Optimizing the Biosynthesis of Therapeutic Compounds in *E. coli* using Computational Modeling

A Technical Report submitted to the Department of Biomedical Engineering

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Benjamin C. Neubert
Spring, 2020

Technical Project Advisors
Jason Papin
Thomas Moutinho

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Benjamin C. Neubert

Approved _____ Date _____
Jason Papin, Department of Biomedical Engineering

Abstract

Genome-scale metabolic network reconstructions (GENREs) are a powerful computational tool for mathematically modeling the metabolic processes within a cell at a systems-level. The development of improved curation methods through strategic data integration would improve our ability to use GENREs to understand metabolic diseases and to inform metabolic engineering[1,2]. Metabolomics aims to identify metabolites within a biological system, which can then be integrated into a GENRE to increase its accuracy[3]. Due to the cost of gathering metabolomics data, there is a need to identify which media conditions would hold the most value for model curation. To this end, we developed a novel data-driven GENRE curation pipeline using a combination of well-established packages and *in vitro* aerobic growth screen data[4–7]. Production sub-networks were created using weighted parsimonious flux balance analysis with different objective functions based upon single products across 44 candidate minimal media conditions with varied carbon sources. We were able to generate a prioritized list of media conditions that induced the greatest variation among ensemble members, representing the conditions for which gathering metabolomics data would be most informative. The resulting data-driven GENRE was applied to determine the optimal dietary input for the generation of therapeutic compounds within the gastrointestinal microbiome. This study developed a novel data-driven GENRE curation pipeline for determining the optimal biosynthesis of therapeutic compounds with reduced uncertainty in network structure and increased curation efficiency.

Keywords: Genome-scale Metabolic Network Reconstructions, Metabolomics, Systems Biology

---

Introduction

The gastrointestinal (GI) microbiome represents a diverse set of organisms that play a critical role in human health by contributing to immune system modulation, metabolic functions,

and other important activities[8–10]. GI microbiome alteration has been linked to the pathogenesis of multiple GI diseases, including playing a role in the pathogenesis of inflammatory bowel disease (IBD)[11]. The reported cases of IBD are on the rise within The United States, with approximately 3 million adults affected as of 2015 which is a 50% increase from cases in 1999[12]. Individuals are typically diagnosed between 15 and 35, but approximately 5% of all IBD cases occur within kids, which corresponds to as many as 80,000 children with the disease[13]. Those affected by IBD experience clinical GI symptoms and emotional burdens as a result of the chronic symptoms, resulting in direct and indirect costs at a population level of between $14.6 and $31.6 billion in the United States as of 2014[14].

Metabolism is one of the primary mechanisms by which the GI microbiome interacts with the host organism[15]. Prebiotics are compounds which confer specific changes to the composition and/or activity of the GI microbiome, which could be used to lessen the burden of IBD on patients[16]. These compounds can be used to increase production of therapeutically relevant metabolites in the gut of the host such as indole, succinate, and acetate. Tryptophan catabolites, such as indole, have been shown to contribute to many processes promoting intestinal homeostasis[17]. Succinate has been shown to positively regulate energy homeostasis and glucose control[10,18]. Short chain fatty acids, including acetate, play a critical role in the maintenance of gut and immune homeostasis[19]. Utilization of prebiotics to increase production of these molecules can result in positive health outcomes, especially in those afflicted with diseases such as IBD[20].

Escherichia coli is one of the most studied and best characterized model organisms[21]. While *E. coli* resides in the gut of most people, research has largely focused on pathogenic strains of the bacteria representing an opportunity for further research into mechanistic host-

microbiome interactions involving the bacteria[22,23]. Lactobacilli and bifidobacteria are often considered when discussing probiotic bacterial strains, however *E. coli* has recently been characterized as a health-promoting bacterium providing its host defense against *P. aeruginosa* colonization and mortality[24]. Specifically, *E. coli* metabolic output was determined to be a critical indicator of resistance to infection. Therefore, *E. coli* represents the ideal organism for the characterization of a novel model development pipeline as it applies to optimizing biosynthesis of therapeutic compounds.

Recent advances in sequencing technology have led to drastic cost reductions in the overall cost of sequencing a genome. Currently it costs approximately $1,000 to sequence a genome, with cost decreases outpacing Moore's Law, which has led to thousands of sequenced genomes being produced[25]. The increased number of high-quality sequenced genomes has allowed for a greater understanding of a host of biological functions. Genome-scale metabolic network reconstructions (GENREs) are a computational framework that utilize genomic data and biochemical network structure to facilitate further understanding of an organism's metabolism. Newly sequenced genomes can be annotated and used in conjunction with biochemical databases and experimental data in order to create high-quality GENREs[26]. The current methodology for manually generating high-quality GENREs requires a great degree of labor and time on the order of several months or even years[27]. Automated methods such as ModelSEED have been able to reduce this figure to approximately 48 hours using only an assembled genome sequence[4]. However, these methods only reach 66% accuracy based upon gene essentiality and Biolog data. Therefore, additional steps must be taken to develop a pipeline through which a high-quality GENRE may be produced, while minimizing the time to achieve this quality. Integration of -omic data including metabolomics, transcriptomics, and proteomics into GENREs is a well-

established method to improve upon model performance and contextualize data, with a variety of methods being developed for these functions[28–31]. Metabolomics aims to identify metabolites within a biological system, which can then be integrated into a GENRE to increase its accuracy. When implementing metabolomics data into the model, a simple glucose minimal media or context-specific media is used during sample preparation. The choice of media condition directly impacts the extracellular metabolome of the bacteria, impacting metabolomics data and, therefore, data integration into the existing model[32]. Due to the cost of gathering metabolomics data, there is a need to identify which media conditions would hold the most value for model curation. The development of an efficient data-driven curation pipeline will allow for a larger application of GENREs leading to advances in our understanding of metabolic diseases, inter-species interactions, evolutionary processes, metabolic network properties, metabolic engineering, and allow for the prediction of cellular phenotypes[1,2].

