Final Report
2017

Grant Number: DE-SC10822882/0008091, "Development of a Knowledgebase (MetRxn) of Metabolites, Reactions and Atom Mappings to Accelerate Discovery and Redesign"
Applicant/Institution: The Pennsylvania State University
Street Address: Office of Sponsored Programs, 110 Technology Center
City/State/Zip: University Park, PA 16802
Principal Investigator: Costas D. Maranas
Address: 126 Land and Water Research Building, University Park PA 16802
Telephone Number: 814-863-9958 Email: costas@psu.edu
DOE/Office of Science Program Office: The Office of Advanced Scientific Computing Research
DOE/Office of Science Program Technical Program Manager Contact: Christine Chalk & Susan Gregurick

A. Overall Technical Summary

With advances in DNA sequencing and genome annotation techniques, the breadth of metabolic knowledge across all kingdoms of life is increasing. The construction of genome-scale models (GSMs) facilitates this distillation of knowledge by systematically accounting for reaction stoichiometry and directionality, gene to protein to reaction relationships, reaction localization among cellular organelles, metabolite transport costs and routes, transcriptional regulation, and biomass composition. Genome-scale reconstructions available now span across all kingdoms of life, from microbes to whole-plant models, and have become indispensable for driving informed metabolic designs and interventions. A key barrier to the pace of this development is our inability to utilize metabolite/reaction information from databases such as BRENDA [1], KEGG [2], MetaCyc [3], etc. due to incompatibilities of representation, duplications, and errors. Duplicate entries constitute a major impediment, where the same metabolite is found with multiple names across databases and models, which significantly slows downs the collating of information from multiple data sources. This can also lead to serious modeling errors such as charge/mass imbalances [4,5] which can thwart model predictive abilities such as identifying synthetic lethal gene pairs and quantifying metabolic flows. Hence, we created the MetRxn database [6] that takes the next step in integrating data from multiple sources and formats to automatically create a standardized knowledgebase. We subsequently deployed this resource to bring about new paradigms in genome-scale metabolic model reconstruction, metabolic flux elucidation through MFA, modeling of microbial communities, and pathway prospecting. This research has enabled the PI's group to continue building upon research milestones and reach new ones (see list of MetRxn-related publications below). We elucidate this using the six aims listed below:

Aim 1: Reaction/Metabolite Data Standardization, Correction, and Congruency (MetRxn)
Aim 2: Development of Novel Database Designs for MetRxn
Aim 3: Incorporation of Atom Mapping Information for all Reaction Entries in MetRxn
Aim 4: Creating Models using MetRxn: Metabolic Model Reconstruction of Plants
Aim 5: Curating Metabolic Models using MetRxn
Aim 6: MetRxn-driven metabolic reconstruction and analysis of microbial communities
Aim 7: Extracting Knowledge using MetRxn: Pathway Prospecting and Synthetic Biology

We have promptly posted on the PI's webpage (http://www.maranasgroup.com/) and broadly disseminated all data as well as the obtained models and computational tools in accordance with DOE's policy. Progress has been made on all fronts since the time of receiving the award and we have met and hopefully surpassed all milestones put forth in the proposal. The work has yielded a number of successful developments both in the area of extension of the MetRxn database and computational platforms to support all of our modeling

aims, and in the area of scientific/technical advances. Below is further information on the results related to the specific individual aims as outlined in the proposal.

The following sections detail the research outcomes of the entire project towards the project aims along a multitude of fronts in the development of MetRxn database and computational tools to analyze, elucidate and redesign biological pathways. The ultimate outcome of the work is a suite of computational aids for analyzing and optimizing the behavior of biological networks built upon the MetRxn database. Below are listed publications of research carried out (partially or completely) based on this research grant support in the last three years.

