration data.

As a final check, I wished to verify that the free energy values I predicted were comparable to those measured experimentally. Unfortunately, there is a shortage of this type of data, particularly for *Methanosarcina* species, but in one particular instance the free energy change in a culture of *M.barkeri* growing on acetate was measured using a calorimeter [75]. Because the free energy prediction in the model of *M.barkeri* was actually a flux (kJ/GDW·h), I used the estimated time of the exponential growth phase (reported as “about 4 days”) to convert this rate to a comparable value. My resulting model-predicted value of -588 kJ/C-mol was of the same order of magnitude as that measured experimentally (-366 kJ/C-mol). This was a very encouraging result because it demonstrated the efficacy of this method for simulating measured values for organism free energy. It is also important to note that this result was achieved without considering a number of factors:

- 1) The time conversion was reported imprecisely, leading to a high degree of uncertainty in the actual time spent in the exponential growth phase during the experiment
- 2) Model predictions for free energies of formation were dependent on maintaining constant concentrations of 1M, whereas concentrations in the experiment were dynamic

- 3) Our assumption was a temperature of 25°C, which differed from the conditions used experimentally (though this would likely cause a much lesser effect than the incorporation of dynamic concentrations)
- 4) Values calculated for ΔG in the paper were reliant on a method [73] that differed from the method I used, and were likely less accurate because they were not rooted in experimentally measured data [70,71]

As my efforts to incorporate thermodynamic constraints move forward, these factors must be considered to achieve the most accurate portrayal of free energy possible. However, the constraints implemented in this work were an important step because they demonstrated that using measured free energy values is an effective method both for determining thermodynamic feasibility of flux solutions and for estimating overall free energy associated with organism growth.

APPLICATION TO METABOLIC ENGINEERING

Updating the *Methanosarcina* GEMs to reflect the best information available was an important step towards improving the understanding of these organisms; however the true test of these models was using them not merely as databases of organism-specific information, but as tools to aid in directing metabolic engineering. To these ends, the models were applied to two metabolic engineering problems with the hope that the predictions made *in silico* could provide valuable guidance to the biologists addressing these challenges *in vivo*. As a first step, both models were engineered to create mutant strains that could utilize new carbon sources, ethanol and pyruvate. These engineered pathways were reflective of actual organisms being grown in the lab, enabling me to use the models to aid in developing and improving the mutants by creating a map of these novel metabolisms. In the second instance, I used a computational method to simulate various reaction knockouts in both wild-type and mutant models and investigate potential knockout targets to improve the productivity of these organisms.

Predicted Growth on Novel Substrates

Ethanol and pyruvate are two prominent components present in waste streams from fermentation processes [76]. Although there are several other compounds in these streams that could potentially be used as carbon sources for production of methane, ethanol and pyruvate are particularly attractive because of their theoretical proximity to the existing acetoclastic pathway in *Methanosarcina* metabolism[25]. The proposed pathways, shown in Figure 6, demonstrate this proximity; ethanol catabolism *in silico* requires 5 reactions (two enzyme-catalyzed and three for modeling purposes) to produce acetyl-CoA and join the established pathway and pyruvate catabolism *in silico* requires only 1 reaction, catalyzed by a pyruvate oxidoreductase (POR2) that is already present in both organisms. Additionally, a pyruvate mutant for *M.barkeri* has already been isolated and grown in another lab [56], demonstrating the viability of using this substrate as an alternative carbon source.

We employed the *Methanosarcina* GEMs to help direct metabolic engineering efforts by adding reactions to the model that are catalyzed by enzymes that can be added *in vivo*. The stoichiometry of these reactions was confirmed using the BiGG database [77]. Five reactions were added to the wild type *Methanosarcina* models to allow for the uptake and utilization of ethanol. The two previously mentioned enzyme-catalyzed reactions responsible for producing acetyl-CoA were NAD-dependent ethanol dehydrogenase (ALCD2x), which converted ethanol to acetaldehyde, and acetaldehyde dehydrogenase (ACALD), which converted the acetaldehyde to acetyl-CoA. Three additional reactions were also required to enable mathematical analysis of the model: an exchange reaction and transport reaction, which were required for ethanol to enter a model and were essential mostly for model completeness; and NADH:NADP transhydrogenase (NADNADP), which uses NADH reduced in both ALCD2x and ACALD to reduce 4 molecules of NADP^+ to NADPH. This transhydrogenase reaction was essential to both the models and the actual mutant organisms grown in lab because it allowed the organisms to switch cofactors needed for production of reduced F_{420} that is required for methanogenesis. *In vivo*, enzymes that performed the ALCD2x and ACALD reactions were taken from *C.kluyveri* and the NADNADP enzyme was taken from *E.coli*, with the activity of all enzymes in the transformed *Methanosarcina* confirmed by assay.

In the models for the pyruvate-utilizing mutant, I took a slightly different course from the pyruvate mutants previously grown. Rather than relying on the already-present POR2 enzyme, which converted pyruvate to CO_2 and acetyl-CoA while reducing ferredoxin, the gene for this enzyme was knocked out *in silico* in favor of a pyruvate formate lyase (PFL), which produced formate instead of CO_2 and did not include ferredoxin. Hence, even though the same primary transformation of pyruvate to acetyl-CoA was present, the reducing equivalents of the reaction were preserved in formate rather than passing to ferredoxin. A formate dehydrogenase (FDH) was also added to the models and was essentially capable of carrying out the second half of the POR2 reaction to produce CO_2 and reduced ferredoxin. Thus, the

model was equivalent to the wild type models in its treatment of pyruvate, but the change to PFL and FDH properly depicted the reactions being transformed into the actual organisms in the lab. For completeness, a formate exchange was also added to allow formate to exit the models, representing the ability of formate to diffuse through the cell membrane in reality, but effectively as a bypass of the FDH reaction for the purposes of simulation.

Once these new models were created, optimal growth yield was calculated using FBA. The simulation results indicated that both models are capable of predicting both growth and methane production. In fact, those models using ethanol as the substrate outperformed all other models, both wild type and pyruvate-consuming, in methane production on a per-mole of carbon basis (Figure 7), as well as on a per-mole of CO₂ basis (Figure 8). The predicted biomass fluxes for the pyruvate models were not as high as the fluxes predicted in simulations of other strains, but the predictions of growth and methane production alone were encouraging signs that there is potential to engineer a pyruvate-consuming *Methanosarcina* strain that is capable of methane production.

Similar to the process of model building and updating, the process of model-directed metabolic engineering is an iterative process. Thus, the next test of these models was to compare the results of these simulations with those found *in vivo*. This comparison itself was difficult to achieve because success with growing these mutant strains in the wet lab has been limited so far. At the time of these model experiments, the only strain successfully grown in lab was an *M.acetivorans* mutant capable of producing methane from ethanol, but incapable of growth on that substrate. Therefore, as promising as the model simulations are, they do not currently reflect observed strain behavior. The other models do not yet have *in vivo* confirmation, but the comparison of the model predictions for the ethanol consuming strain with *in vivo* observation illustrates two salient points with regard to the current state of model-guided synthetic biology: first, it is critically important that models be subjected to validation

and that they be constantly updated and compared to biological data in order to make them depict real-life organisms as accurately as possible; and secondly, it is vital to keep in mind that the results of *in silico* models cannot be treated in the same manner as experimental evidence gathered in a lab using actual organisms. With respect to the results described here, the deviation from experimental results thus far should serve as a catalyst to drive more updates to the models, a process which will also help elucidate engineering strategies to improve *Methanosarcina* strains to achieve both growth and methane production.

Optimization of Methane Production

Much of the current utility of GEMs in driving metabolic engineering is derived from the role they can play in suggesting potential knockout targets in wild-type organisms. Several algorithms have been developed along these lines with the general principles of finding reaction or gene knockouts that are simulated to increase production of a desired substance while simultaneously encouraging a high growth yield [78-81]. In this study, I used Genetic Design through Local Search (GDLS) [81], implemented in version 2.05 of the COBRA Toolbox [82] to make these predictions. In addition to its ease of use through direct integration with the latest version of the toolbox, GDLS was chosen over comparable methods because its strategy of using multiple search paths to speed simulations and effectively explore alternative knockout targets offered an attractive combination of swiftness and thoroughness.

Each wild-type model, as well as each mutant strain model developed for utilization of ethanol and pyruvate, was optimized for maximum methane production, with a maximum of 4 knockouts allowed and a minimum growth requirement of 10% of wild-type growth. The algorithm was also tuned to allow for two separate “search paths”, meaning that for each additional knockout target, the algorithm selected two candidates for knockout rather than one. Specifying this parameter allowed me to broaden the scope of the optimizations because the use of multiple search paths naturally made for a more comprehensive investigation of reaction knockout targets. Full results from this experiment,

including predicted growth rates, methane production rates, and knockout targets, are contained in Table 3 and graphical comparisons of wild type models versus predicted knockout mutants are shown in Figure 9 and Figure 10.