Within this project, we have generated a novel data-driven curation pipeline for GENREs, which will help to guide future model reconstruction projects within the field. Further, we have applied this curation pipeline to create a well-curated metabolic reconstruction of *E. coli* K-12 allowing for an increased understanding of its metabolism within an aerobic environment through *in silico* predictions and *in vitro* growth data. The curated model and curation pipeline were further applied to effectively identify metabolic strategies to optimize production of therapeutic metabolites.

Results

Model Curation Pipeline

A novel data-driven curation pipeline for GENREs was developed within this project that incorporates several well-developed tools (Figure 1)[5,6,33,34]. The complete genome sequence of *E. coli* K-12 as well as the complete annotation of its genome was used as a starting point for the curation

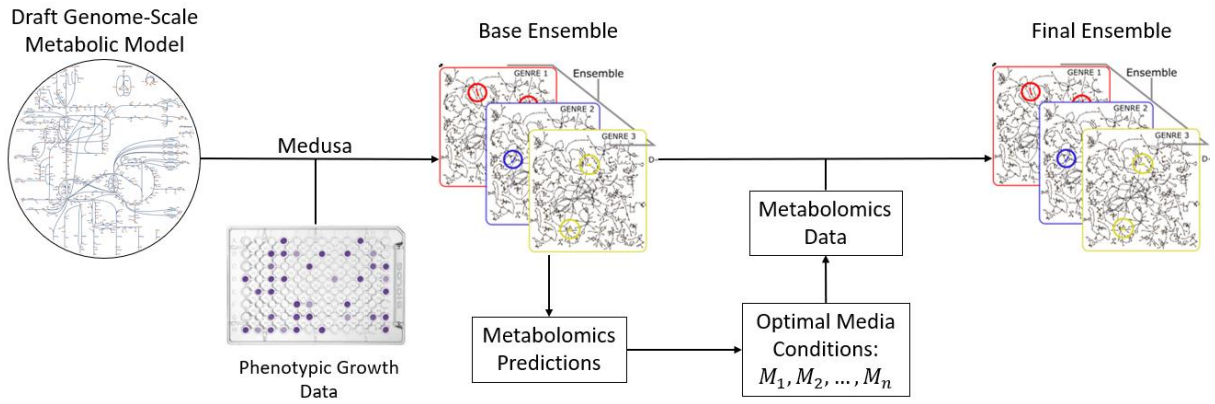


Figure 1. Novel Data-driven GENRE Curation Pipeline. A draft metabolic model is created using an annotated genome from the PATRIC database input into the automated reconstruction tool ModelSEED. Phenotypic growth data on a variety of minimal media conditions is integrated into the model using Medusa to create an ensemble of models, each of which being an equally feasible solution. Metabolomics predictions are generated using the base ensemble network structure to determine the optimally informative media condition for metabolomics data generation. Metabolomics data is generated for the optimal media conditions and then used to constrain the base ensemble to create a final data-driven ensemble that is consistent with genetic, phenotypic growth, and metabolomics data.

pipeline[35,36]. We have utilized the annotated genome of *E. coli* K-12 in conjunction with ModelSEED to generate a draft reconstruction[5]. The reconstruction was transferred to COBRApy using Mackinac to enable further development of the model and the application of advanced analyses provided by COBRApy[33,34]. *In vitro* growth data from *E. coli* K-12 substr. MG1655 grown aerobically in Biolog microarray plates was gathered from the EcoCyc database[37]. Data covered four different microarray plates resulting in growth data across a variety of minimal media conditions with varied carbon, nitrogen, sulfur, and phosphorus sources. We integrated the growth data using a python package called Medusa to generate an ensemble of 53 GENREs, which manage uncertainty in network structure and improve the model's predictive capabilities[6,38]. In this pipeline, the curated ensemble is then used in conjunction with a predictive algorithm to determine the optimal media conditions for metabolomics data generation and further curation. The selection of the optimal media conditions is governed by the model simulations that indicate which conditions will allow for the optimal curation of the ensemble. Metabolomics data is particularly useful for model curation because it provides data that accounts for both the metabolic inputs and outputs. Simulations are performed on the base ensemble using the media conditions for which metabolomics was gathered. The models are constrained such that there is required production of

metabolites detected through the metabolomics data. Reactions are added to the model from a universal reaction bag until the model is capable of producing all excreted metabolites and biomass. The resulting ensemble represents a data-driven set of models that recapitulate observed growth phenotypes and metabolomics data.

