Exploring Hydrogenotrophic Methanogenesis: A Genome Scale Metabolic Reconstruction of *Methanococcus maripaludis S2*

**Matthew A. Richards1,2, Juan Zhang3, Thomas J. Lie4, John A. Leigh4\* , Nathan D. Price2\***

1Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL USA

2Institute for Systems Biology, Seattle, WA, USA

3Jiangnan University, China

4Department of Microbiology, University of Washington, Seattle, WA, USA

\*Corresponding authors: [nprice@systemsbiology.org](mailto:nprice@systemsbiology.org), [leighj@u.washington.edu](mailto:leighj@u.washington.edu)

# Abstract

Methanogenic archaea are a crucial part of the global carbon cycle, producing about 1 billion tons of methane each year. We have constructed a genome-scale metabolic model for the model archaeon *Methanococcus maripaludis S2* that is the first model to accurately portray hydrogenotrophic methanogenesis*.* Our model contains the complete Wolfe cycle, the central catabolic pathway of our organism, including the crucial electron bifurcation step that completes the cycle. This model serves as a knowledgebase of *M. maripaludis* metabolism and provides a platform for generating hypotheses for strain designs.

# Introduction

Methane is a greenhouse gas that plays a critical role in the global carbon cycle and is 21 times more potent than carbon dioxide (1). However, it can also be used as a “green” energy fuel source that burns relatively cleanly compared to conventional fuels such as coal (2). Furthermore, there is growing interest in biologically converting of methane to high energy density liquid fuels (3). The greatest biological contributor of methane gas is a specialized group of anaerobic microorganisms known as methanogenic Archaea or methanogens, which produce about 1 Gt of methane gas per year through their metabolic activity (4). Methanogens employ unusual metabolic pathways with biologically unique cofactors and enzymes to produce methane from 1-C or 2-C carbon sources (5, 6). Studying methanogenesis gives us a window through which we can better understand these organisms and learn ways to metabolically engineer them to enhance methane production or produce commodity chemicals.

Though diverse, methanogens can be separated into two main groups based on the presence or absence of cytochromes (4). Metabolically, the cytochrome-lacking methanogens have a narrow substrate range limited to only H2 or formate as electron donors that reduce CO2 to methane; hence they are known as hydrogenotrophic methanogens {ref}. By contrast, methanogens with cytochromes utilize acetate and methylated compounds for methanogenic growth and only a minority can also use H2 and CO2 {ref}; they are known as methylotrophic methanogens. Although both groups have similar central pathways of CO2 reduction, they vary in their energy conservation approaches due to disparities in the number of coupling sites for generating ion gradients. For example, in the last step of methylotrophic methanogenesis, H2 reduces a heterodisulfide as part of a membrane bound hydrogenase-heterodisulfide complex that conserves energy via generation of a proton gradient (4, 7). The same complex in hydrogenotrophs, however, is cytoplasmic and thus does not provide a coupling site for the generation of a proton gradient (8). Instead, this group conserves energy via flavin linked electron bifurcation, which links the final heterodisulfide reduction step to the first step of CO2 reduction by using the latter step as an electron sink for the former step (9). The resulting circular pathway is known as the Wolfe Cycle (10) and is key not only for understanding hydrogenotrophic methanogenesis, but also as prime example of electron bifurcation, a non-traditional mechanism that has emerged in a variety of organisms (11).

*Methanococcus maripaludis* is an anaerobic hydrogenotrophic methanogen originally isolated from a salt marsh in 1983 (12). Its genome is comprised of only 1722 protein coding genes (13) and it has a relatively simple metabolism, using electrons from formate or hydrogen to reduce carbon dioxide to methane and build an ion motive force that drives ATP synthesis {SOURCE NEEDED}. *M. maripaludis* grows rapidly with a generation time of 2 hours (ref) and is genetically tractable (14), making it an ideal candidate for studying methanogenesis and for creating novel strain designs that produce industrially relevant products. Additionally, the establishment of anaerobic chemostats (ref) allows for the ability to make large scale steady state studies (ref) or for production of heterologous proteins (ref??).

Genome scale metabolic reconstructions are powerful tools that can aid in mapping metabolic pathways and serve as platforms for generating hypothetical strain designs. Additionally, they serve as organism knowledge bases and can be made into models that predict growth phenotypes for potential wet lab experiments by simulating steady-state growth via flux balance analysis (FBA) {SOURCE NEEDED}. Their valuable ability to model metabolism has helped guide metabolic engineering efforts for production of industrial biochemical in multiple organisms (15). Constructing a genome scale metabolic reconstruction for *M. maripaludis* would therefore have promise for better understanding methanogenesis and for guiding metabolic engineering efforts that harness the unique metabolism of our hydrogenotrophic methanogen.

*M. maripaludis* has already been metabolically reconstructed as part of a mutualistic community model with *D. vulgaris* (16)and as an isolate (17). In the former case, the model of *M. maripaludis* represented only core metabolism and was used primarily to investigate interactions between the two different species rather than probe the depths of one organism’s metabolism (16). The latter case was the first genome-scale metabolic reconstruction of *M. maripaludis S2*(17), an important step towards understanding *M. maripaludis* metabolism.