Publications

1. Gopalakrishnan, S., Pakrasi, H.B., and Maranas, C.D., Elucidation of Photoautotrophic Carbon Flux Topology in Synechocystis PCC 6803 using Genome-scale Carbon Mapping Models Metabolic Engineering (Under Review).
2. Chan, S., Cai, J., Wang, L., Simons-Senftle, M. and Maranas, C.D. (2017). Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. *Bioinformatics*, 33(22), pp.3603-3609.
3. Chan, S., Simons, M. and Maranas, C.D. (2017). SteadyCom: Predicting microbial abundances while ensuring community stability. *PLOS Computational Biology*, 13(5), e1005539.
4. Kumar A., Wang L., Ng C.Y., and Maranas C.D. (2017) Pathway design using de novo steps through uncharted biochemical spaces. *Nature Communications*, in press.
5. Wang, L., Dash, S., Ng, C. and Maranas, C. D. (2017). A review of computational tools for design and reconstruction of metabolic pathways. *Synthetic and Systems Biotechnology*.
6. Gopalakrishnan, S. and Maranas, C.D. (2015). 13C metabolic flux analysis at a genome-scale. *Metabolic Engineering*, 32, pp.12-22.
7. Gopalakrishnan, S. and Maranas, C.D. (2015). Achieving Metabolic Flux Analysis for S. cerevisiae at a Genome-Scale: Challenges, Requirements, and Considerations. *Metabolites*, 5(3), pp.521-535.
8. Chowdhury R, Chowdhury A, Maranas C.D. Using Gene Essentiality and Synthetic Lethality Information to Correct Yeast and CHO Cell Genome-Scale Models. Metabolites 2015;5:536–70. doi:10.3390/metabo5040536.
9. Chowdhury, A. and Maranas, C.D. (2015). Designing overall stoichiometric conversions and intervening metabolic reactions. *Scientific Reports*, 5(1).
10. Ng, C., Khodayari, A., Chowdhury, A. and Maranas, C.D. (2015). Advances in de novo strain design using integrated systems and synthetic biology tools. *Current Opinion in Chemical Biology*, 28, pp.105-114.
11. Ng, C., Farasat, I., Maranas, C.D. and Salis, H. (2015). Rational design of a synthetic Entner–Doudoroff pathway for improved and controllable NADPH regeneration. *Metabolic Engineering*, 29, pp.86-96.
12. Khodayari, A., Chowdhury, A. and Maranas, C.D. (2015). Succinate Overproduction: A Case Study of Computational Strain Design Using a Comprehensive Escherichia coli Kinetic Model. *Frontiers in Bioengineering and Biotechnology*, 2.
13. Chowdhury, A., Zomorrodi, A. and Maranas, C.D. (2015). Bilevel optimization techniques in computational strain design. *Computers & Chemical Engineering*, 72, pp.363-372.
14. Saha, R.S., A. Chowdhury and C.D. Maranas (2014), "Recent advances in the reconstruction of metabolic models and integration of omics data", *Current Opinion in Biotechnology*, **29**, 39-45.
15. Zomorrodi, A.R., M.M. Islam and C.D. Maranas (2014), "d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities", *ACS Synthetic Biology*, http://dx.doi.org/10.1021/sb4001307
16. Zomorrodi, A. R., and C.D. Maranas (2014), "Coarse-grained optimization-driven design and piecewise linear modeling of synthetic genetic circuits", *European Journal of Operational Research*, http://dx.doi.org/10.1016/j.ejor.2014.01.054