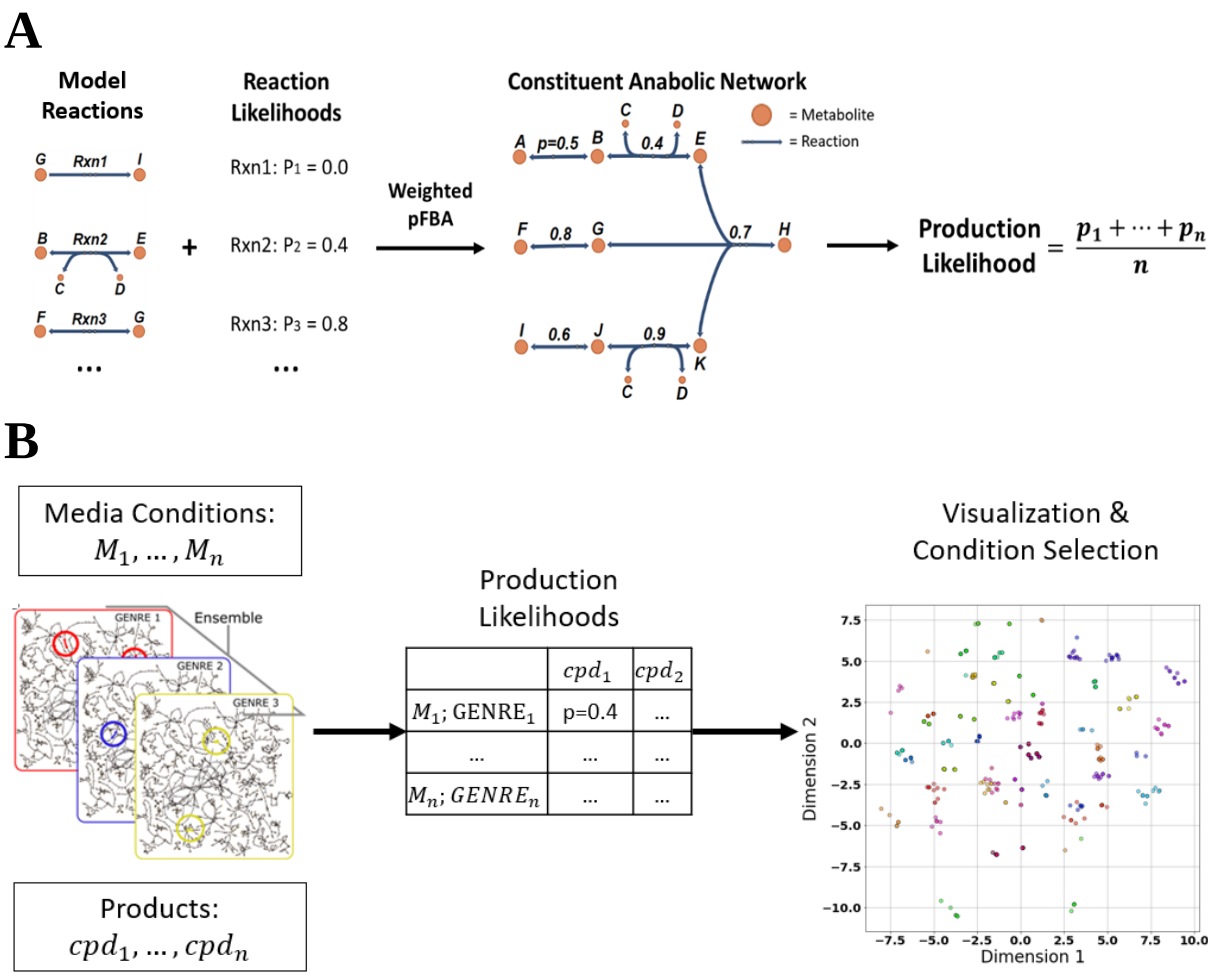


Figure 2. Process of Generating Data to Evaluate Optimal Metabolomics Media Conditions.
(A) A modified version of Parsimonious Flux Balance Analysis was performed using genomic likelihoods from ProbAnnoPy to generate constituent anabolic networks[47,48]. Reactions with greater likelihoods were preferentially included. Each sub-network has input metabolites which represent a media condition and a demand reaction for a certain metabolic product. The production likelihood metric is an average of all the reaction likelihoods associated with the reactions included in the sub-constituent network. This metric is used as a summary statistic that allows for the comparison of sub-constituent networks across GENREs, where a higher production likelihood corresponds with greater genetic evidence for that particular constituent anabolic network. **(B)** Production likelihoods are generated for each member of the ensemble with each possible combination of media condition and product[33]. Each ensemble member and media condition pair are treated as a sample with the production likelihood for each product as its features. This dataset is then visualized and used to develop a metric quantifying the variation between ensemble members to ultimately select ideal media conditions for metabolomics data generation.

Optimal Media Condition Prediction Algorithm

In order to determine the media condition for which generating metabolomics data would be most valuable, we developed a novel algorithm quantifying variation in network structure. Sub-networks were generated to capture portions of network structure by optimizing towards production of a single metabolite, while minimizing overall flux through the network and weighting reactions by their genetic likelihoods[7,39,40] (Figure 2A). A total of 1,515 out of 1,681 reactions within the ensemble had a ProbAnnoPy-associated genetic likelihood, while 1,076 of these reactions have likelihoods greater than 0 (Supplementary Figure 1). This result demonstrates the limitation of genome annotations even for a well-studied model organism, thus requiring additional data types to characterize the metabolism of the organism. The genetic likelihood of every reaction within the constituent anabolic network is averaged to summarize each sub-network, termed a production likelihood. The distribution of production likelihood scores for each simulated metabolite can be seen in Supplementary Figure 2.

A high-dimensional dataset was then created using the production likelihood metric for a combination of metabolites and minimal media conditions in order to sample various network states (See methods for additional details; Figure 2B). The minimal media conditions used were only varied by the carbon source as a proof of concept, but the algorithm can analyze entirely different media conditions. Carbon sources were chosen from Biolog microarray plates for which growth was observed in *E. coli* and biochemical knowledge of reactions associated with the metabolite existed. The resulting dataset represented a 33-dimensional feature space, which was visualized in two dimensions through multidimensional scaling (MDS; Figure 3). From the visualization, samples cluster more similarly between media conditions than between ensemble members indicating media conditions drive greater differences in network structure than exists inherently between ensemble members. Additionally, multiple media conditions cluster similarly