The genome of the *M. maripaludis S2* has been sequenced (ref) and a transcriptome (ref) as well as its proteome (18) has been published. Together with the metabolic model (ref), we will have a complete approach to understanding the unique biology of this organism as be able to study selected pathways through biochemical (ref) or genetic approaches (ref). In our metabolic model

iMR533, we have sought to include the crucial Wolfe cycle (10) which includes the electron bifurcation step in our metabolic reconstruction (9)(11). This helps explain the ability for this organism to grow despite the lack of membrane related energy coupling sites (ref). Other improvements include eliminating methanophenazine utilization, which is known not to occur in hydrogenotrophic methanogens (4), replacing sulfate with sulfide, the primary sulfur source in *M. maripaludis* {SOURCE NEEDED}, and adding biosynthesis pathways for all of the unique coenzymes involved in methanogenesis (19). We further increased genome coverage and homology by employing likelihood-based gapfilling, a technique that fills reaction gaps based on probability rather than on parsimony (20). Our reconstruction is the first manually-curated genome scale reconstruction to employ likelihood based gap filling.

As an additional measure, we have included free energies of formation to our reconstruction. This is an especially salient consideration for methanogenic archaea, which can grow close to the thermodynamic limits that support life (21). A well-established method of applying thermodynamic constraints involves applying the second law of thermodynamics to metabolic models to restrict reaction directionalities in the direction of negative free energy generation (22, 23). Rather than apply thermodynamic constraints to every metabolic reaction, we devised a novel method of adding free energies of formation to our model to predict the overall free energy generated during steady state growth based solely on standard free energies and concentrations of external metabolites. Our method allows us to evaluate and restrict our FBA solutions to only those that have a negative net free energy of formation, but does not bound the solution space by placing reaction directionality constraints on internal reactions where metabolite concentrations are difficult to measure. In combining these novel thermodynamic considerations with stoichiometric information, iMR533 provides a means to predict energetically feasible strain designs, enhancing our metabolic engineering capabilities with *M. maripaludis.*

# Methods

## Genome Scale Reconstruction Procedure

The process of genome scale metabolic network reconstruction has been reviewed previously (24) and begins with annotating an organism genome using gene-protein-reaction (GPR) relationships stored in a reaction database. Several databases are available for this purpose (25–27); we chose the Department of Energy Systems Biology Knowledgebase (Kbase; www.kbase.us) interface, a suite of tools that includes the Model SEED reaction database (27). We created our first draft reconstruction using the stored Kbase genome for *M. maripaludis S2* (kb|g.575) and the automated reconstruction method (“Reconstruct Genome-scale Metabolic Model”). For this initial reconstruction, we chose the default “Gram Negative Biomass” definition of cell mass and gapfilled using likelihood based gapfilling (method currently not available through Kbase Narrative Interface). This yielded the first full draft of the metabolic reconstruction that could be converted to a simulateable model.

We expanded and refined the model by manually adding information from literature sources. In cases where reactions from literature were part of the Model SEED database, we labeled the reactions using SEED identifiers, names, subsystems, and EC numbers. For other cases where we discovered reactions that were not part of the Model SEED, we created our own unique reaction identifiers and names, then added subsystem information based on our knowledge of the metabolic network. We also adhered to SEED identifiers, names, formulas, and charges for metabolites whenever possible and had very few cases where we specified our own values. Metabolites were compartmentalized using standard tags for cytosol (“c0”) and extracellular (“e0”) compartments. These tags additionally identify *M. maripaludis* as “Organism 0” in the possible future case where we could consider adding other organisms to create a community model. Exchange reactions used for introducing metabolites to the extracellular compartment were standardized in “EX\_{metabolite ID}\_e0” format. Comprehensive information on the reactions, metabolites, and genes in our reconstruction can be found in Supplementary Materials.

## Model Simulations with Flux Balance Analysis

A genome scale metabolic reconstruction and genome scale models are often mentioned in the same breath, but they differ in that the former is a network whereas the latter is a simulatable structure {reference}. A model does not, for instance, encompass the dead end reactions and metabolites present in the reconstruction because these components of the network by definition cannot carry flux {reference?}. To make rigorous quantitative growth predictions, a genome scale metabolic reconstruction must be converted to a simulatable model. Reactions and their participating metabolites in the metabolic network are connected via the stoichiometric matrix (S), which contains the stoichiometric coefficients for each metabolite (row) in each reaction (column). The S-matrix, which was all present in the reconstruction, is converted to a model via the principles of metabolite mass conservation by recognizing that time-dependent accumulation of metabolites in the system (b) is equivalent to the product of the S-matrix and the vector of reaction fluxes (v)

|  |  |
| --- | --- |
|  | Equation 1 |

In flux balance analysis (FBA), we further simplify this differential system by assuming our organism is in steady state growth, thus b=0 and the system is linear {reference; Nathan paper}. This assumption bounds our model system to a large solution space that can further be constrained by applying upper and lower bounds to each reaction flux:

|  |  |
| --- | --- |
|  | Equation 2 |

To find a feasible flux distribution within this solution space, we solved our model by optimizing for maximum biomass yield, or the biomass objective function {reference}. All model simulations were performed using the COBRA toolbox 2.0 (29) in MATLAB [7.14.0.739] (The MathWorks Inc., Natick, MA). We

It is vital that metabolic reconstruction efforts remain transparent and that resulting models be easily usable by other groups (28). An important part of building a model is ensuring that it is easy to simulate. To assist in this process, we have created functions specifically written to simulate model growth, make systematic changes to the model, and replicate all of the results reported in this manuscript. These codes are included in supplementary data and current versions can be found on Github (https://github.com/marichards/methanococcus)

maxGrowthOn\_\_ codes

simulateKOPanel code

switchToFormate code

switchToSpecificFerredoxins code

Others available on GitHub

## Gene Knockout Phenotype Simulations

We performed gene knockout simulations using the “deleteModelGenes.m” function in the COBRA Toolbox 2.0 (29) for MATLAB. We incorporated knockout predictions across six previous publications detailing wet lab hydrogenase knockout experiments (30–35). Experimental gene knockout data phenotypes were pFor gene knockouts, we evaluated our model’s performance using the Matthews correlation coefficient (MCC)(36). The formula for the MCC is given as follows:

## Thermodynamic Calculations

In a novel approach, we added free energies of formation from the Equilibrator database (37) to all exchanges reactions for which these values could be calculated via the group contribution method (38). We expanded the standard structure of the model to include free energy values for a standard 1 mM concentration, temperature of 25 C, pressure of 1 bar, pH of 7, and ionic strength of 0.1 M. Our “optimizeThermoModel.m” code (See Supplementary Materials) allows us to specify concentrations of exchange metabolites, alter these standard free energies to reflect differences in concentration, and estimate the overall free energy generated during growth by multiplying each metabolite’s free energy of formation by its exchange flux.