Examining these results revealed that, as a whole, the models could only be optimized for a slight increase in methane production, with maximum of 13.7% flux increase (*M.acetivorans* on pyruvate). These small gains also came at a steep growth cost generally in the range of 55-90%, including a 79.9% decrease in growth yield in the case of the aforementioned *M.acetivorans* model simulated to grow on pyruvate. Several models (both on acetate, *M.barkeri* on H_2/CO_2) achieved particularly modest methane production gains of <3.5%, indicating that regardless of growth yield, the wild type models were already operating very close to optimal methane production, at least in comparison to any knockout strains. In all other cases, the optimal knockout models predicted a maximum methane flux gain of approximately 10%, suggesting that in their current states, the wild type models are also operating fairly close to the optimal methane production that can be attained through reaction knockouts. These results would seem to indicate that reaction knockouts would be an ineffective method for achieving large gains in methane production, particularly if the resulting mutants were expected to grow at rates similar to the wild type strains. Based on this assessment, efforts to increase methane production without substantially sacrificing mutant growth rates may be best served by targeting other methods of strain transformation, such as enzyme addition or changing cofactor preference of native enzymes. It is possible that this is because the organisms are operating near peak efficiency, leaving little room for improvement of methane production by gene deletion. However, as the models are updated in the future when more information regarding these metabolic networks becomes available, changes to the models could potentially reveal currently-unknown pathways and alter the results found here. Thus, although this optimization strategy was relatively unsuccessful in this instance, it could prove to be a valuable tool once I possess better knowledge of my target organisms.

MATERIALS AND METHODS

Models were mathematically represented in MATLAB [7.14.0.739] (The MathWorks Inc., Natick, MA) using the COBRA toolbox [82] defined data structure, which contains lists of model reactions, metabolites, and genes, as well as the relationships between these data types. The central structure of each of these models was the stoichiometric matrix (S), an $m \times n$ matrix where each of the m rows represented one metabolite and each of the n columns represented one reaction. Thus, for any entry of the matrix, S_{ij} , the value of that entry would denote the stoichiometry of metabolite i in reaction j . To generate growth predictions for a model, the S -matrix was used to perform flux balance analysis (FBA)[83], whereby the model was subjected to the constraint:

$$Sv = 0$$

Here, v denoted the $n \times 1$ vector of reaction fluxes, encompassing every reaction in the model. Thus, FBA solutions to a model under this constraint were representative of the steady-state solutions of the model. In general, solutions to this equation are an infinite set of linear combinations upon a convex solution space, the “null space” of the S -matrix. Additional constraints were applied to reduce this solution space, including restricting solutions to optimize to maximize the biomass objective function [84] and setting upper and lower bounds on all reaction fluxes. For the majority of reactions, these bounds were:

$$-1000 < v_{j,rev} < 1000$$

$$0 < v_{j,irrev} < 1000$$

where $v_{j,rev}$ was a reversible reaction and $v_{j,irrev}$ was irreversible. The exceptions to these bounds were substrate uptake reactions, which were given set negative lower bounds to allow metabolites to enter the reaction network, and the ATP maintenance reaction, which was set to an absolute bound of 2 based on the previous models [36,37]. As a final measure for choosing between the many remaining

solutions that satisfied these new constraints, the FBA solver was set to select the solution with the minimum total network flux.

Choices for cofactor reactions were made primarily through use of the BRENDA database [27] by choosing the enzyme type (NAD or NADP preference) based on relative levels of enzyme activity in related organisms. In cases where comparisons to other organisms yielded conflicting results, phylogenetic distances were assessed using trees available through the SEED-viewer [85] to determine the closest relative to the *Methanosarcina* genus. All reactions in the cofactor pairs that were not chosen as the best associations, as well as any other reactions lacking sufficient biological evidence, were removed from the models.

Free energy constraints were incorporated into the existing models by adding an array ($m \times 1$) of standard free energies of metabolite formation (1 M, 25°C, pH=7) to the standard COBRA model structures.

Unlike previous methods that assigned free energy values to every reaction in the metabolic network [29,30], the method used here affected only the substrates and products of a given model. Hence, only metabolites that were allowed to enter and leave the reaction network were subjected to these constraints. Overall model free energy was calculated in the same manner used to calculate reaction free energy by summing free energy of formation for all reaction components according to stoichiometry. This value was calculated and, in accordance with the second law of thermodynamics, was constrained to always be negative:

$$\Delta G_r = \sum_{products} \Delta G_{if} - \sum_{reactants} \Delta G_{if}$$

$$\Delta G_r < 0$$

Here, ΔG_r represented overall model free energy, and ΔG_{if} represented the free energy of formation for some metabolite i . Solutions that did not meet this constraint were not permitted, which effectively reduced the possible solution space for a given model. Sensitivity analyses were performed using 6 different metabolites (CH_4 , CO_2 , H_2O , HCO_3^- , biomass, substrate) as independent variables and the overall model free energy as the dependent variable. The free energy of formation for each of these 6 metabolites was individually altered to +/- 50% of the standard value and the resulting model was simulated with FBA to obtain the model ΔG_r .

Mutant strains for uptake of ethanol and pyruvate were constructed for each *Methanosarcina* model by adding 5 and 3 reactions, respectively, to allow for new substrate to enter the model. GDLS was performed directly in MATLAB with the following specifications: 4 knockouts maximum; 10% of wild-type growth minimum; 2 search paths. All model manipulations, including computing FBA solutions, adding/removing reactions, applying thermodynamic constraints, and simulating knockout optimizations, were performed using the latest version of the COBRA toolbox for the MATLAB environment [82].

CONCLUSIONS

Metabolic models hold tremendous potential as tools to guide metabolic engineering efforts, though this potential greatly hinges on model accuracy with respect to actual biology. To fulfill their immense promise, existing models must not only be iteratively updated to incorporate novel biological knowledge, but also must generate predictions that adhere to the laws of thermodynamics. The work described here, using the most recently published *Methanosarcina* models, iMB745 and iMG746, represents a step towards realizing this goal. Each of these models was updated to reflect the most up-to-date knowledge available, in particular with regard to reducing model redundancy by eliminating reactions with identical primary metabolites that differed only in their utilization of different cofactors (NAD and NADP). Free energy constraints were also added to the models to ensure that model solutions produced negative Gibbs free energy, in accordance with the second law of thermodynamics. These new models were then employed as metabolic engineering tools by introducing new model pathways for uptake of ethanol and pyruvate to simulate mutants being created in lab. Additionally, wild type and mutant models were searched for possible knockout targets to optimize production of methane while maintaining acceptable levels of growth.

Immediate results from these efforts to direct the manipulation of *Methanosarcina* strains were unsuccessful in mirroring the mutant strains transformed in lab or in finding viable knockout targets for maximizing methane production. However, rather than discouraging further experimentation with the models, these results serve as motivation to continue the process of updating these models. My hope is that by extending this process of model improvement, I can both tailor the existing models to match what is seen in lab and use my models to suggest ways in which the organisms might be modified to improve methane production. As systems biology moves forward, there is an ever-increasing drive to increase the speed of model generation by automating more steps of the curation process, but as this work demonstrated, it is chiefly important to ensure that the information contained in these models is

of the utmost accuracy. In this regard, though the metabolic engineering experiments carried out in this instance were not able to seamlessly serve as a blueprint for organism improvement, the updated models themselves, including the added thermodynamic data, were certainly a step in the right direction. Most importantly, my application of measured thermodynamic data to genome-scale models generated reasonable flux and free energy predictions, representing an important step in developing methods for using these data to improve existing models. Moving forward, I will continue my investigation of *M.acetivorans* and *M.barkeri* to augment my knowledge of these organisms so that I can incorporate new information into these models. I also plan to extend my treatment of thermodynamic constraints, principally by adding concentration data to my current method and applying my constraints to dynamic FBA calculations. Ultimately, I hope that my models can become a vehicle to advance both the understanding of *Methanosarcina* metabolism and the widespread application of genome-scale models to metabolic engineering challenges.

REFERENCES

1. Edwards JS, Palsson BO (1999) Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype. *Journal of Biological Chemistry* 274: 17410-17416.
2. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* 5: 93-121.
3. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2008) Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology* 7: 129-143.
4. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Molecular systems biology* 5.
5. Milne CB, Kim PJ, Eddy JA, Price ND (2009) Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnol J* 4: 1653-1670.
6. Xu C, Liu L, Zhang Z, Jin D, Qiu J, et al. (2013) Genome-scale metabolic model in guiding metabolic engineering of microbial improvement. *Applied microbiology and biotechnology* 97: 519-539.
7. Park JH, Lee KH, Kim TY, Lee SY (2007) Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proceedings of the National Academy of Sciences* 104: 7797-7802.
8. Lee KH, Park JH, Kim TY, Kim HU, Lee SY (2007) Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Molecular systems biology* 3.
9. Brochado AR, Matos C, Møller BL, Hansen J, Mortensen UH, et al. (2010) Improved vanillin production in baker's yeast through in silico design. *Microbial cell factories* 9: 84.
10. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in genetics* 24: 133.
11. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature biotechnology* 26: 1135-1145.
12. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57-63.
13. Edwards J, Palsson B (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences* 97: 5528-5533.
14. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology* 3.
15. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, et al. (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. *Molecular systems biology* 7.
16. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* 4: R54.
17. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research* 13: 244-253.
18. Duarte NC, Herrgård MJ, Palsson BØ (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research* 14: 1298-1309.
19. Kuepfer L, Sauer U, Blank LM (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome research* 15: 1421-1430.
20. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, et al. (2008) The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC systems biology* 2: 71.

21. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, et al. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology* 26: 1155-1160.
22. Mo ML, Palsson BØ, Herrgård MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC systems biology* 3: 37.
23. Heavner BD, Smallbone K, Barker B, Mendes P, Walker LP (2012) Yeast 5—an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Systems Biology* 6: 55.
24. Seth-Smith H (2007) A more convenient truth. *Nature Reviews Microbiology* 5: 248-250.
25. Deppenmeier U (2002) The unique biochemistry of methanogenesis. *Prog Nucleic Acid Res Mol Biol* 71: 223-283.
26. Kohler PR, Metcalf WW (2012) Genetic manipulation of *Methanosarcina* spp. *Frontiers in Microbiology* 3.
27. Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Research* 30: 47-49.
28. Ghosh A, Zhao H, Price ND (2011) Genome-scale consequences of cofactor balancing in engineered pentose utilization pathways in *Saccharomyces cerevisiae*. *PLoS One* 6: e27316.
29. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. *Biophys J* 92: 1792-1805.
30. Hoppe A, Hoffmann S, Holzhutter HG (2007) Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst Biol* 1: 23.
31. Beard DA, Liang SD, Qian H (2002) Energy balance for analysis of complex metabolic networks. *Biophys J* 83: 79-86.
32. Fleming RM, Thiele I (2011) von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. *Bioinformatics* 27: 142-143.
33. Müller AC, Bockmayr A (2013) Fast Thermodynamically Constrained Flux Variability Analysis. *Bioinformatics*.
34. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J* 95: 1487-1499.
35. Soh KC, Hatzimanikatis V (2010) Network thermodynamics in the post-genomic era. *Curr Opin Microbiol* 13: 350-357.
36. Benedict MN, Gonnerman MC, Metcalf WW, Price ND (2012) Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon *Methanosarcina acetivorans* C2A. *J Bacteriol* 194: 855-865.
37. Gonnerman MC, Benedict MN, Feist AM, Metcalf WW, Price ND (2013) Genomically and biochemically accurate metabolic reconstruction of *Methanosarcina barkeri* Fusaro, iMG746. *Biotechnology journal*.
38. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 27: 29-34.
39. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research* 32: D438-D442.
40. Weimer P, Zeikus J (1979) Acetate assimilation pathway of *Methanosarcina barkeri*. *Journal of bacteriology* 137: 332-339.
41. Koga Y, Morii H (2007) Biosynthesis of ether-type polar lipids in archaea and evolutionary considerations. *Microbiology and molecular biology reviews* 71: 97-120.
42. Li X, Li Y, Wei D, Li P, Wang L, et al. (2010) Characterization of a broad-range aldehyde dehydrogenase involved in alkane degradation in *Geobacillus thermodenitrificans* NG80-2. *Microbiological research* 165: 706-712.

43. Meng Z, Liu Z, Lou Z, Gong X, Cao Y, et al. (2009) Purification, characterization and crystallization of pyrroline-5-carboxylate reductase from the hyperthermophilic archaeon *Sulfolobus Solfataricus*. Protein expression and purification 64: 125-130.
44. Lazzarini RA, Atkinson DE (1961) A triphosphopyridine nucleotide-specific nitrite reductase from *Escherichia coli*. Journal of Biological Chemistry 236: 3330-3335.
45. Chatwell L, Krojer T, Fidler A, Romisch W, Eisenreich W, et al. (2006) Biosynthesis of riboflavin: structure and properties of 2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5'-phosphate reductase of *Methanocaldococcus jannaschii*. J Mol Biol 359: 1334-1351.
46. Roje S, Chan SY, Kaplan F, Raymond RK, Horne DW, et al. (2002) Metabolic Engineering in Yeast Demonstrates That S-Adenosylmethionine Controls Flux through the Methylenetetrahydrofolate Reductase Reaction in Vivo. Journal of Biological Chemistry 277: 4056-4061.
47. Laupitz R, Hecht S, Amslinger S, Zepeck F, Kaiser J, et al. (2004) Biochemical characterization of *Bacillus subtilis* type II isopentenyl diphosphate isomerase, and phylogenetic distribution of isoprenoid biosynthesis pathways. European Journal of Biochemistry 271: 2658-2669.
48. Yilmaz EI, Çaydasi AK, Özcengiz G (2008) Targeted disruption of homoserine dehydrogenase gene and its effect on cephamycin C production in *Streptomyces clavuligerus*. Journal of industrial microbiology & biotechnology 35: 1-7.
49. Timm S, Nunes-Nesi A, Parnik T, Morgenthal K, Wienkoop S, et al. (2008) A cytosolic pathway for the conversion of hydroxypyruvate to glycerate during photorespiration in *Arabidopsis*. Plant Cell 20: 2848-2859.
50. Pearce FG, Sprissler C, Gerrard JA (2008) Characterization of dihydrodipicolinate reductase from *Thermotoga maritima* reveals evolution of substrate binding kinetics. J Biochem 143: 617-623.
51. Yoneda K, Kawakami R, Tagashira Y, Sakuraba H, Goda S, et al. (2006) The first archaeal L-aspartate dehydrogenase from the hyperthermophile *Archaeoglobus fulgidus*: gene cloning and enzymological characterization. Biochim Biophys Acta 1764: 1087-1093.
52. Summer H (2009) Improved approach for transferring and cultivating *Methanosarcina acetivorans* C2A (DSM 2834). Letters in applied microbiology 48: 786-789.
53. Sowers KR, Nelson MJ, Ferry JG (1984) Growth of acetotrophic, methane-producing bacteria in a pH auxostat. Current Microbiology 11: 227-229.
54. Rother M, Metcalf WW (2004) Anaerobic growth of *Methanosarcina acetivorans* C2A on carbon monoxide: an unusual way of life for a methanogenic archaeon. Proceedings of the National Academy of Sciences of the United States of America 101: 16929-16934.
55. Hutten TJ, Bongaerts HCM, Drift C, Vogels GD (1980) Acetate, methanol and carbon dioxide as substrates for growth of *Methanosarcina barkeri*. Antonie Van Leeuwenhoek 46: 601-610.
56. Bock A-K, Prieger-Kraft A, Schönheit P (1994) Pyruvate — a novel substrate for growth and methane formation in *Methanosarcina barkeri*. Archives of Microbiology 161: 33-46.
57. Bomar M, Knoll K, Widdel F (1985) Fixation of molecular nitrogen by *Methanosarcina barkeri*. FEMS Microbiology Letters 31: 47-55.
58. Smith MR, Mah RA (1978) Growth and methanogenesis by *Methanosarcina* strain 227 on acetate and methanol. Applied and environmental microbiology 36: 870-879.
59. Bock A-K, Schönheit P (1995) Growth of *Methanosarcina barkeri* (Fusaro) under nonmethanogenic conditions by the fermentation of pyruvate to acetate: ATP synthesis via the mechanism of substrate level phosphorylation. Journal of bacteriology 177: 2002-2007.
60. Kulkarni G, Kridelbaugh DM, Guss AM, Metcalf WW (2009) Hydrogen is a preferred intermediate in the energy-conserving electron transport chain of *Methanosarcina barkeri*. Proceedings of the National Academy of Sciences 106: 15915-15920.

61. Welander PV, Metcalf WW (2005) Loss of the mtr operon in *Methanosarcina* blocks growth on methanol, but not methanogenesis, and reveals an unknown methanogenic pathway. *Proc Natl Acad Sci U S A* 102: 10664-10669.
62. Weimer P, Zeikus J (1978) One carbon metabolism in methanogenic bacteria. *Archives of microbiology* 119: 49-57.
63. Sauerwald A, Zhu W, Major TA, Roy H, Palioura S, et al. (2005) RNA-dependent cysteine biosynthesis in archaea. *Science* 307: 1969-1972.
64. Wang S, Huang H, Moll J, Thauer RK (2010) NADP⁺ reduction with reduced ferredoxin and NADP⁺ reduction with NADH are coupled via an electron-bifurcating enzyme complex in *Clostridium kluyveri*. *Journal of bacteriology* 192: 5115-5123.
65. Huang H, Wang S, Moll J, Thauer RK (2012) Electron bifurcation involved in the energy metabolism of the acetogenic bacterium *Moorella thermoacetica* growing on glucose or H₂ plus CO₂. *Journal of Bacteriology* 194: 3689-3699.
66. Buckel W, Thauer RK (2013) Energy conservation via electron bifurcating ferredoxin reduction and proton/Na⁺ translocating ferredoxin oxidation. *Biochim Biophys Acta* 1827: 94-113.
67. Wang S, Huang H, Kahnt J, Thauer RK (2013) A reversible electron-bifurcating ferredoxin-and NAD-dependent [FeFe]-hydrogenase (HydABC) in *Moorella thermoacetica*. *Journal of bacteriology* 195: 1267-1275.
68. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. *BMC bioinformatics* 10: 421.
69. Flamholz A, Noor E, Bar-Even A, Milo R (2012) eQuilibrator—the biochemical thermodynamics calculator. *Nucleic acids research* 40: D770-D775.
70. Alberty RA (2003) *Thermodynamics of biochemical reactions*. Hoboken, NJ: John Wiley & Sons. 397 p.
71. Alberty RA (2006) *Biochemical thermodynamics: Applications of Mathematica*. Hoboken, N.J.: Wiley-Interscience. 464 p.
72. Morowitz HJ (1968) *Energy flow in biology*: Academic Press New York.
73. Roels J (1980) Application of macroscopic principles to microbial metabolism. *Biotechnology and Bioengineering* 22: 2457-2514.
74. Von Stockar U, Liu JS (1999) Does microbial life always feed on negative entropy? Thermodynamic analysis of microbial growth. *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1412: 191-211.
75. Liu JS, Marison IW, Von Stockar U (2001) Microbial growth by a net heat up-take: A calorimetric and thermodynamic study on acetotrophic methanogenesis by *Methanosarcina barkeri*. *Biotechnology and bioengineering* 75: 170-180.
76. Parnaudeau V, Condom N, Oliver R, Cazevielle P, Recous S (2008) Vinasse organic matter quality and mineralization potential, as influenced by raw material, fermentation and concentration processes. *Bioresource technology* 99: 1553-1562.
77. Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *Bmc Bioinformatics* 11: 213.
78. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 84: 647-657.
79. Patil KR, Rocha I, Forster J, Nielsen J (2005) Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 6: 308.
80. Tepper N, Shlomi T (2010) Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics* 26: 536-543.
81. Lun DS, Rockwell G, Guido NJ, Baym M, Kelner JA, et al. (2009) Large-scale identification of genetic design strategies using local search. *Mol Syst Biol* 5: 296.

82. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 6: 1290-1307.
83. Varma A, Palsson BO (1994) Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/technology* 12.
84. Feist AM, Palsson BO (2010) The biomass objective function. *Current opinion in microbiology* 13: 344.
85. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research* 33: 5691-5702.

APPENDIX A – FIGURES

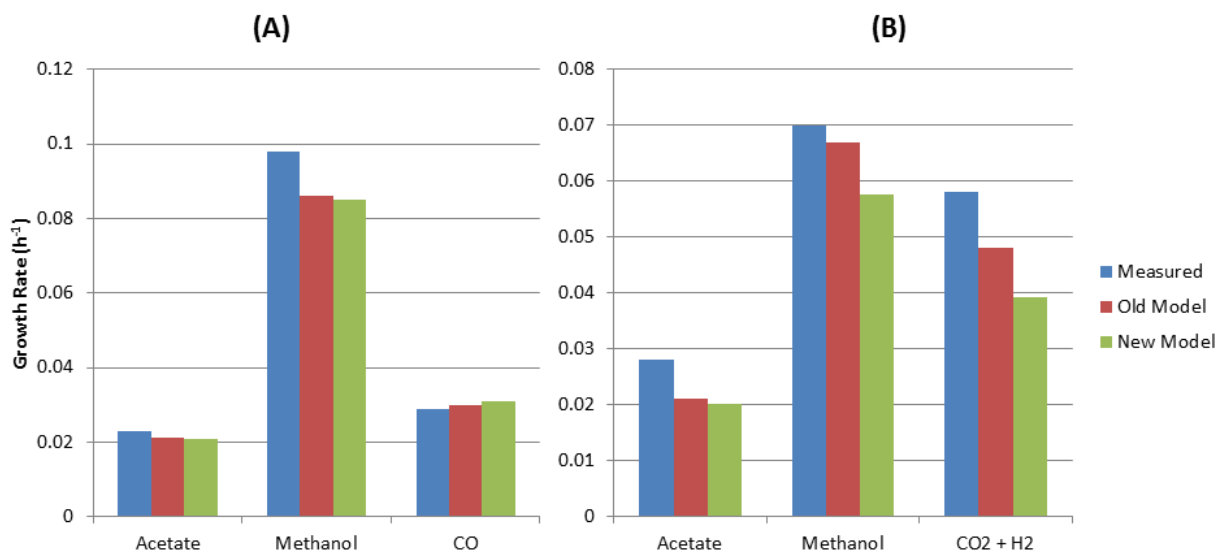


Figure 1: Graphs show growth comparisons between measured experimental rates and rates predicted by model simulations, using both the original model (“Old Model”) and the updated model (“New Model”). Plot (A) shows these comparisons in *M. acetivorans* and Plot (B) shows the comparisons in *M. barkeri*

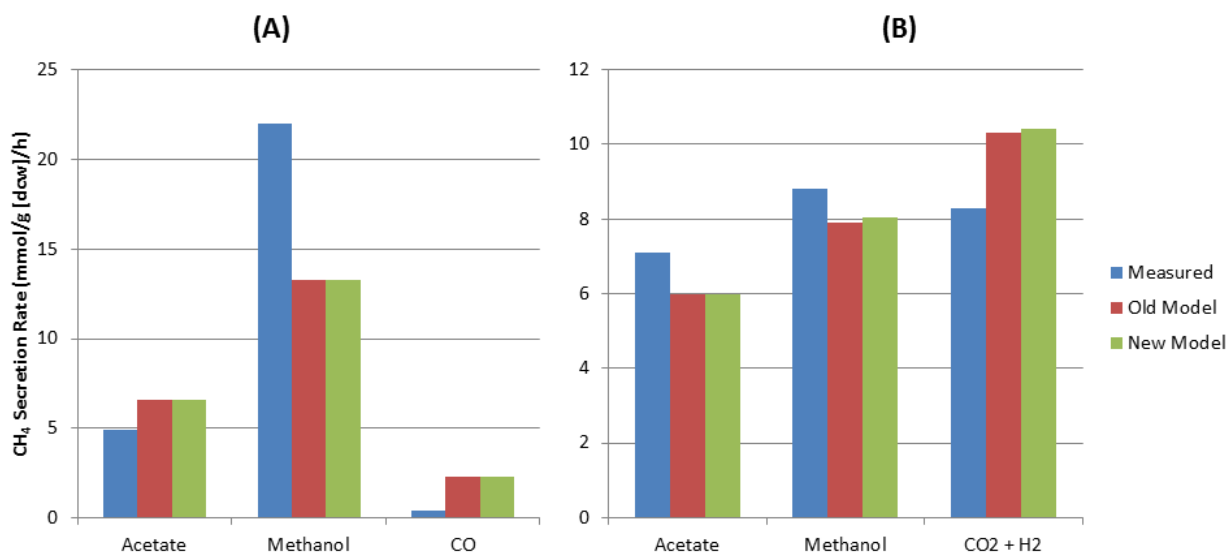


Figure 2: Graphs show methane secretion comparisons between measured experimental rates and rates predicted by model simulations, using both the original model (“Old Model”) and the updated model (“New Model”). Plot (A) shows these comparisons in *M. acetivorans* and Plot (B) shows the comparisons in *M. barkeri*