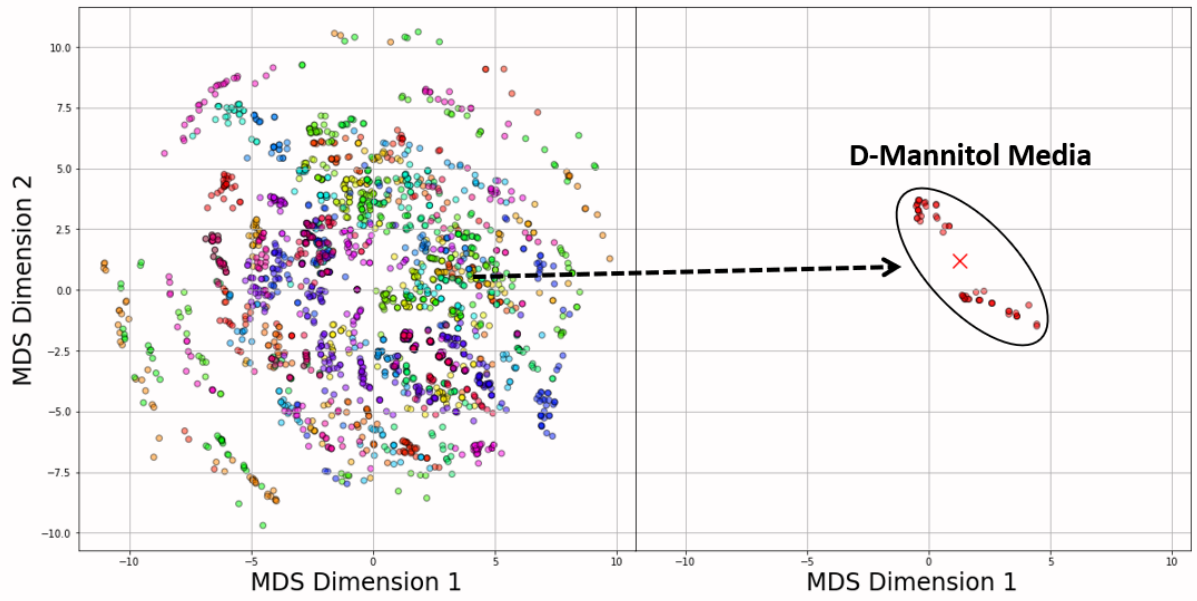


Figure 3. Multidimensional Scaling of Production Likelihood Data. MDS of the underlying data shows a 2-d imensional representation of variation in ensemble members. Each color represents a media condition and points on the graph represent ensemble members. For each media condition there are 53 corresponding ensemble member points. D-Mannitol is highlighted as an example of this, with an X representing its centroid which is used in the variation metric developed in this study.

to one another, likely due to accessing similar portions of the network structure. For instance, D-glucose and galactose cluster similarly due to occupying similar portions of *E. coli*'s metabolism.

Next, we developed a metric that quantifies the variation between ensemble members for a particular media condition. Increased variation in ensemble members represented increased flux through areas of the network with increased uncertainty. In order to quantify induced variation, a centroid was calculated for each media condition using production likelihoods for each product as coordinates and ensemble members as samples (D-Mannitol in Figure 3). The sum of distances to the centroid in high-dimensional space was used as a metric for the variation induced by a media condition (Figure 4). The media conditions with the highest sum of distances value represent the ideal targets for metabolomics generation. These results will be used for future metabolomics data generation and integration into the GENRE. The top three media conditions based upon this metric are deoxyadenosine, alpha-D-lactose, and L-threonine.

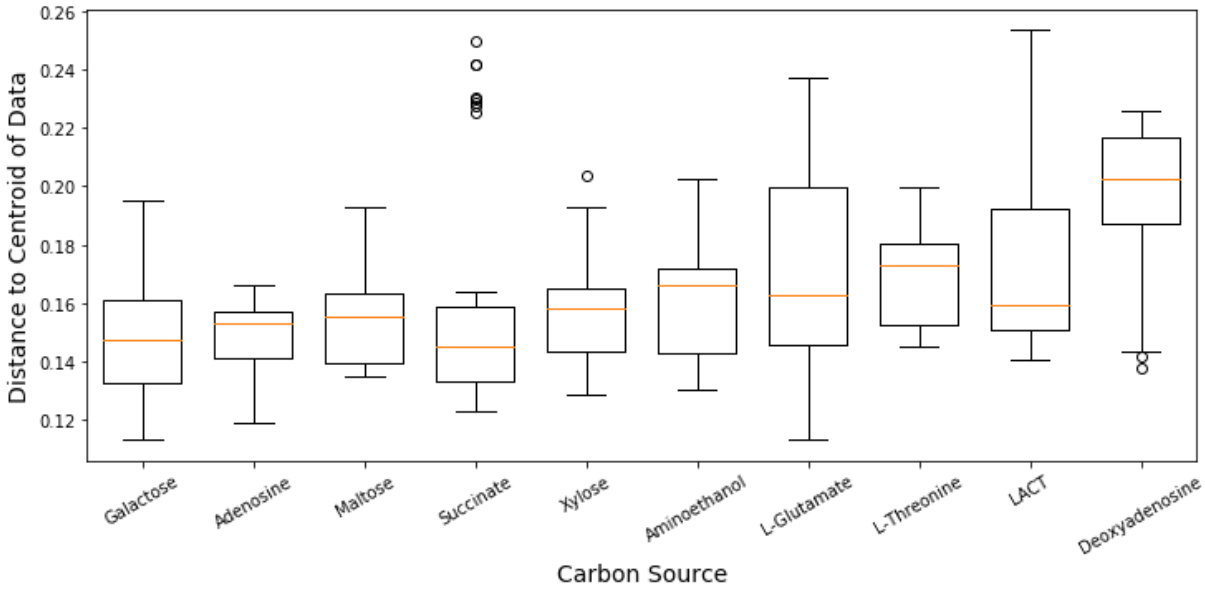


Figure 4. Distance Distribution of Media Condition Targets. Individual data points within a media condition's boxplot represent the distance of an ensemble member to its media condition's centroid. The plots are organized in ascending order by priority with the largest distance sums on the right.