## Dry Cell Weight Measurements

We grew *M. maripaludis S2* cells in a chemically defined media (See Supplementary Materials) using a 1-L chemostat under anaerobic conditions. We operated the chemostat in continuous mode with gas flows of 110 L/h H2, 15 L/h CO2, 15 L/h N2, and 15 L/h H2S, with a dilution rate of 0.0833 h-1.

We measured dry cell weight via cell filtering. 100 mL aliquots of cells in media were filtered through 25 nM pore filters to remove all non-cellular components. The wet filters were then dried in a 50 degree oven and their weight was measured daily until it stabilized, giving the final dry cell weight.

# Results

## Reconstruction Statistics

The basic statistics for iMR544 are displayed in Table 1 compared to iMM518, the previously published genome-scale metabolic reconstruction (17). Although our reconstruction was generated independently of the iMM518 reconstruction, many of the differences in models represent an expansion of the reaction network with regard to gene coverage and biochemistry. We slightly increased gene coverage by 26 genes, but because iMR544 was not built directly from the previous reconstruction and shares only 434 of its genes, we added 110 novel genes to our reconstruction and chose to exclude 84 genes for which we did not find sufficient literature evidence. Despite a relatively modest increase in gene coverage, our network contains over 100 more gene-associated reactions and, as a result, over 90% of the internal reactions in our reconstruction are associated with at least one gene. We suspect that a major reason for our increase in gene-associated reactions was our use of likelihood based gapfilling, which resulted in the automated addition of 66 genes to our reconstruction before we began to curate it manually. Furthermore, we relied heavily on biochemical knowledge from literature sources, particularly regarding recently-elucidated biosynthesis pathways that were not available in annotation databases. Combined, our use of maximum likelihood gapfilling and reliance on published literature sources directly resulted in more consistent ties to gene homology.

Notably, when compared to iMM518 our reconstruction has 75 more internal reactions and over 100 more of both internal metabolites and dead-end metabolites that cannot be synthesized or consumed. Thus, although our reconstruction contains more metabolites and reactions, the size of the mathematical model generated by removing these dead ends is comparable to the previous model. These metabolites and their reactions may not be part of our mathematical model, but we have included them in our reconstruction because they are all gene-associated; our reconstruction contains are no dead-end internal reactions lacking genes. Thus, we have evidence that each of these reactions and the involved metabolites should be involved in metabolism, but we have not yet elucidated full synthesis or consumption pathways. They represent excellent candidates for further exploration of *M. maripaludis* metabolism, particularly as this reconstruction is updated and expanded in the future.

Conversely, our reconstruction contains 85 internal reactions that lack genes, many of which were added during the automated gap filling phase but some of which were added manually. All of our reactions are annotated with subsystems from Model SEED database, allowing us to assess where each reaction, including those without genes, fits into metabolism. Figure 1 shows a breakdown of these reactions lacking genes, where the subsystems have been manually grouped into broader categories (e.g. “Amino Acid Biosynthesis” instead of “Glycine Biosynthesis”). The largest group of these reactions is the “Unique Coenzyme Syntheses”, which includes reactions that synthesize coenzyme M, coenzyme B, tetrahydromethanopterin (H4MPT), methanofuran, coenzyme F420, and coenzyme F430. Although these 24 reactions lack genes, all of them were added manually as hypothetical steps to complete essential biosynthetic pathways and are based on information from biochemical literature. These are distinct from, for example, the 11 reactions encompassed by “Vitamin and Cofactor Synthesis” that were all added to fill biosynthesis gaps but have no supporting literature evidence. In total, of the 85 reactions lacking genes 46 (54%) are gapfilling reactions and the remaining 39 (46%) are classified as “hypothetical”, with no known genes currently but with literature evidence pointing to their inclusion in the reconstruction. We expect that as biochemical research pushes forward, many reactions in this latter group will become gene-associated whereas the gapfilling reactions, much like dead end reactions and metabolites, point us toward areas of metabolism that are poorly understood in our organism and require more in depth investigation.

Furthermore, our use of likelihood based gap filling not only directs us toward unknown portions of metabolism, but also lets us assign likelihood scores for many of the reactions in the reconstruction. These scores provide a novel metric of evaluating our confidence in the model because of these reactions is annotated with a confidence score ranging from 0-1. We can then use the scores allow us to quickly hone in on both reactions that lack genes and gene-associated reactions with low gene homology as possible targets for more experimental investigation. They also provide a logical starting point for future users looking to expand upon and improve the existing reconstruction.