17. Chowdhury A., A.R. Zomorrodi and C.D. Maranas (2014), "k-OptForce: Integrating Kinetics with Flux Balance Analysis for Strain Design". *PLoS Computational Biology*, 10(2):e1003487. PMID: 24586136.
18. Tee,T.W., A. Chowdhury, C.D. Maranas and J.V. Shanks (2014) "Systems metabolic engineering design: Fatty acid production as an emerging case study", *Biotechnology and Bioengineering*. doi: 10.1002/bit.25205, PMID: 24481660.
19. Grisewood, M.J., N.P. Gifford, R.J. Pantazes, Y. Li, P.C. Cirino, M.J. Janik and C.D. Maranas (2013), "OptZyme: Computational Enzyme Redesign Using Transition State Analogues", *PLoS ONE*, 8(10):e75358, doi:10.1371/journal.pone.0075358, PMID: 24116038.
20. Mueller, T.J., B.M. Berla, H.B. Pakrasi, and C.D. Maranas (2013), "Rapid construction of metabolic models for a family of Cyanobacteria using a multiple source annotation workflow", *BMC Systems Biology*, 7:142, doi:10.1186/1752-0509-7-142, PMID: 24369854.
21. Berla, B.M., Saha, R., Immethun, C.M., Maranas, C.D., Moon, T.S. and Pakrasi, H.B. (2013), "Synthetic biology of cyanobacteria: unique challenges and opportunities", *Frontiers in Microbiology*, **4**, doi: 10.3389/fmicb.2013.00246, PMID: 24009604.
22. Zomorrodi A.R, J.G. Lafontaine Rivera, J.C. Liao and C.D. Maranas (2013), "Optimization-driven identification of genetic perturbations accelerates the convergence of model parameters in ensemble modeling of metabolic networks," *Biotechnol. J.*, doi: 10.1002/biot.201200270.
23. Saha R., A.T. Verseput, B.M. Berla, T.J. Mueller, H.B. Pakrasi and C.D. Maranas (2012), "Reconstruction and Comparison of the Metabolic Potential of Cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803," *PLoS ONE*, Vol. 7, Issue 10, e48285.
24. Suthers, P.F. and C.D. Maranas (2012), "Ochestrating hi-fi annotations," *Nature Chemical Biology*, Vol. 8, 810-811.
25. Brochado, A.R., S. Andrejev, C.D. Maranas and K.R. Patil (2012), "Impact of Stoichiometry Representation on Simulation of Genotype-Phenotype Relationships in Metabolic Networks," *PLoS Comput. Biol*., Vol. 8, Issue 11, e1002758.
26. Zomorrodi A.R., P.F. Suthers, S. Ranganathan and C.D. Maranas (2012), "Mathematical optimization applications in metabolic networks," *Metabolic Engineering*, Vol. 14, 672-686.
27. Ranganathan, S., T.W. Tee, A. Chowdhury, A.R. Zomorrodi, J.M. Yoon, Y. Fu, J.S. Shanks, C.D. Maranas (2012), "An integrated computational and experimental study for overproducing fatty acids in *Escherichia coli*," *Metabolic Engineering*, Vol. 14, 687-704.
28. Copeland, W.B., B.A. Bartley, D. Chandran, M. Galdzicki, K.H. Kim, S.C. Sleight, C.D. Maranas and H.M. Sauro (2012), "Computational tools for metabolic engineering," *Metabolic Engineering*, Vol. 14, 270-280.
29. Zomorrodi, A.R. and C.D. Maranas (2012), "OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities," *PLoS Comput. Biol.,* Vol. 8, Issue 2, e1002363.
30. Kumar, A., P. Suthers and C.D. Maranas (2012), "MetRxn: A Knowledgebase of Metabolites and Reactions Spanning Metabolic Models and Databases," BMC Bioinformatics, Vol. 13, Issue 6, doi:10.1186/1471-2105-13-6.

Conference Presentations

1. Gopalakrishnan S. and Maranas C.D., "Genome-scale mapping models and algorithms for stationary and instationary MFA-based flux elucidation", Biochemical and Molecular Engineering XX, Newport Beach, CA, July 2017
2. Chan S.H.J., Cai, J., Wang L., Simons-Senftle M.N. & Maranas C.D. Uncovering and Correcting the Effect of Biomass Molecular Weight Discrepancies in FBA Calculations. American Institute of Chemical Engineering (AIChE) Annual Meeting, Minneapolis, Minnesota, Oct 2017
3. Chan S.H.J., Simons-Senftle M.N., Maranas C.D. Metabolic modeling of the gut microbiome. The 36th Summer Symposium in Molecular Biology, State College, Pennsylvania, Jun 2017