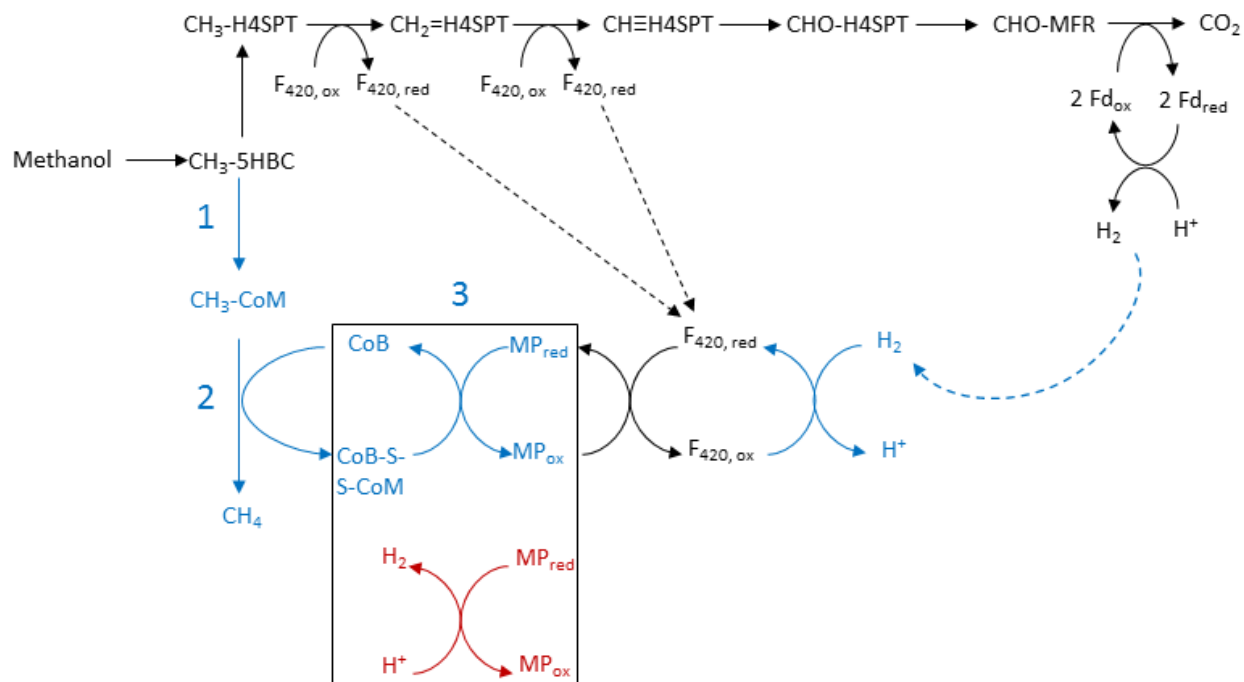
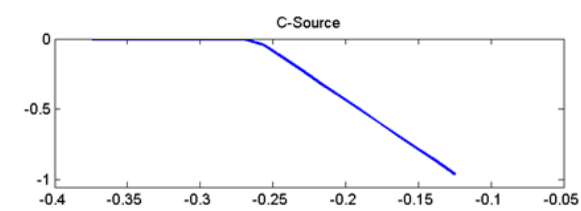
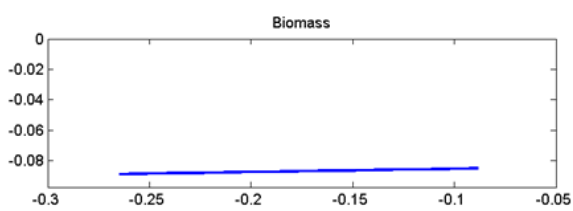
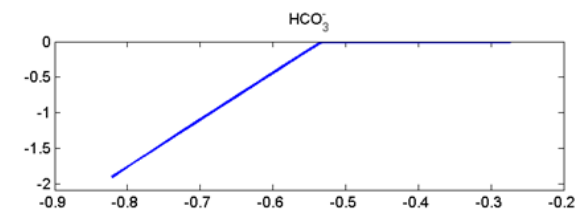
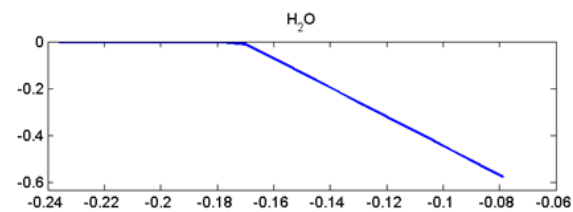
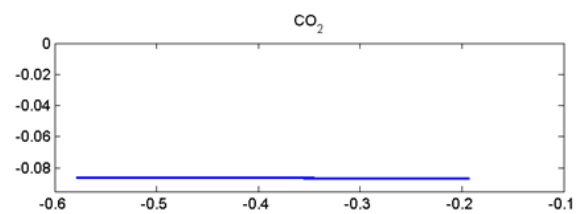
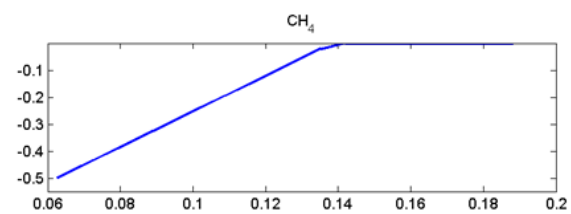
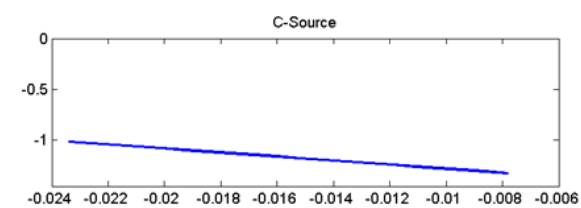
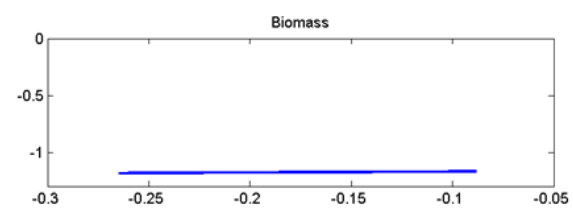
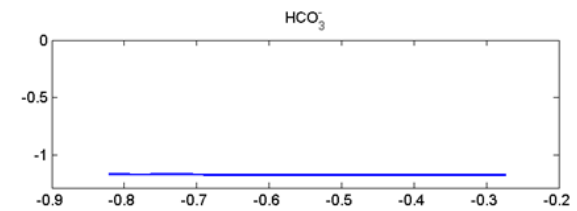
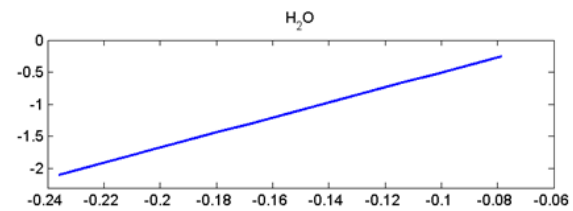
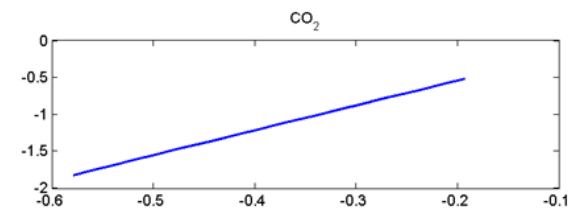
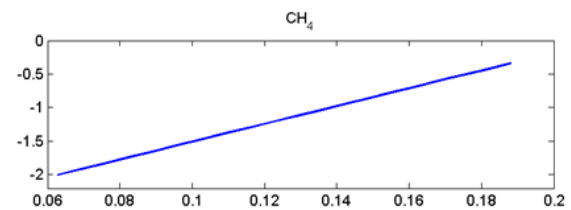


Figure 3: Reaction pathway diagram showing the possible model pathways for catabolism of methanol in *M.barkeri*. Numbered reactions (1,2,3) were predicted as non-essential for growth *in silico*. Pathways in blue denote reactions that were predicted as active in the wild-type models, but inactive when any numbered reaction was knocked out *in silico*. Pathways in red denote reactions that were predicted as inactive in the wild-type models, but active when any numbered reaction was knocked out *in silico*. Pathways in black denote reactions that were predicted as active regardless of knockouts. The boxed reactions demonstrate the switch that occurs between the wild-type and knockout model predictions, particularly pertaining to how the models oxidize methanophenazine (MP). Use of the reaction in red produces hydrogen gas that is predicted as one of the end products (along with CO₂) of the knockout pathway. (Abbreviations used: F420 – coenzyme ferredoxin 420; Fd – ferredoxin; H4SPT – tetrahydrosarcinapterin; 5HBC - 5-hydroxybenzimidazolylcob(I)amide; MFR – methanofuran; MP – methanophenazine; CoA – coenzyme A; CoB – coenzyme B; CoM – coenzyme M; CoB-S-S-CoM – heterodisulfide; Pi – phosphate; red/ox – denote reduced or oxidized form of a cofactor species)

(A)



(B)



(C)