## Biochemistry Improvements

The most noteworthy addition to our reconstruction was including electron bifurcation to complete the Wolfe Cycle, resulting in what is, to our knowledge, the first accurate depiction of central carbon catabolism in hydrogenotrophic methanogenesis. Because it is a relatively novel mechanism of energy conservation, this heterodisulfide reduction step was not part of our annotation database and thus was not included in the automated first reconstruction draft. Rather, the default mechanism of energy conservation matched methylotrophic methanogens and utilized methanophenazine, an electron carrier known to be absent from *M. maripaludis* and other hydrogenotrophic methanogens. We replaced the methylotrophic pathway with the correct electron bifurcation pathway, linking heterodisulfide reduction with electrons from H2 to carbon dioxide reduction via reduced ferredoxin. This incident serves as an example of a commonly-encountered reconstruction pitfall, in which information available in annotation databases does not sufficiently represent known biochemical phenomena. Therefore, it is paramount that even as automated reconstruction methods improve, we take the time to carefully evaluate biochemical literature and ensure that all pieces of the reconstruction match experimentally-verified pathways as much as possible.

Interestingly, *M. maripaludis* uses multiple forms of ferredoxin as electron carriers {ref} and may link multiple steps, including electron bifurcation, using specific ferredoxins. Previous gene knockout experiments suggested ferredoxin specificity may contribute to the ability of the energy-conserving hydrogenases, Eha and Ehb, to substitute for one another {ref}. Presently, the full extent of this phenomenon is not well understood and requires more experimental investigation. However, in an effort to represent ferredoxin specificity in our model, we have included a function (see Supplemental Materials) that changes promiscuous ferredoxins to specific ferredoxins for the Eha hydrogenase, heterodisulfide reductase, and formylmethanfuran dehydrogenase (carbon dioxide reduction). Using this function, which also provides a specific ferredoxins with the ability to act as promiscuous ferredoxins, tightens the coupling between the aforementioned reactions by restricting them all to one pool of electron carriers and allows us to predict changes to possible model flux distributions.

A major part of our manual curation was the effort put into adding biosynthesis pathways for the methanogenic coenzymes. *M. maripaludis* utilizes these coenzymes directly as electron carriers (methanofuran, H4MPT, coenzyme F420, coenzyme B, coenzyme M) and vital pieces of catabolic enzymes (coenzyme F430) during methanogenesis {reference}. Hence, we were adamant about including synthesis pathways for these metabolites and requiring their synthesis by including them in our biomass definition. Although several of these pathways were completely included in the Model SEED database, many reactions were missing and nearly all of the reactions were added manually. Similarly, we also added achaellin sugar and archaeol membrane lipid syntheses to the biomass.

In a similar vein, we sought to accurately represent sulfur assimilation, a poorly understood pathway in *M. maripaludis* beyond the fact that it does not uptake sulfate {reference}. Because sulfate is generally the default sulfur source for most microorganisms, our initial reconstruction included a sulfate transporter and sulfate reduction pathway. We removed the sulfate transporter and instead added a transporter for sulfide, the primary sulfur source for *M. maripaludis.* Our updated sulfur assimilation pathway includes a pathway for sulfide oxidation to sulfite, which is used in multiple biosynthetic pathways, via a hypothesized Dsr-LP enzyme. Taken together with additional coenzyme syntheses, these modifications demonstrated the need for rigorous manual curation to add known biochemical pathways that were not part of the automated reconstruction and remove pathways that are known not to function in the organism. By employing these methods and by working in a team of experienced metabolic modelers and biochemical experts, we have created a reconstruction that is qualitatively more consistent with accumulated biochemical knowledge of our organism.

## Model Validation

Evaluating the metabolic network reconstruction by qualitatively comparing it to known biochemical phenomena is a useful way to gauge how close the network can represent actual biochemistry. To make more quantitative comparisons, we must convert the reconstruction to a metabolic model by imposing flux constraints on the network, enforcing mass balance on all metabolites, and optimizing to an objective function. A common way of quantitatively evaluating the resulting model is to simulate maximum cell growth under steady-state conditions and compare its growth yield predictions to experimentally-determined values. Due to the narrow range of possible substrates for our hydrogenotrophic system, our comparison was limited to two conditions: H2-limiting and formate-limiting. Previous work yielded YCH4 (grams [cell mass] per mole of CH4 produced) values of 2.86 ± 0.58 for H2 limitation and 2.31 ± 0.58 for formate limitation (39).

However, we were unsure of the accuracy of these experimental yields based on disparate reported values for the conversion factor between optical density and dry cell weight. To mitigate our concerns, we re-measured this conversion factor (see Methods) and calculated a value of #VALUE#, as compared to the value of #VALUE# used to calculate the original yields. Using our newly-measured conversion factor, we calculated YCH4 values of 2.86 ± 0.58 for H2 limitation and 2.31 ± 0.58 for formate limitation.

Growth yield predictions can also vary considerably in response to model ATP maintenance energies (40). Our default biomass equation obtained through Kbase specified growth associated maintenance (GAM) as 40.11 (mmol per grams [cell mass]) and the first draft model specified no non-growth associated maintenance (NGAM). Though these figures gave what we considered to be reasonable growth yield predictions, we chose to measure GAM and NGAM ourselves with respect to methane production rate (see Methods). Following our growth experiments, we altered these values to #VALUE# and #VALUE# in H2-limiting conditions and #VALUE # and #VALUE# in formate-limiting conditions for GAM and NGAM, respectively.

Using our recalibrated growth yields and ATP maintenance values, we compared our model’s FBA predictions of maximum growth yields to the experimentally-derived measurements as shown in Figure 2. Both of our computational values agreed closely with the experimental values, falling within the uncertainty range in each case. Though growth yield validation is not an absolute measure of model performance, our model’s ability to closely reproduce experimental results without manual overfitting suggested a high propensity for generating viable growth predictions. This result bodes well for our model’s utility as a predictive tool as we look to use it to generate quantitatively feasible growth hypotheses for novel strain designs.