Optimal Biosynthesis of Therapeutic Products

To apply the model resulting from the developed curation pipeline, we sought to determine the optimal dietary input for maximizing production of therapeutic products in *E. coli.* The therapeutic compounds: indole, succinate, and acetate were focused on within this analysis. Due to an inability to generate experimental data, a final ensemble using the pipeline could not be created. Therefore, we utilized the base ensemble with integrated growth screen data created through Medusa. A set of 600 media conditions were created using a combination of carbon sources, nitrogen sources, and the same base used while integrating growth data as described in Plata et al.[41]. 25 nitrogen sources and 24 carbon sources were randomly selected from the set of positive phenotypic microarray conditions for *E. coli.* A joint optimization problem was formed for each therapeutic compound where flux was maximized through a biomass objective function and export of the therapeutic compound. Further, the model was constrained to 90% of its maximum growth to ensure a potential physiologically possible solution rather than biasing the

model too heavily towards the production of the therapeutic compound. Simulations were completed across all media conditions and ensemble members. Media conditions produced varied results across ensemble members. Therefore, the average flux through the therapeutic compound export reaction was used to asses media conditions. Through this analysis, a set of media conditions to optimize synthesis of indole, succinate, and acetate were identified (Figure 5).

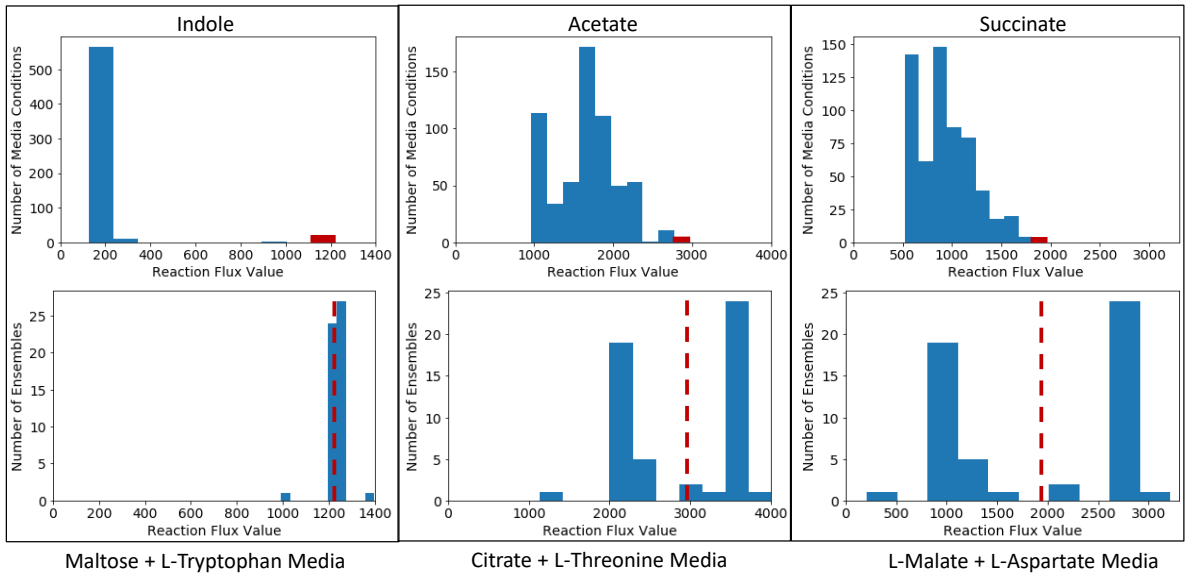


Figure 5. Optimal Media Conditions for Biosynthesis of Therapeutics. Model simulations were run on each ensemble member across 600 candidate media conditions with varied nitrogen and carbon sources. The uppermost histograms represent the average therapeutic export flux across the ensemble as the result of each of the 600 media conditions. The maximum average flux through the therapeutic export reaction was used to determine the optimal media condition for biosynthesis, which was located at the rightmost portion of the distribution. The lower histograms represent the distribution of reaction flux values through the export reaction for each of the 53 ensemble members used to complete the simulation. The red dashed line represents the mean of this distribution. A combination of Maltose and L-Tryptophan was found to result in maximum production of indole. Citrate and L-threonine media was identified for acetate. L-malate and L-aspartate media was identified as optimal for succinate production.

Discussion

As a result of this project, we have generated a metabolomics data-driven pipeline, which will help to guide future model reconstruction projects within the field. Further, we have created a well-curated metabolic reconstruction of *E. coli* K-12 allowing for an increased understanding of its metabolism within an aerobic environment through *in silico* predictions and *in vitro* growth

data. Through a predictive algorithm, we were able to determine the optimal media conditions for metabolomics data generation and subsequent integration. Thus, draft network structure can be used to inform curation efforts efficiently. Further, the created curation pipeline represents an effective compromise between automated techniques and manual curation efforts. Minimal required data generation and a predictive metabolomics algorithm reduces model curation time, enabling researchers to more effectively curate GENREs. The curated model allowed for a better understanding of metabolic pathways related to indole, acetate, and succinate. The improved confidence in metabolic capabilities related to the production of these therapeutic compounds allowed for the identification of the optimal dietary input to maximize production of these metabolites. Optimal dietary input for *E. coli* can be used to inform future prebiotic design with the potential to design a probiotic/prebiotic combination to maximize production of therapeutic metabolites within the GI microbiome.

Due to coronavirus-related changes, experimental work was unable to be completed during the Spring semester. Consequently, *in vitro* phenotypic microarray data was unable to be gathered within an anaerobic environment. Anaerobic data of this form could be found from two studies which were compiled in the EcoCyc database, but provided inconsistent results[37]. Therefore, we instead performed aerobic simulations using the wealth of aerobic data available on EcoCyc from multiple studies. Performing this analysis in an aerobic environment modified the resulting metabolomics and biosynthesis predictions which would likely differ in the context of anaerobic metabolism. Future work will focus on gathering this data so that more accurate predictions of *E. coli* K-12 metabolism within the GI microbiome can be created.