4. Chan S.H.J., Simons-Senftle M.N., Maranas C.D. Steadycom: Modeling Microbial Communities Under Steady-State Growth. American Institute of Chemical Engineering (AIChE) Annual Meeting, Minneapolis, Minnesota, Oct 2017
5. Ng C.Y., Wang L., Kumar A., Chan S.H.J., Simons M., Chowdhury A., and Maranas C.D., Application of the MetRxn database to highlight multi-tissue/organisms and expansion to include algorithms for predicting novel reactions and pathways. Genomic Science Annual Contractor-Grantee Meeting. February 6-8, Tysons Corner, VA, 2017.
6. Chan S.H.J., Simons-Senftle M.N., Maranas C.D. Metabolic modeling of the gut microbiome. Microbiome Centre, State College, Pennsylvania, Mar 2017
7. Gopalakrishnan S. and Maranas C.D., "13C-Assisted flux elucidation using genome-scale carbon mapping models" DOE Genomic Sciences Annual Contractor-Grantee Meeting, Arlington, VA, February 2017
8. Chan S.H.J., Simons-Senftle M.N., Maranas C.D. Metabolic modeling of the gut microbiome. Penn State Microbiome Research Netwroking Event, State College, Pennsylvania, Oct 2016
9. Gopalakrishnan S., Pakrasi H.B., and Maranas C.D., "Cyanobacterial genome-scale carbon mapping models for ^{13}C-MFA", 12th Workshop on Cyanobacteria, Tempe, AZ, May 2016
10. Kumar A., Gopalakrishnan S., Simons M., Chan S.H.J., and Maranas C.D., "Development of a knowledgebase of reactions, metabolites, and atom mappings to accelerate discovery and redesign", DOE Genomic Sciences Annual Contractor-Grantee Meeting, Tyson corner, VA, March 2016
11. Gopalakrishnan S. and Maranas C.D., "^{13}C Metabolic flux analysis at the genome-scale", 249th Annual ACS meeting, Denver, CO, March 2015
12. Kumar A., Gopalakrishnan S., and Maranas C.D., "MetRxn 2.0: Integrating atom mapping information for pathway comparisons and metabolic flux elucidation through MFA", DOE Genomic Sciences Annual Contractor-Grantee Meeting, Tyson corner, VA, February 2015
13. Kumar A., Gopalakrishnan S., and Maranas C.D., "Using MetRxn for flux elucidation and model reconstruction", Metabolic Engineering X, Vancouver, BC, Canada, June 2014
14. Kumar A., Gopalakrishnan S., and Maranas C.D., "Integrating atom mapping information within MetRxn: Application to metabolic flux elucidation using MFA" ,3rdConference on Constraint-Based Analysis and Reconstruction, Charlottesville, VA, May 2014
15. Kumar A., Gopalakrishnan S., and Maranas C.D., "MetRxn 2.0: Integrating atom mapping information for pathway comparisons and metabolic flux elucidation through MFA" DOE Genomic Sciences Annual Contractor-Grantee Meeting, Tyson corner, VA, February 2014

B. Specific Aim 1: Reaction/Metabolite Data Standardization, Correction, and Congruency (MetRxn)

B.1. Background

MetRxn [1] is a knowledgebase that includes standardized metabolite and reaction descriptions by integrating information from 8 highly accessed databases including BRENDA [2], KEGG [3], MetaCyc [4], Reactome.org [5] and recently published metabolic models into a single unified data set. All metabolite entries have matched synonyms, resolved protonation states, and are linked to unique structures. All reaction entries are elementally and charge balanced. This is accomplished through the use of a workflow of lexicographic, phonetic, and structural comparison algorithms. MetRxn allows for the download of

standardized versions of existing genome-scale metabolic models and the use of metabolic information for the rapid reconstruction of new ones.