Gene knockout experiments present something of a “gold standard” for validating a metabolic reconstruction based on its model. At its core, a constraint-based model is built around gene-protein-reaction relationships that connect genotype to growth phenotype. Thus, comparing model predictions of gene knockout lethality provide an excellent way to quantitatively measure the qualitative content of the model. This process naturally hinges on the availability of gene knockout data for the organism being modeled, ideally with the abundance of data found for traditional model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* {reference}. In the case of *M. maripaludis* transposon mutagenesis has been used to calculate an essentiality index of all genes in *M. maripaludis* (41), but there is relatively little data where gene knockout experiments have been systematically carried out *in vivo.* Because much of methanogenesis revolves around the function of different hydrogenases, the bulk of available gene knockout data involves hydrogenase knockouts on different media. For our test set, we were able to assemble a knockout panel of 30 binary growth phenotypes based on previous publications (30–35). Though the breadth of these knockout genotypes is limited, they are all vital pieces of central carbon metabolism and therefore, they give us a good idea of how well our model can predict knockouts in central catabolism. In comparing with these data, as shown in Figure 3, our model achieved 90% prediction accuracy and a Matthew’s correlation coefficient of 0.67. These high values suggested that our model is an excellent predictor of growth phenotype based on genotype changes in central carbon metabolism. This result was particularly encouraging because we avoided training our model on this dataset in the interest of preventing overfitting our model to the validation set.

It is also worth noting that all 3 incorrect predictions had regulatory bases. In these cases, knockouts of 5 or 6 hydrogenases are lethal in formate-grown cells with downregulation of the glyceraldehyde-3-phosphate ferredoxin oxidoreductase (GAPOR) cycle. When the GAPOR cycle is upregulated (as in Δ6H2asesupp and Δ7H2asesupp), it provides anaplerotic electrons that allow cell growth in formate media, even when missing up to 7 hydrogenases. Upregulation of the GAPOR cycle is a regulatory difference, and we have chosen to leave the cycle upregulated as a default to allow for a richer spectrum of possible flux distributions. Thus, although our default model does not predict knockout growth phenotypes with 100% accuracy, it can easily be tuned to reflect regulatory differences.

## Thermodynamic Calculations

Free energy plays a key role in biochemistry as all biological systems must have a sufficiently low overall free energy to support growth {reference}. When simulating optimal growth using a metabolic model we expect the same rules to apply to our system, hence we can apply thermodynamic constraints to the model based on metabolite free energies of formation. In a previous study, free energies of formation were used to constrain reversibility of all internal model reactions based on the second law of thermodynamics (22). This method, while rigorous, is highly dependent on concentration and can be overly restrictive with regard to predicted flux distributions; thus it is most effective when paired with metabolite concentration data (23). Lacking extensive concentration data for *M. maripaludis,* we chose to represent free energy constraints in a novel approach where we add free energies only to exchange reactions, the set of metabolites that can be taken up or produced by the model. These metabolites effectively represent the overall biochemical “reaction” being performed by the model, therefore it is reasonable to expect this overall reaction to produce a negative overall free energy.

To calculate overall model free energy, we added free energies of formation from the Equilibrator database (37) to all exchanges reactions for which these values could be calculated via the group contribution method (38). We expanded the standard structure of the model to include standard free energy values for a 1 mM concentration, temperature of 25 C, pressure of 1 bar, pH of 7, and ionic strength of 0.1 M. During flux balance analysis, these standard values are modified based on specified concentrations of exchange metabolites and temperature, allowing us to estimate the overall free energy generated during growth by multiplying each metabolite’s free energy of formation by its exchange flux. The feasible model solution space can also be further reduced by restricting overall free energy to be negative, the equivalent of imposing the second law of thermodynamics on the organism itself.

We expect that this straightforward calculation will be a useful addition to our model, particularly as we aim to use it as a platform for generating possible strain designs. With regard to free energy, methanogens are particularly notable in that they subsist close to the thermodynamic limit to support growth (21). It follows that for any potential strain design, we must pay particular attention to the overall free energy of our system, lest it dip below this vital threshold. It may also provide a metric for differentiating between multiple feasible strain designs by ranking them in order of thermodynamic feasibility. At the very least, it serves as an additional capability of our model and as a checkpoint to ensure that our overall stoichiometry matches up with overall free energy. We have included example functions for adding metabolite free energies to our model and performing FBA with an additional free energy calculation (see Supplementary Materials).

## Reconstruction and Model Distribution

Reconstructing a metabolic network is an iterative process and therefore, it is paramount that reconstructions be as clear as possible to encourage future updates and expansions (28). We have strived for clarity in both our nomenclature and in our decision making process for including each reaction present in our reconstruction. Reactions and metabolites in our model are based upon identifiers and names found in Kbase, but also include crosslinks to KEGG identifiers (25), enzyme commission numbers, and reaction subsystems where available. Each reaction and gene in the model is justified and connected to its literature and/or database source (see Supplementary Materials).

Additionally, we have sought to maximize usability of both our reconstruction and the resulting model. The systems biology markup language (SBML) is a standard medium for distributing metabolic reconstructions (42); thus, we have included our reaction network in SBML. In our experience using models from other groups, we have found a wide range of model usability, from models that can easily be imported and simulated to those that are difficult to use and interpret. In the interest of making our simulations and results easy to reproduce, we have included our model in MATLAB data structure format and all of our codes for simulating model growth on different media and gene knockout phenotypes (see Supplementary Materials). In the interest of transparency and to support the iterative process of reconstructing metabolic networks, we have also made our codes and reconstruction available on Github (https://github.com/marichards/methanococcus).