Additionally, metabolomics data was not able to be gathered due to coronavirus-related changes. As a result, proper validation of the metabolomics algorithm could not be performed

due to a lack of extracellular metabolomics data for *E. coli* K-12 across multiple minimal media conditions. In the future, metabolomics data will be gathered to determine the accuracy of metabolomics predictions. A general overview of how the metabolomics algorithm will be validated is described within the Methods. Due to a lack of metabolomics data, we were unable to integrate metabolomics data into our base ensemble. Analyses within this project were thus constrained to the output base ensemble from Medusa. Future validation of the curation pipeline will be conducted to gain a better understanding of the utility of different steps within the pipeline as well as the quality of the output model as a whole. We will make gene essentiality predictions with the curated *E. coli* metabolic model from the curation pipeline, and then we will validate these predictions using experimentally-derived gene essentiality data to assess the accuracy of the generated metabolic model[42]. In addition, we will generate growth predictions across all Biolog phenotypic microarray conditions and compare these to experimental results. Positive growth conditions were utilized within the data integration step, but negative growth conditions cannot be included in the Medusa pipeline. Therefore, growth predictions represent an excellent test set to assess if the model contains metabolic capabilities not representative of *E. coli*. Validation through growth predictions and gene essentiality will be performed on each step of the curation pipeline to assess the value gained at each step. Comparisons will be made to the existing manually-curated model, iJO1366, for *E. coli* K-12[43].

Materials and Methods

In Vitro Single-Carbon Source Utilization Screen Data Collection and Processing

Aerobic phenotype-microarray data was collected through the EcoCyc database, which contained a compilation of five separate studies that performed the Biolog *in vitro* phenotype microarray on *E. coli* K-12 substr. MG1655[37,44–46]. Data covered four phenotype microarray

plates: PM1, PM2A, PM3, and PM4 which covered a variety of different carbon, nitrogen. phosphorus, and sulfur sources. Four out of five studies contained data on all plates, while one only provided data on PM1. Growth was classified into four categories: no growth, low growth, growth and inconsistent results based upon the results of the underlying studies. For the purposes of this analysis, the distinction between low and high growth did not matter so both of these categories were considered growth. Instances where the underlying study results conflicted were classified by the majority result. Cases where there was not a majority result were classified as no growth. Further, media conditions for which the varied source was not present within the universal reaction bag were excluded from the analysis due to limited biochemical knowledge of reactions incorporating the metabolite.

Model Curation Pipeline

The sequenced genome of *E. coli* K-12 was taken from the PATRIC database and annotated using their genome annotation service[5]. The annotated genome was input into the PATRIC metabolic reconstruction service which uses the ModelSEED algorithm to create a draft GENRE[4]. Growth data was gathered from the EcoCyc database and processed as described above.

Growth data was integrated into Medusa to produce a base ensemble[6]. The Biolog base medium utilized in Medusa was the same as described in Plata et al.[41]. The ModelSEED universal reaction bag was used for gap filling reactions in the Medusa pipeline. Metabolites from the Biolog data which were missing from the model were added, including an exchange reaction to allow for movement of the media into the system. Instances where the Biolog media condition was not available in the universal reaction bag were excluded from the analysis. 110 cycles of iterative gap filling were used to create a total of 53 ensemble members in this analysis.

154 total media conditions were used to perform gap filling, pulled from the phenotypic microarray data as described above.

The base ensemble output from Medusa was used to predict the optimal media conditions for metabolomics integration as described below. The following methodology were not followed within this analysis due to an inability to conduct experimental work. Selected media conditions are used to generate metabolomics data. Metabolomics data is filtered to determine the metabolites that are secreted by the organism. The ensemble is constrained such that there is forced production and excretion of the detected metabolites then gap filling is performed until the model is capable of reproducing the experimental data and producing biomass. The resulting ensemble represents a high-quality data-driven model that recapitulates physiological metabolomic data and growth phenotypes.

Constituent Anabolic Network Generation

Probabilistic pFBA-based constituent anabolic network generation was accomplished using three Python packages, Cobrapy, Mackinac, and ProbAnnoPy[7,33,34]. The complete ModelSEED universal reaction bag was downloaded from the GitHub repository and filtered based on the annotation quality 330 score, including all reactions with an 'OK' quality status or better[4]. For each reaction in the ModelSEED universal reaction bag, we used ProbAnnoPy to generate a reaction likelihood based on the FASTA file for *E.* coli K-12 obtained from the PATRIC database[5]. The Cobrapy implementation of Parsimonious Enzyme Usage Flux Balance Analysis (pFBA) was altered to allow for each reaction's linear constraint to be set individually based on the reaction likelihood. The linear constraint for each reaction was set to one minus the reaction likelihood (a value between 0 and 1). There were reactions included in the universal reaction bag that were lacking from the ProbAnnoPy template model, therefore resulting in

several gene-associated reactions lacking reaction likelihood scores. The reactions without likelihoods were left at a full minimization penalty (linear constraint value of 1). We chose to penalize the reactions without likelihoods to bias our results towards the construction of networks for which all reactions had evidence of presence. The linear constraints applied to each reaction based on likelihood acted as a weighting (inclusion penalty) for the minimization step in pFBA, resulting in the reactions with greater likelihood having a lower penalty for carrying flux; therefore, the reactions had a higher likelihood of being included in the constituent anabolic networks. Using this methodology, we generated constituent anabolic networks by setting a certain input media condition and constraining flux through the single metabolite objective function. We ran our likelihood-weighted pFBA flux minimization across each ensemble member and isolated the reactions that carried flux to get the desired product. The resulting networks consist of the direct reactions that would be part of a production pathway as might be shown in a typical biosynthesis pathway figure, while also accounting for all of the secondary and energy metabolites. Additionally, this algorithm is optimizing for three core characteristics in the constituent networks: 1) minimum flux through the network (loosely, the minimum number of reactions), 2) maximum average reaction likelihood across the constituent network, and 3) output flux within 90% of the optimal yield of the metabolic product.