In this work, we describe the development and highlight applications of the web-based resource MetRxn that integrates, using internally consistent descriptions, metabolite and reaction information from eight databases and 112 metabolic models. Since its creation, the MetRxn knowledgebase (as of March 2014) has increased its data set to contain over 120,000 metabolites and 55,000 reactions (including unresolved entries) that are charge and elementally balanced. By conforming to standardized metabolite and reaction descriptions, MetRxn enables users to efficiently perform queries and comparisons across models and/or databases. For example, common metabolites and/or reactions between models and databases can rapidly be generated along with connected paths that link source to target metabolites. The workflow followed in the creation of the MetRxn knowledgebase (see Figure B.1) identified a number of naming and structure inconsistencies in metabolites and reactions. For instance, the same metabolite name may map to molecules with different numbers of repeat units (e.g., lecithin) or completely different structures (e.g., AMP could refer to either adenosine monosphate or ampicillin). Notably, even for the most well-curated metabolic model, *E. coli i*AF1260 [6] , we found minor errors or omissions (a total of 17) arising from inconsistencies or incompleteness of representation in the data culled from other sources. For example, the metabolite abbreviation arbtn-fe3 was mistakenly associated with the KEGG ID and structure of aerobactin instead of ferric-aerobactin. In the *Corynebacterium glutamicum* model [7], 7,8-aminopelargonic acid (DAPA) has no associated structural information. Reaction matching found the same reaction in the *E. coli i*AF1260 model:

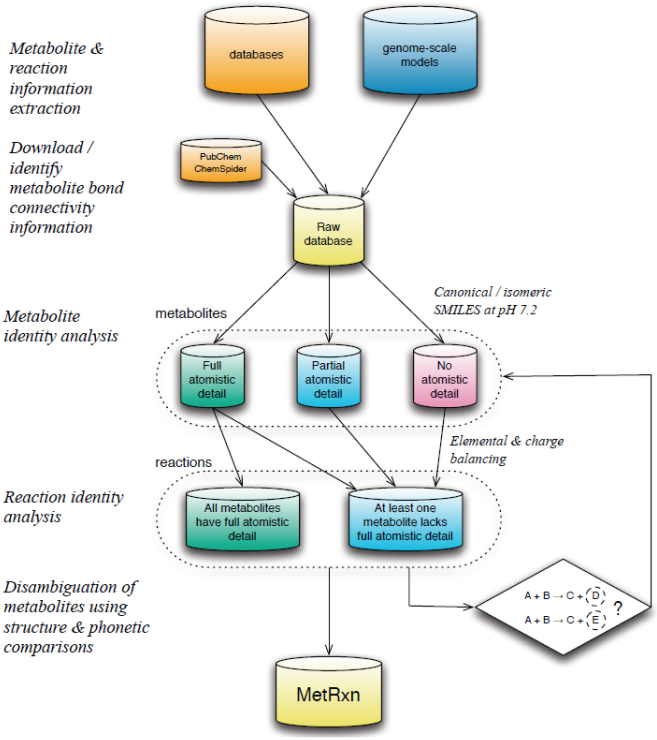


Figure B.1. Outline of the workflow of MetRxn curation procedure: After download of primary sources of data from databases and models, we integrated metabolite and reaction data, followed by calculation and reconciliation of structural information. By identifying overlaps between metabolite and reaction information, we generated elemental and charge balancing of reactions. The procedure for developing MetRxn was iterative with subsequent passes making use of previous associations to resolve remaining ambiguities.

***C. glutamicum*:** DAPA + ATP + CO2 < = > DTBIOTIN + ADP + PI
***i*AF1260 [c] :** atp + co2 + dann → adp + dtbt + (3) h + pi

which, implies that 7,8-aminopelargonic acid (DAPA) is identical to 7,8-Diaminononanoate (dann). Examination of pelargonic acid and nonanoate reveals that they were indeed known synonyms. In many cases, we were also able to assign stereo-specific information to metabolite entries in models (e.g., stipulate the L-lysine isomer for lysine). We made use of an iterative approach that allowed us to map structures from models with explicit links to structures (e.g. to KEGG or CAS numbers) to models that only provided metabolite names. Furthermore, by using a phonetic algorithm along with Jaro Winkler similarity that uses tokens for equivalent strings in metabolite names (e.g., '-ic acid' and '-ate' are equivalent) we were able to resolve an additional 159 metabolites. For example, phonetic searches flagged cis-4-coumarate and COUMARATE in the *Acinetobacter baylyi* model [8] as potentially identical compounds. Additional