# Discussion

Metabolic reconstructions provide us with a computational approach to studying the complexity of genome-scale biochemistry. With iMR544, we have created the most comprehensive reconstruction of *M. maripaludis* currently available, the first manually-curated reconstruction built on top of likelihood based gapfilling, and a straightforward new method to add thermodynamic constraints to metabolic models.

Our reconstruction process hinged first and foremost on literature sources, which provided high quality biochemical information. Through manual curation from these sources, we created a model with an accurate representation of the Wolfe Cycle, added pathways to synthesize methanogenic coenzymes, included archaellin and archaeol lipid biosyntheses, and corrected the sulfur assimilation pathway. The efficacy of rigorous manual curation is shown not only in the qualitative accuracy of our reconstruction, but also in the formidable quantitative capabilities of the resulting model.

This reconstruction is also, to our knowledge, the first genome scale metabolic network created using likelihood-based gapfilling. Favoring this method over a maximum parsimony method helped us to automatically identify 66 more genes, increasing the gene coverage of our reconstruction prior to the start of manual curation. Simultaneously, we were also able to assign reaction likelihood scores for many of the reactions, lending a measure of confidence level for including each reaction in the network.

For a methanogen living close to the edge of thermodynamic feasibility, we also thought it salient to include some calculation of overall free energy when simulating our model. We have thus introduced a novel method of predicting overall model free energy generation based solely on standard free energies and concentrations of exchange metabolites. Though a much more trivial method than previous efforts, our method gives a quick assessment of whether a predicted flux distribution is thermodynamically possible and could prove a particularly useful tool for guiding future metabolic engineering designs.

We recognize that our reconstruction effort represents only an incremental step toward understanding *M. maripaludis* metabolism and that many other users may follow in our footsteps. With these considerations in mind, we strived for maximum transparency in our metabolic network to make our reconstruction decisions apparent to future users and to make our results easily reproducible. In addition to distributing our model in SBML format, we have also included numerous functions for modifying the reconstruction, simulating growth, and predicting gene knockout lethality. By following this model of reconstruction distribution, we hope to encourage increased transparency and user-friendliness in all metabolic network reconstructions.

Taken together, this work demonstrates both the importance of doing extensive manual curation and the effectiveness of using automated gene homology based methods when reconstructing metabolic networks. Automated methods such as likelihood based gapfilling are invaluable in their ability to quickly expand and enrich the reaction network and literature from experimental studies serves as the gold standard for adding new pieces to a reconstruction. Above all, the goal of a metabolic network reconstruction is to create a realistic representation of organism metabolism that can then be used to push forward biological discovery and fuel innovations in strain design. With this reconstruction, we have contributed a tool to help understand the unique biochemistry of hydrogenotrophic methanogens and unlock their potential as metabolic engineering targets.

# Acknowledgements

We would like to thank Eliora Gachelet for assisting in chemostat growth experiments and Matthew Benedict for his expertise and advice concerning methanogenic archaea, likelihood-based gapfilling, and metabolic model construction.

# References

1. **Haynes CA**, **Gonzalez R**. 2014. Rethinking biological activation of methane and conversion to liquid fuels. Nat Chem Biol **10**:331–339.

2. **Zhang X**, **Myhrvold NP**, **Caldeira K**. 2014. Key factors for assessing climate benefits of natural gas versus coal electricity generation. Environ Res Lett **9**:114022.

3. **Mueller TJ**, **Grisewood MJ**, **Nazem-Bokaee H**, **Gopalakrishnan S**, **Ferry JG**, **Wood TK**, **Maranas CD**. 2014. Methane oxidation by anaerobic archaea for conversion to liquid fuels. J Ind Microbiol Biotechnol **42**:391–401.

4. **Thauer RK**, **Kaster A-K**, **Seedorf H**, **Buckel W**, **Hedderich R**. 2008. Methanogenic archaea: ecologically relevant differences in energy conservation. Nat Rev Microbiol **6**:579–591.

5. **DiMarco AA**, **Bobik TA**, **Wolfe RS**. 1990. Unusual coenzymes of methanogenesis. Annu Rev Biochem **59**:355–394.

6. **Deppenmeier U**. 2002. The unique biochemistry of methanogenesis. Prog Nucleic Acid Res Mol Biol **71**:223–283.

7. **Heiden S**, **Hedderich R**, **Setzke E**, **Thauer RK**. 1993. Purification of a cytochrome b containing H2:heterodisulfide oxidoreductase complex from membranes of Methanosarcina barkeri. Eur J Biochem **213**:529–535.

8. **Setzke E**, **Hedderich R**, **Heiden S**, **Thauer RK**. 1994. H2: heterodisulfide oxidoreductase complex from Methanobacterium thermoautotrophicum. Eur J Biochem **220**:139–148.

9. **Kaster A-K**, **Moll J**, **Parey K**, **Thauer RK**. 2011. Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. Proc Natl Acad Sci **108**:2981–2986.

10. **Thauer RK**. 2012. The Wolfe cycle comes full circle. Proc Natl Acad Sci **109**:15084–15085.

11. **Buckel W**, **Thauer RK**. 2013. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. Biochim Biophys Acta BBA - Bioenerg **1827**:94–113.

12. **Jones WJ**, **Paynter MJB**, **Gupta R**. 1983. Characterization of Methanococcus maripaludis sp. nov., a new methanogen isolated from salt marsh sediment. Arch Microbiol **135**:91–97.