Optimal Media Condition Prediction Algorithm

The base ensemble created from integrating minimal media growth screen data with Medusa captures the uncertainty underlying the draft metabolic reconstruction. Using this ensemble of models, constituent anabolic networks were generated across 44 minimal media conditions and 33 products. Minimal media conditions within this analysis were defined through the base Biolog media characterized by Plata et. al with an additional carbon source being varied

from condition to condition[47]. The chosen carbon sources were selected from Biolog PM1 for which there was evidence of *E. coli* growth. Further, carbon sources were constrained to those that were present within the existing model. The methodology presented here is capable of being expanded to include any number of media conditions and variations, the chosen media conditions were purely a demonstration of the technique limited by the computational demands of these simulations. The total number of constituent anabolic networks generated is the product of the number of products, media conditions, and ensemble members (76,956 in this analysis). Constituent anabolic networks were summarized by calculating the average genetic likelihood of reactions contained within the network. Media conditions and ensembles were grouped together as samples with the production likelihood for each product as features. In order to identify the media condition which induced the greatest variation in the ensemble, a centroid was calculated for each media condition based upon the ensemble member's features. The summed distance of each ensemble member to the centroid for a particular media condition was used as an assessment of the variation induced in underlying network structure. Media conditions with the greatest value of this metric represented targets for metabolomics data integration.

Figure Generation

MDS visualizations were created using Scikit-learn package with a Euclidean dissimilarity metric in Python[48]. Ensemble images were repurposed from a previous paper[38]. Pathway diagrams were created within Escher[49]. Remaining figures were created using Python and Matplotlib[50].

Code and Data Availability

All data and code utilized within this project are available with the following GitHub: https://github.com/ben-neubert/ecoli.

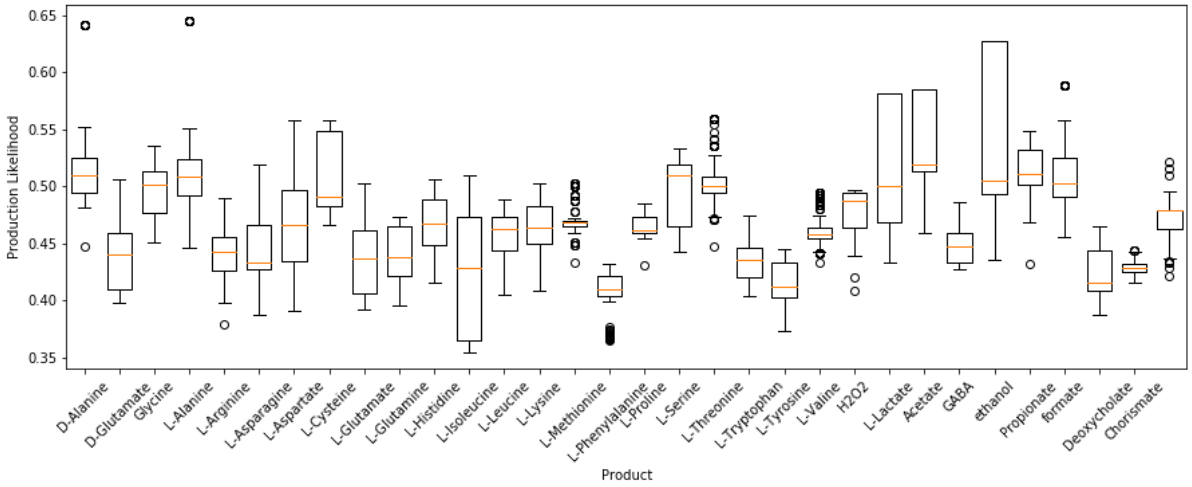
References

1. Oberhardt, M. A. & Gianchandani, E. P. Genome-scale modeling and human disease: an overview. *Front. Physiol.* **5**, (2015).
2. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* **5**, 320 (2009).
3. Töpfer, N., Kleessen, S. & Nikoloski, Z. Integration of metabolomics data into metabolic networks. *Front Plant Sci* **6**, (2015).
4. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology* **28**, 977–982 (2010).
5. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* **42**, D581–D591 (2014).
6. Medlock, G. L. & Papin, J. Medusa: software to build and analyze ensembles of genome-scale metabolic network reconstructions. *bioRxiv* (2019) doi:10.1101/547174.
7. King, B. ProbAnnoWeb and ProbAnnoPy: probabilistic annotation and gap-filling of metabolic reconstructions | Bioinformatics | Oxford Academic. *Bioinformatics* **34**, 1594–1596 (2018).
8. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nature Medicine* **24**, 392–400 (2018).
9. Morrison, D. J. & Preston, T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189–200 (2016).
10. Round, J. L. & Mazmanian, S. K. The gut microbiome shapes intestinal immune responses during health and disease. *Nat Rev Immunol* **9**, 313–323 (2009).
11. Khan, I. *et al.* Alteration of Gut Microbiota in Inflammatory Bowel Disease (IBD): Cause or Consequence? IBD Treatment Targeting the Gut Microbiome. *Pathogens* **8**, (2019).
12. Inflammatory Bowel Disease Prevalence (IBD) in the United States. *Center for Disease Control and Prevention* https://www.cdc.gov/ibd/data-statistics.htm (2019).
13. The Facts About Inflammatory Bowel Diseases. *Crohns and Colitis Foundation of America* https://www.crohnscolitisfoundation.org/sites/default/files/2019-02/Updated%20IBD%20Factbook.pdf (2014).
14. Mehta, F. Report: Economic Implications of Inflammatory Bowel Disease and Its Management. *AJMC* https://www.ajmc.com/journals/supplement/2016/importance_of_selecting_appropriate_therapy_inflammatory_bowel_disease_managed_care_environment/importance_of_selecting_appropriate_therapy_inflammatory_bowel_disease_managed_care_environment_report_economic_implications_ibd (2016).
15. Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nature Communications* **10**, 1–10 (2019).
16. Roberfroid, M. Prebiotics: The Concept Revisited. *J Nutr* **137**, 830S-837S (2007).
17. Roager, H. M. & Licht, T. R. Microbial tryptophan catabolites in health and disease. *Nature Communications* **9**, 3294 (2018).
18. Soty, M., Gautier-Stein, A., Rajas, F. & Mithieux, G. Gut-Brain Glucose Signaling in Energy Homeostasis. *Cell Metabolism* **25**, 1231–1242 (2017).
19. Tan, J. *et al.* Chapter Three - The Role of Short-Chain Fatty Acids in Health and Disease. in *Advances in Immunology* (ed. Alt, F. W.) vol. 121 91–119 (Academic Press, 2014).