checks revealed that indeed both metabolites should map to the same structure. A more complex matching example involved 1-(5'-Phosphoribosyl)-4-(N-succinocarboxamide)-5-aminoimidazole from the *Bacillus subtilis* model [9] and 1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole from the *Aspergillus nidulans* model [10]. We note that the phonetic algorithm only makes suggestions and orders the possible matches for the curator.

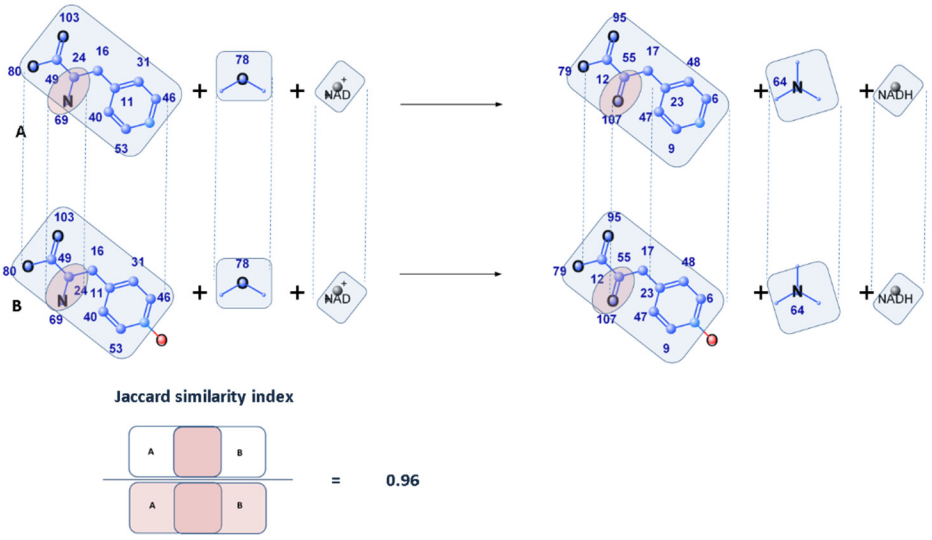


Figure B.2. Reaction comparison: We compare two reactions directly here. The numbers above show the equivalent substructures between both the reactions and not the reaction atom transitions. Common substructures are highlighted in cyan for visual clarity. Subfigure A shows the reaction L-phenylalanine:NAD+ oxidoreductase and subfigure B shows the reaction L-Tyrosine:NAD+ oxidoreductase. The common reaction center is also highlighted in pink. The Jaccard similarity calculated for the number of atoms in the common substructures was 0.96.

B.2. Results

The number of inconsistencies is significantly higher for less curated metabolic models. We used a variety of procedures to disambiguate the identity of metabolites lacking structural information ranging from reaction matching to phonetic searches. We applied the standardization procedure on 22 newly published metabolic models and latest versions of 7 metabolic databases since March 2013. Twelve of the metabolic models published since the last update belong to the bacteria taxon, 2 to Eudicots, 2 to Archaea and the