13. **Hendrickson EL**, **Kaul R**, **Zhou Y**, **Bovee D**, **Chapman P**, **Chung J**, **Macario EC de**, **Dodsworth JA**, **Gillett W**, **Graham DE**, **Hackett M**, **Haydock AK**, **Kang A**, **Land ML**, **Levy R**, **Lie TJ**, **Major TA**, **Moore BC**, **Porat I**, **Palmeiri A**, **Rouse G**, **Saenphimmachak C**, **Söll D**, **Dien SV**, **Wang T**, **Whitman WB**, **Xia Q**, **Zhang Y**, **Larimer FW**, **Olson MV**, **Leigh JA**. 2004. Complete Genome Sequence of the Genetically Tractable Hydrogenotrophic Methanogen Methanococcus maripaludis. J Bacteriol **186**:6956–6969.

14. **Sarmiento FB**, **Leigh JA**, **Whitman WB**. 2011. Chapter three - Genetic Systems for Hydrogenotrophic Methanogens, p. 43–73. *In* Ragsdale, ACR and SW (ed.), Methods in Enzymology. Academic Press.

15. **Milne CB**, **Kim P-J**, **Eddy JA**, **Price ND**. 2009. Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. Biotechnol J **4**:1653–1670.

16. **Stolyar S**, **Van Dien S**, **Hillesland KL**, **Pinel N**, **Lie TJ**, **Leigh JA**, **Stahl DA**. 2007. Metabolic modeling of a mutualistic microbial community. Mol Syst Biol **3**:92.

17. **Goyal N**, **Widiastuti H**, **Karimi IA**, **Zhou Z**. 2014. A genome-scale metabolic model of Methanococcus maripaludis S2 for CO2 capture and conversion to methane. Mol Biosyst **10**:1043–1054.

18. **Xia Q**, **Wang T**, **Hendrickson EL**, **Lie TJ**, **Hackett M**, **Leigh JA**. 2009. Quantitative proteomics of nutrient limitation in the hydrogenotrophic methanogen Methanococcus maripaludis. BMC Microbiol **9**:149.

19. **Graham DE**, **White RH**. 2002. Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics. Nat Prod Rep **19**:133–147.

20. **Benedict MN**, **Mundy MB**, **Henry CS**, **Chia N**, **Price ND**. 2014. Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. PLoS Comput Biol **10**:e1003882.

21. **Jackson BE**, **McInerney MJ**. 2002. Anaerobic microbial metabolism can proceed close to thermodynamic limits. Nature **415**:454–456.

22. **Henry CS**, **Broadbelt LJ**, **Hatzimanikatis V**. 2007. Thermodynamics-Based Metabolic Flux Analysis. Biophys J **92**:1792–1805.

23. **Hoppe A**, **Hoffmann S**, **Holzhütter H-G**. 2007. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. BMC Syst Biol **1**:23.

24. **Thiele I**, **Palsson BØ**. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc **5**:93–121.

25. **Kanehisa M**, **Goto S**. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res **28**:27–30.

26. **Caspi R**, **Altman T**, **Dale JM**, **Dreher K**, **Fulcher CA**, **Gilham F**, **Kaipa P**, **Karthikeyan AS**, **Kothari A**, **Krummenacker M**, **Latendresse M**, **Mueller LA**, **Paley S**, **Popescu L**, **Pujar A**, **Shearer AG**, **Zhang P**, **Karp PD**. 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **38**:D473–D479.

27. **Henry CS**, **DeJongh M**, **Best AA**, **Frybarger PM**, **Linsay B**, **Stevens RL**. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol **28**:977–982.

28. **Heavner BD**, **Price ND**. 2015. Transparency in metabolic network reconstruction enables scalable biological discovery. Curr Opin Biotechnol **34**:105–109.

29. **Schellenberger J**, **Que R**, **Fleming RMT**, **Thiele I**, **Orth JD**, **Feist AM**, **Zielinski DC**, **Bordbar A**, **Lewis NE**, **Rahmanian S**, **Kang J**, **Hyduke DR**, **Palsson BØ**. 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc **6**:1290–1307.

30. **Porat I**, **Kim W**, **Hendrickson EL**, **Xia Q**, **Zhang Y**, **Wang T**, **Taub F**, **Moore BC**, **Anderson IJ**, **Hackett M**, **Leigh JA**, **Whitman WB**. 2006. Disruption of the Operon Encoding Ehb Hydrogenase Limits Anabolic CO2 Assimilation in the Archaeon Methanococcus maripaludis. J Bacteriol **188**:1373–1380.

31. **Lie TJ**, **Costa KC**, **Lupa B**, **Korpole S**, **Whitman WB**, **Leigh JA**. 2012. Essential anaplerotic role for the energy-converting hydrogenase Eha in hydrogenotrophic methanogenesis. Proc Natl Acad Sci **109**:15473–15478.

32. **Lupa B**, **Hendrickson EL**, **Leigh JA**, **Whitman WB**. 2008. Formate-Dependent H2 Production by the Mesophilic Methanogen Methanococcus maripaludis. Appl Environ Microbiol **74**:6584–6590.

33. **Costa KC**, **Lie TJ**, **Jacobs MA**, **Leigh JA**. 2013. H2-Independent Growth of the Hydrogenotrophic Methanogen Methanococcus maripaludis. mBio **4**:e00062–13.

34. **Costa KC**, **Wong PM**, **Wang T**, **Lie TJ**, **Dodsworth JA**, **Swanson I**, **Burn JA**, **Hackett M**, **Leigh JA**. 2010. Protein complexing in a methanogen suggests electron bifurcation and electron delivery from formate to heterodisulfide reductase. Proc Natl Acad Sci **107**:11050–11055.

35. **Hendrickson EL**, **Leigh JA**. 2008. Roles of Coenzyme F420-Reducing Hydrogenases and Hydrogen- and F420-Dependent Methylenetetrahydromethanopterin Dehydrogenases in Reduction of F420 and Production of Hydrogen during Methanogenesis. J Bacteriol **190**:4818–4821.