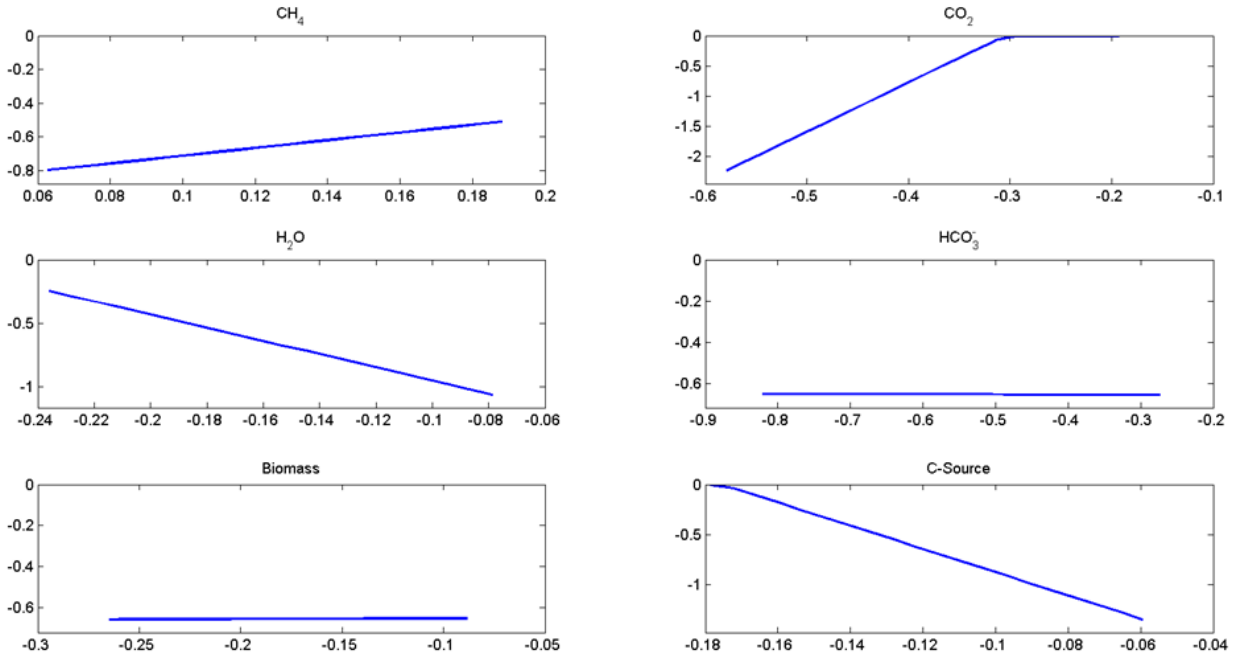
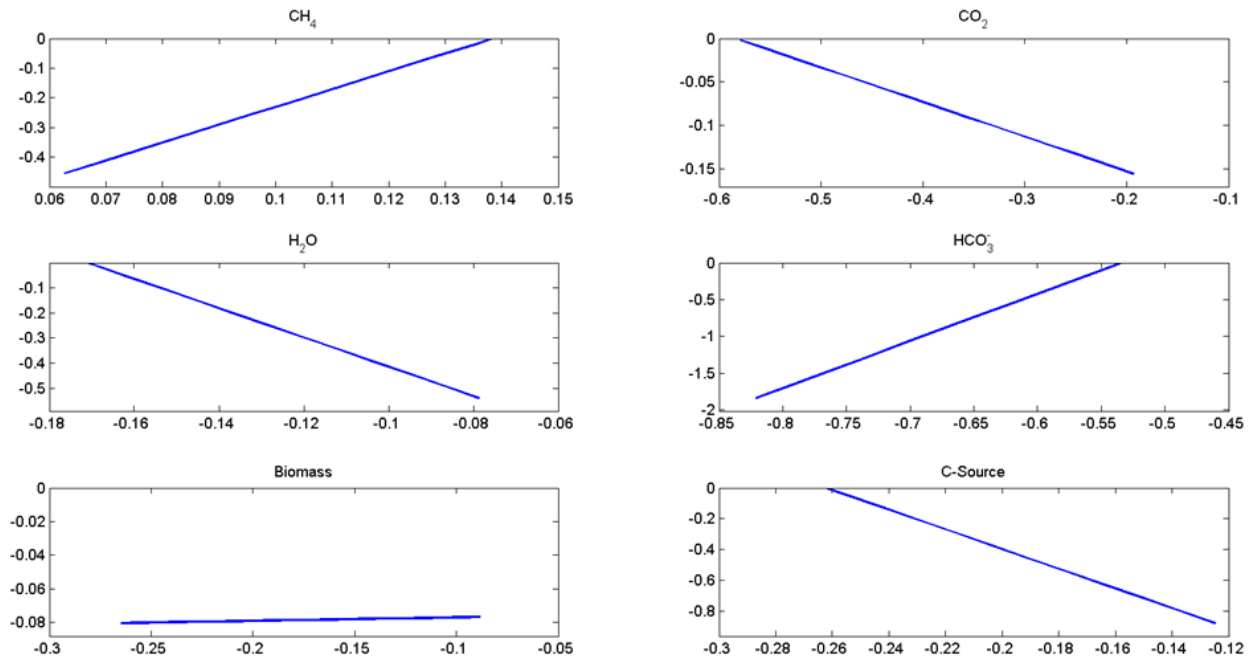
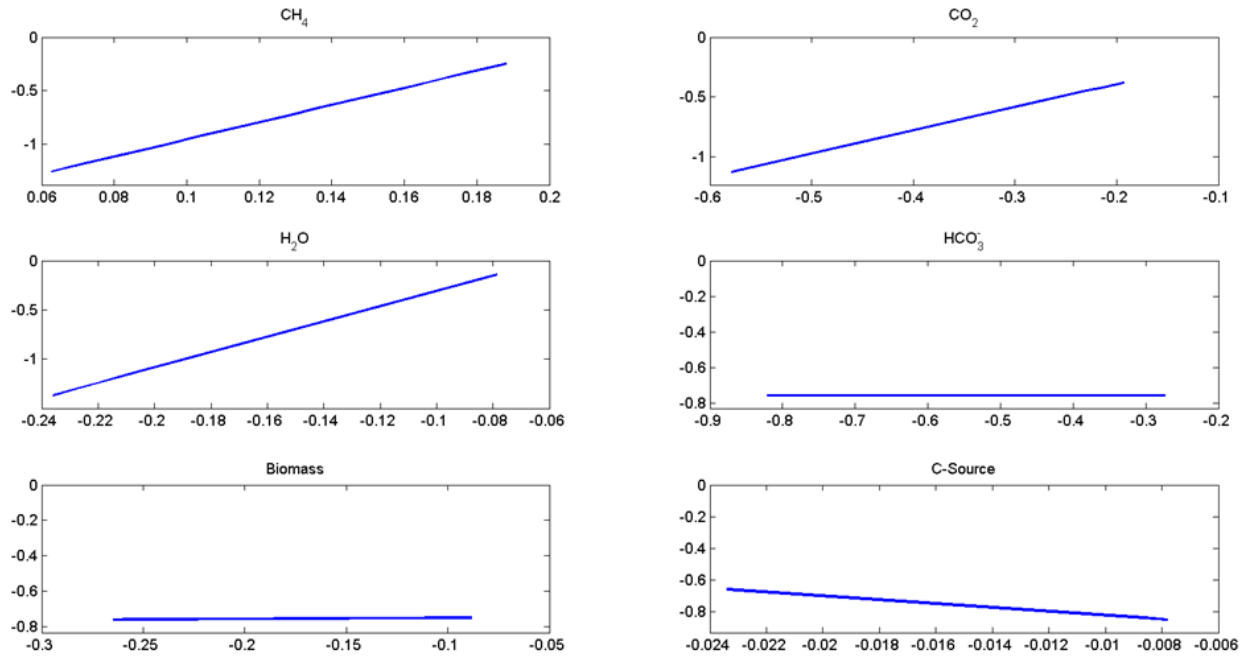


Figure 4(A-C): Free energy sensitivity analyses for *M. acetivorans* model growing on (A) acetate, (B) methanol, (C) carbon monoxide. For each plot, the Y-axis is the overall free energy (ΔG_r) flux [kJ/GDW/h] predicted by simulating the model for a given value of free energy of formation (ΔG_f) [kJ/mmol] for the component denoted. “C-source” represents the main carbon substrate for a given simulation, in this case corresponding to either (A) acetate, (B) methanol, or (C) carbon monoxide.

(A)



(B)



(C)

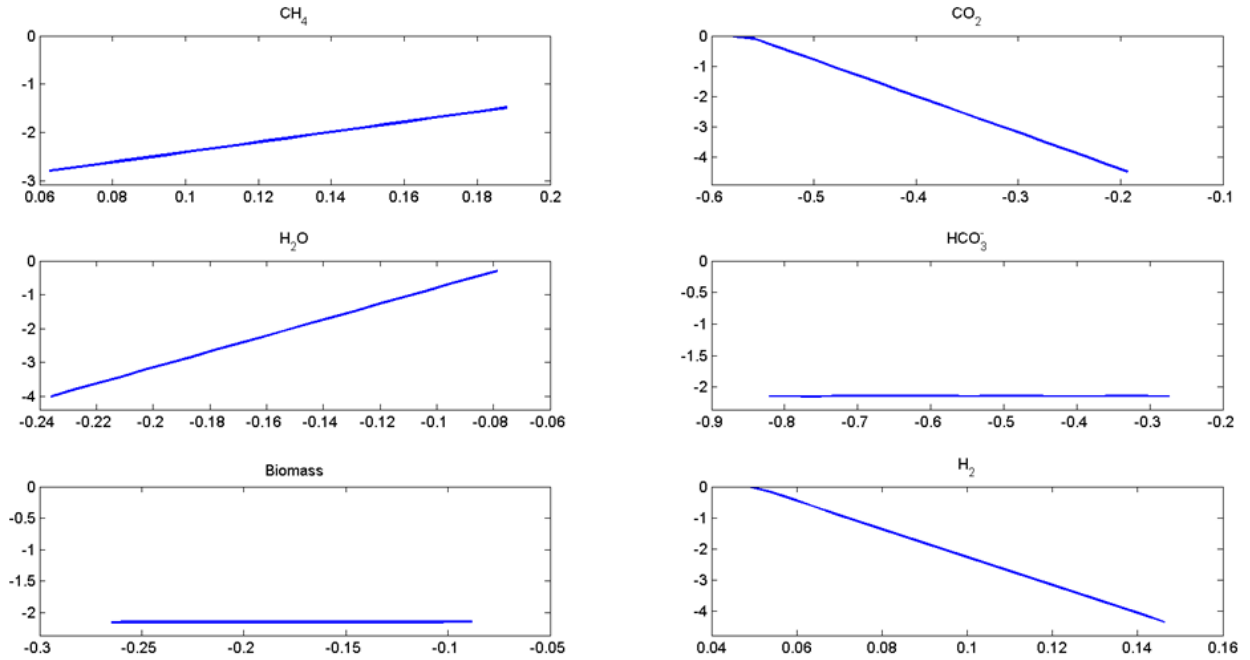


Figure 5(A-C): Free energy sensitivity analyses for *M. barkeri* model growing on (A) acetate, (B) methanol, (C) carbon dioxide plus hydrogen. For each plot, the Y-axis is the overall free energy (ΔG_r) flux [kJ/GDW/h] predicted by simulating the model for a given value of free energy of formation (ΔG_f) [kJ/mmol] for the component denoted. “C-source” represents the main carbon substrate for a given simulation, in this case corresponding to either (A) acetate or (B) methanol.