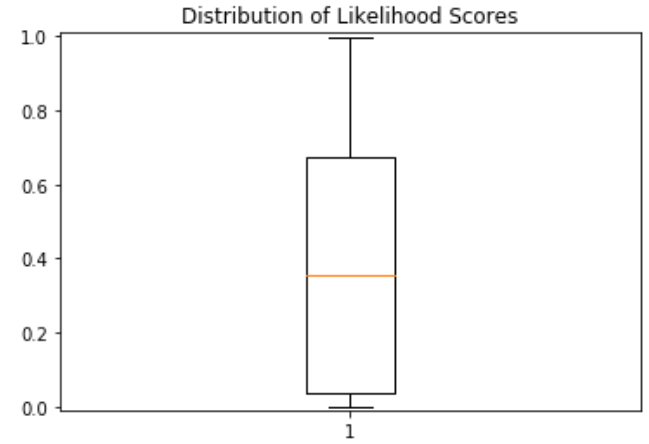
20. Parada Venegas, D. *et al.* Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Front Immunol* **10**, (2019).
21. Riley, M. *et al.* Escherichia coli K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res* **34**, 1–9 (2006).
22. Kaper, J. B., Nataro, J. P. & Mobley, H. L. T. Pathogenic *Escherichia coli*. *Nature Reviews Microbiology* **2**, 123–140 (2004).
23. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
24. Christofi, T., Panayidou, S., Dieronitou, I., Michael, C. & Apidianakis, Y. Metabolic output defines Escherichia coli as a health-promoting microbe against intestinal Pseudomonas aeruginosa. *Scientific Reports* **9**, 1–13 (2019).
25. Wetterstrand, K. DNA Sequencing Costs: Data. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.
26. Haggart, C. R., Bartell, J. A., Saucerman, J. J. & Papin, J. A. Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods Enzymol* **500**, 411–433 (2011).
27. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols* **5**, 93–121 (2010).
28. Jensen, P. A. & Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* **27**, 541–547 (2011).
29. Sajitz-Hermstein, M., Töpfer, N., Kleessen, S., Fernie, A. R. & Nikoloski, Z. iReMet-flux: constraint-based approach for integrating relative metabolite levels into a stoichiometric metabolic models. *Bioinformatics* **32**, i755–i762 (2016).
30. Schmidt, B. J. *et al.* GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics* **29**, 2900–2908 (2013).
31. Zur, H., Ruppin, E. & Shlomi, T. iMAT: an integrative metabolic analysis tool. *Bioinformatics* **26**, 3140–3142 (2010).
32. Kell, D. B. *et al.* Metabolic footprinting and systems biology: the medium is the message. *Nat Rev Microbiol* **3**, 557–565 (2005).
33. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology* **7**, 74 (2013).
34. Mundy, M., Mendes-Soares, H. & Chia, N. Mackinac: a bridge between ModelSEED and COBRApy to generate and analyze genome-scale metabolic models. *Bioinformatics* **33**, 2416–2418 (2017).
35. Riley, M. *et al.* Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res.* **34**, 1–9 (2006).
36. Blattner, F. R. *et al.* The Complete Genome Sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).
37. Keseler, I. M. *et al.* EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* **41**, D605–D612 (2013).
38. Biggs, M. B. & Papin, J. A. Managing uncertainty in metabolic network structure and improving predictions using EnsembleFBA. *PLOS Computational Biology* 25 (2017).
39. Moutinho, T. J. *et al.* Functional Anabolic Network Analysis of Human-associated Lactobacillus Strains. *bioRxiv* 746420 (2019) doi:10.1101/746420.

40. Lewis, N. E. *et al.* Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology* **6**, 390 (2010).
41. Plata, G., Henry, C. S. & Vitkup, D. Long-term phenotypic evolution of bacteria. *Nature* **517**, 369–372 (2015).
42. Xavier, J. C., Patil, K. R. & Rocha, I. Metabolic models and gene essentiality data reveal essential and conserved metabolism in prokaryotes. *PLOS Computational Biology* **14**, e1006556 (2018).
43. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011 | Molecular Systems Biology. https://www.embopress.org/doi/full/10.1038/msb.2011.65.
44. Yoon, S. H. *et al.* Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome Biology* **13**, R37 (2012).
45. AbuOun, M. *et al.* Genome Scale Reconstruction of a Salmonella Metabolic Model. *J Biol Chem* **284**, 29480–29488 (2009).
46. The evolution of metabolic networks of E. coli. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3229490/.
47. Long-term phenotypic evolution of bacteria | Nature. https://www.nature.com/articles/nature13827.
48. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
49. King, Z. A. *et al.* Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLOS Computational Biology* **11**, e1004321 (2015).
50. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9**, 90–95 (2007).

Supplementary Figures



Supplementary Figure 1. Distribution of Production Likelihoods Across Products. Each boxplot represents the production likelihood value across all media condition and ensemble member combinations. Production likelihood values vary heavily from product to product, highlighting the differences between ensemble members and media conditions.



Supplementary Figure 2. Distribution of Likelihood Scores. Boxplot showing all non-zero likelihood values across the *E. coli* model. Likelihood scores were generated using ProbAnnoPy in conjunction with an annotated genome of *E. coli* K-12.