36. **Matthews BW**. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta BBA - Protein Struct **405**:442–451.

37. **Flamholz A**, **Noor E**, **Bar-Even A**, **Milo R**. 2011. eQuilibrator—the biochemical thermodynamics calculator. Nucleic Acids Res gkr874.

38. **Jankowski MD**, **Henry CS**, **Broadbelt LJ**, **Hatzimanikatis V**. 2008. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. Biophys J **95**:1487–1499.

39. **Costa KC**, **Yoon SH**, **Pan M**, **Burn JA**, **Baliga NS**, **Leigh JA**. 2013. Effects of H2 and Formate on Growth Yield and Regulation of Methanogenesis in Methanococcus maripaludis. J Bacteriol **195**:1456–1462.

40. **Gonnerman MC**, **Benedict MN**, **Feist AM**, **Metcalf WW**, **Price ND**. 2013. Genomically and biochemically accurate metabolic reconstruction of *Methanosarcina barkeri* Fusaro, iMG746. Biotechnol J **8**:1070–1079.

41. **Sarmiento F**, **Mrázek J**, **Whitman WB**. 2013. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon Methanococcus maripaludis. Proc Natl Acad Sci **110**:4726–4731.

42. **Hucka M**, **Finney A**, **Sauro HM**, **Bolouri H**, **Doyle JC**, **Kitano H**, **Forum and the rest of the S**, **Arkin AP**, **Bornstein BJ**, **Bray D**, **Cornish-Bowden A**, **Cuellar AA**, **Dronov S**, **Gilles ED**, **Ginkel M**, **Gor V**, **Goryanin II**, **Hedley WJ**, **Hodgman TC**, **Hofmeyr J-H**, **Hunter PJ**, **Juty NS**, **Kasberger JL**, **Kremling A**, **Kummer U**, **Novère NL**, **Loew LM**, **Lucio D**, **Mendes P**, **Minch E**, **Mjolsness ED**, **Nakayama Y**, **Nelson MR**, **Nielsen PF**, **Sakurada T**, **Schaff JC**, **Shapiro BE**, **Shimizu TS**, **Spence HD**, **Stelling J**, **Takahashi K**, **Tomita M**, **Wagner J**, **Wang J**. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**:524–531.

# Tables and Figures

|  |  |  |
| --- | --- | --- |
| ***Methanococcus maripaludis S2* model comparison** | | |
| Model | iMM518 | iMR544 |
| Protein Coding Genes | 518 | 544 |
| % ORF Coverage | 30 | 32 |
| Intra/Extracellular Metabolites | 556/49 | 665/53 |
| Dead End Metabolites | 163 | 265 |
| Internal Reactions | 570 | 645 |
| Exchange Reactions | 49 | 59 |
| Gene-Associated Reactions | 464 | 590 |
| % Reactions Associated with Genes (non-exchange) | 75 | 91 |
|  | | |

Table 1: A comparison between iMR544 and iMM518 indicates that our model covers slightly more of the genome, including over 125 more gene-associated reactions. Our model includes more internal reactions and metabolites, as well as approximately 100 more dead end metabolites. Though these metabolites represent the portion of metabolism that cannot carry flux, all of our model's dead end metabolites are part of gene-associated reactions and thus represent promising avenues for future model expansion

Figure 1: A chart showing broad subsystem groupings of the 85 reactions in iMR544 that are not associated with any genes. Reactions falling underneath the “None” subsystem grouping were present in the Model SEED database but had no subsystems listed there and no obvious membership in another subsystem. Reactions grouped within “Other” were dissimilar both from the other categories and from one another, thus we felt they did not merit creation of multiple additional categories.

Figure 2: Comparing our models’ growth yield predictions on hydrogen and formate, respectively, to experimental data we found that our predictions fell within the experimental error of the measured values. This comparison demonstrates that our models can reproduce these values, suggesting that the models and the reconstruction they were derived from can reliably predict growth yields are quantitatively consistent with *in vivo* results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Genotype | H2 | Formate | H2 + Formate | Formate + CO |
| ∆hmd | N | N | N | N |
| ∆mtd | N | N | N | N |
| ∆frcA | N | N | N | N |
| ∆fruA | N | N | N | N |
| ∆frcA∆fruA | N | N | N | N |
| ∆vhcAU∆vhuA | N | N | N | N |
| ∆hdrB2 | N | N | N | N |
| ∆fdhA1 | N | N | N | N |
| ∆fdhA2 | N | N | N | N |
| ∆fdhA1∆fdhA2 | N | L | N | L |
| ∆fdhA2∆fdhB2 | N | N | N | N |
| ∆ehbF | N | N | N | N |
| ∆3H2ase | N | N | N | N |
| ∆5H2ase | L | N | N | N |
| ∆6H2ase | L | N | N | N |
| ∆6H2ase∆cdh | L | N | N | N |
| ∆6H2asesupp | L | N | N | N |
| ∆7H2asesupp | L | N | N | N |
| **Total Correct:** | **10 of 10** | **14 of 16** | **2 of 2** | **1 of 2** |

Figure 3: Knockout lethality predictions from running FBA on our models show close agreement with experimental results of hydrogenase knockouts. Green boxes indicate growth phenotypes where our models correctly replicated experimental results; red boxes indicate growth phenotypes where our models were incorrect; white boxes indicate growth phenotypes where we lacked experimental validation data. Across the full spectrum of conditions, our models correctly predicted 27 of 30 conditions (90%) accurately, resulting in a strong Matthews Correlation Coefficient of 0.67. This suggests that our reconstruction produces models that accurately depict the effects of genotype alterations on growth phenotypes.