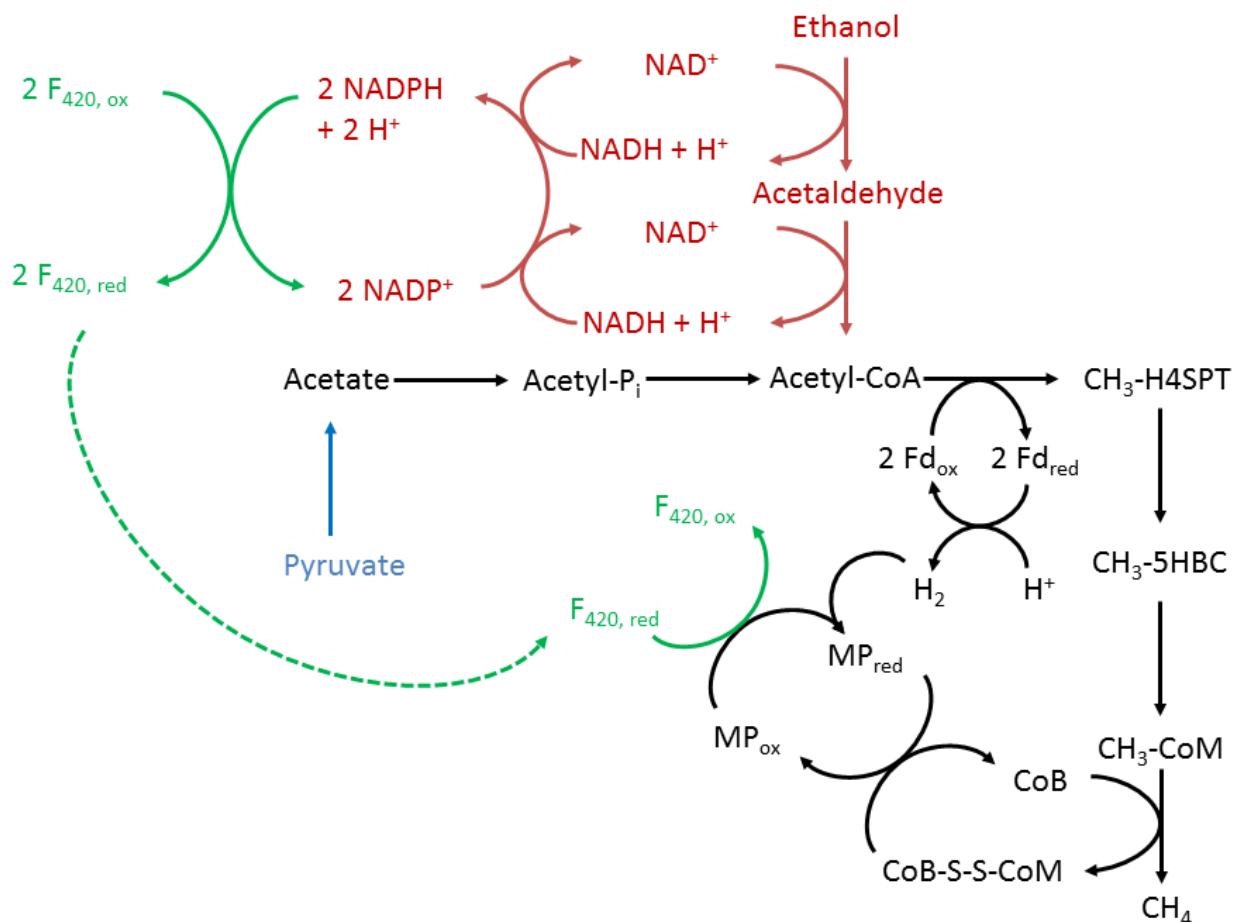


Figure 6: Reaction pathway diagram showing the model pathways for methanogenesis using acetate, ethanol, and pyruvate. Pathways in black indicate that these reactions are used in the wild-type models for catabolism of acetate. Pathways in blue indicate reactions added to the models for catabolism of pyruvate. Pathways in red indicate reactions added to the models for catabolism of ethanol. Pathways in green indicate reactions present in wild-type models that are used for catabolism of ethanol, but not the other two substrates. Not shown here is the remaining portion of the methanogenesis pathway present in the wild-type models. This other pathway is also used for methane production during catabolism of ethanol, but is not shown here because it is not part of aceticlastic methanogenesis. Abbreviations used: F₄₂₀ – coenzyme ferredoxin 420; Fd – ferredoxin; H₄SPT – tetrahydrosarcinapterin; 5HBC - 5-hydroxybenzimidazolylcob(I)amide; MP – methanophenazine; CoA – coenzyme A; CoB – coenzyme B; CoM – coenzyme M; CoB-S-S-CoM – heterodisulfide; Pi – phosphate; red/ox – denote reduced or oxidized form of a cofactor species

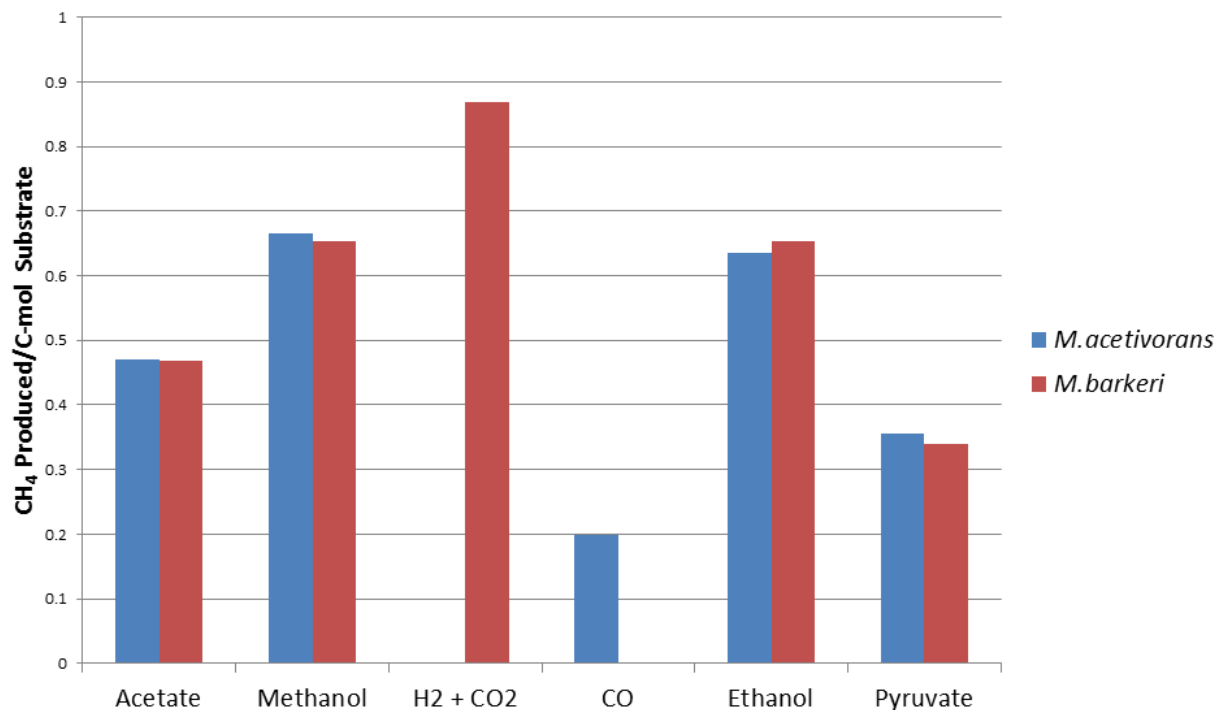


Figure 7: Model predictions for methane production on a per-carbon mole of substrate basis.

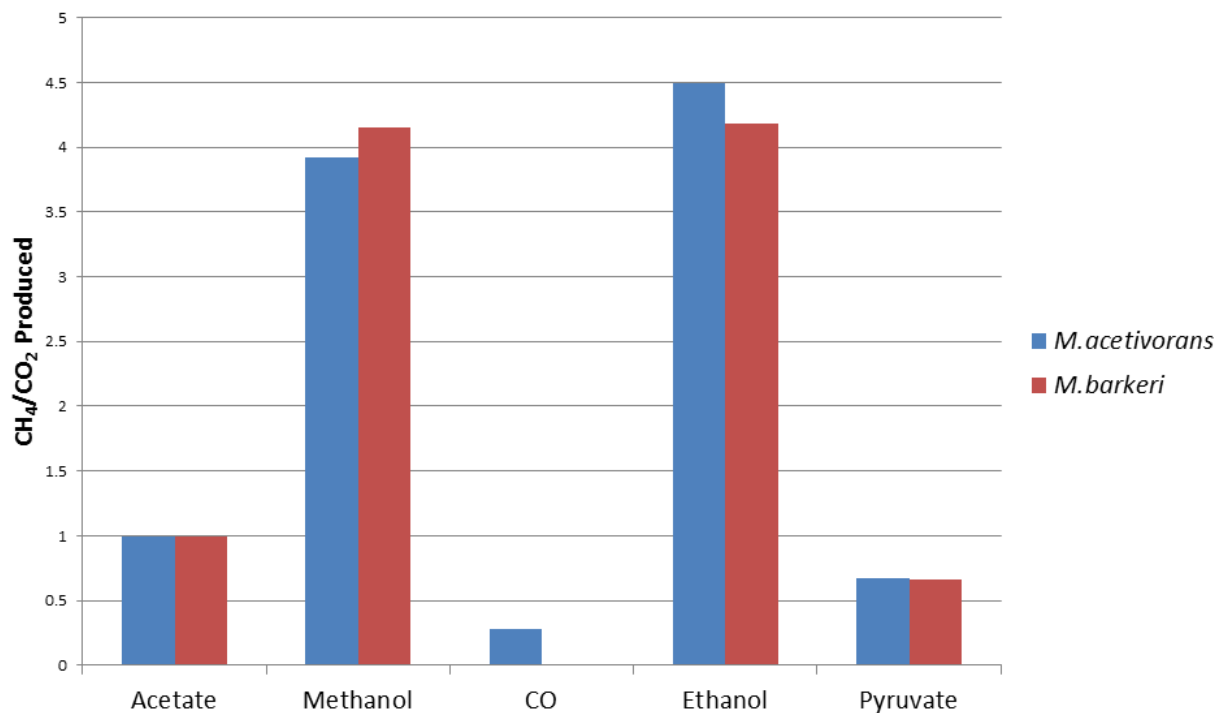


Figure 8: Model predictions for methane production per mole of carbon dioxide produced. Here, “CO₂” encompasses not only carbon dioxide itself, but also bicarbonate (HCO₃⁻) produced in model simulations. This is to account for the fact that these 2 substances can freely interconvert, a fact not well-represented by model flux distributions. Predictions for *M. barkeri* on a mixture of CO₂+H₂ are not shown here because for that model, there is no net CO₂ produced.

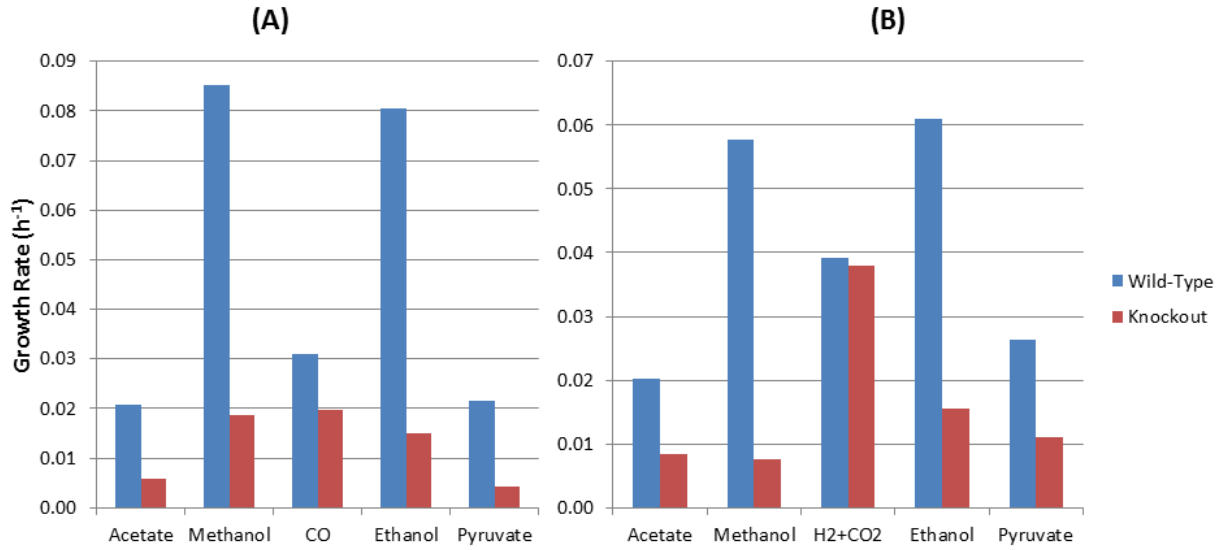


Figure 9: Predicted growth rates in GDLS knockout optimizations for (A) *M. acetivorans* and (B) *M. barkeri*. “Wild-Type” indicates models with no knockouts; “Knockout” indicates models with the knockouts predicted by the GLDS algorithm

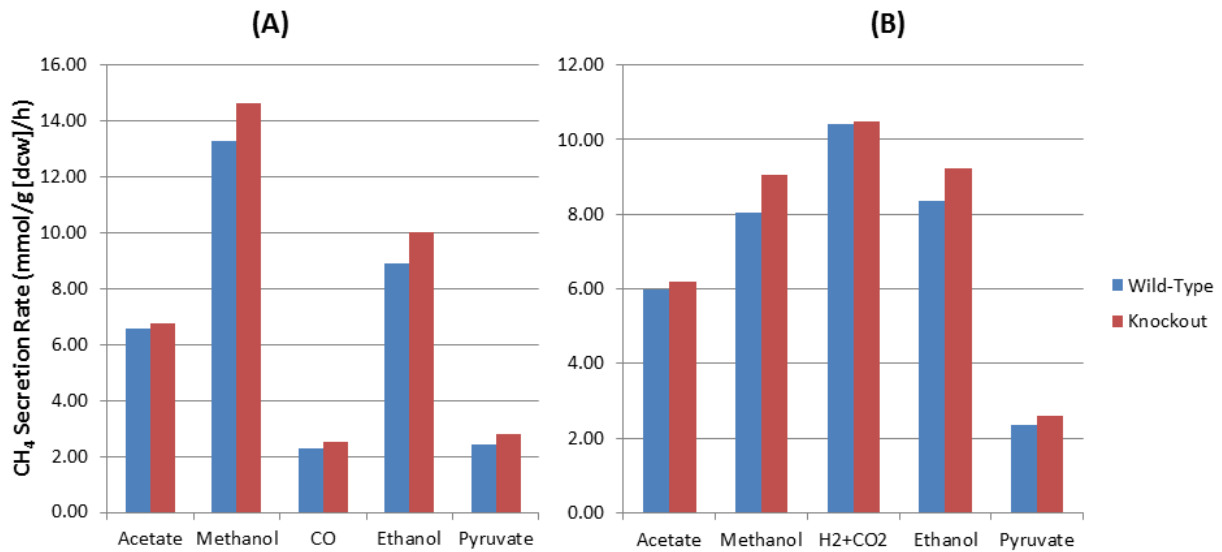


Figure 10: Predicted methane production rates in GDLS knockout optimizations for (A) *M. acetivorans* and (B) *M. barkeri*. “Wild-Type” indicates models with no knockouts; “Knockout” indicates models with the knockouts predicted by the GLDS algorithm

APPENDIX B - TABLES

	<i>Methanosarcina barkeri</i> (iMG746)		<i>Methanosarcina acetivorans</i> (iMB745)	
Model Version	<u>Original</u>	<u>New</u>	<u>Original</u>	<u>New</u>
Genes	750	751	745	747
Reactions	816	800	825	804
Metabolites	718	716	715	713

Table 1: A comparison of the original versions of the *Methanosarcina* models, before any updates, and the new versions, which encapsulate all of the updates described in this work.

Organism	Substrate	Essential w/o ΔG	Essential w/ ΔG	ΔG_r (kJ/g[dcw]/h)
<i>M.acetivorans</i>	Acetate	340	340	-0.086
<i>M.acetivorans</i>	Methanol	343	343	-1.171
<i>M.acetivorans</i>	CO	336	336	-0.653
<i>M.barkeri</i>	Acetate	333	333	-0.078
<i>M.barkeri</i>	Methanol	326	331	-0.752
<i>M.barkeri</i>	CO ₂ +H ₂	337	337	-2.139

Table 2: A list of reactions predicted as essential by performing single reaction knockout tests on each model, both with and without constraints on free energy. As shown, only *M.barkeri* growing on methanol had changes in reaction essentiality. Also shown here is the predicted overall model free energy (ΔG_r) for each simulation.

Model	Substrate	Predicted Knockouts	Growth Rate (h ⁻¹)			CH ₄ Rate (mmol/g[dcw]/h)		
			WT	KO	% Dec.	WT	KO	% Inc.
<i>M.acetivorans</i>	Acetate	2	0.0208	0.0057	72.4%	6.58	6.78	2.9%
<i>M.acetivorans</i>	Methanol	2	0.0852	0.0186	78.1%	13.30	14.63	10.0%
<i>M.acetivorans</i>	CO	4	0.0309	0.0199	35.7%	2.30	2.51	9.4%
<i>M.acetivorans</i>	Ethanol	3	0.0804	0.0151	81.3%	8.90	10.05	13.0%
<i>M.acetivorans</i>	Pyruvate	2	0.0216	0.0044	79.9%	2.45	2.79	13.7%
<i>M.barkeri</i>	Acetate	1	0.0202	0.0084	58.4%	5.99	6.19	3.4%
<i>M.barkeri</i>	Methanol	2	0.0577	0.0077	86.7%	8.05	9.07	12.7%
<i>M.barkeri</i>	CO ₂ +H ₂	1	0.0391	0.0380	2.9%	10.43	10.47	0.4%
<i>M.barkeri</i>	Ethanol	2	0.0611	0.0156	74.4%	8.35	9.21	10.3%
<i>M.barkeri</i>	Pyruvate	2	0.0264	0.0110	58.3%	2.35	2.61	11.2%

Table 3: Results from GDLS knockout optimization experiments. WT denotes “wild-type”, the full model without any knockouts. KO denotes “knockout”, the strain with the described number of knockouts predicted by the GDLS algorithm. % Dec. and % Inc. represent the growth rate decrease and methane production rate increase, respectively, caused by incorporating the knockouts.