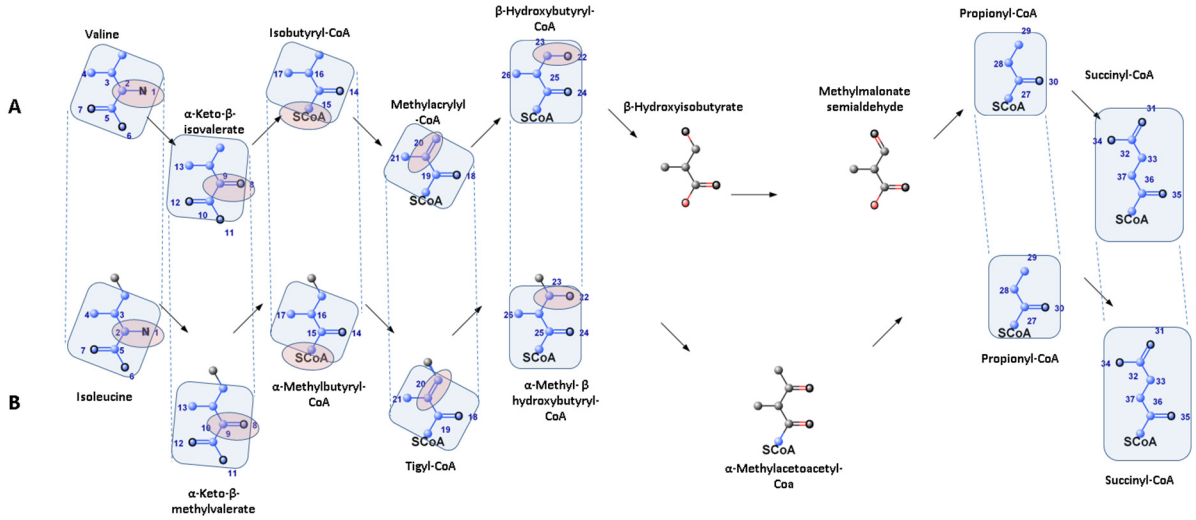


Figure B.3. Pathway comparison: We compare the two branched chain amino acid degradation pathways for Valine (A) and Isoleucine (B) degradation. Common subgraphs between the two pathways are in boxes and highlighted in cyan while the common reaction centers are identified by the oval. The first three reactions have a common enzyme, branched-chain amino acid aminotransferase, branched-chain α-keto acid dehydrogenase and acyl-CoA dehydrogenase. Valine degradation continues with enoyl-CoA hydratase and β-hydroxyisobutyryl-CoA while isoleucine degradation continues with enoyl-CoA hydratase and β-hydroxyacyl-CoA dehydrogenase. Both Valine and Isoleucine finally end up as the Citric Acid Cycle intermediate Succinyl-CoA. Co-factors such as NAD and FADH2 have been omitted for visual clarity

remaining ones to Fungi. Data from the 8 metabolic databases are kept up to date as when the latest versions are made available. The latest versions of the 7 databases are KEGG 70.0, MetaCyc 18, BRENDA 2014.1, RHEA 50, ChEBI 114, Reactome 48 and HMDB 3.5. However, many metabolites downloaded from the aforementioned datasets contain incomplete atomistic details. Details about some of the atoms of the molecules are suppressed by representing them as part of generic sides (-R group). Currently in MetRxn, 102,336 out of 115,512 metabolites and 38,132 out of 42,965 reactions have complete atomistic details.

MetRxn's primary dataset aggregates information from various metabolic resources. A major drawback we faced was in the quality of annotations provided from the primary datasets. A high quality metabolic model provides additional reaction annotations such as EC number, subsystem/pathway classification, deltaG and reaction direction. Such annotations are invaluable since they assist development of metabolic models of newly sequenced organisms from phylogenetically related metabolic models. Large numbers of reactions from metabolic models as well as databases lack the aforementioned annotations making reconstruction of quality genome-scale metabolic models a challenging task. Better annotations lead to a compilation of reactions encompassing the entire chemistry repertoire of a specific organism. It must be noted that these models are not necessarily predictive but instead have a scoping nature by allowing us to assess what is metabolically feasible.

We automate reaction annotation using a novel maximum substructure algorithm called CLCA (Canonical Labelling for Clique Approximation). CLCA is polynomial runtime algorithm capable of identifying common subgraph isomorphs between two graphs. We utilize atom connectivity information available for all reacting substrates to produce input graphs for CLCA. EC number classification is a semantic classification of the underlying reaction mechanism. Reaction mechanism can be identified by the reaction center i.e. the atoms and bonds involved in electron transfer between/within each reacting substrate. EC numbers also at times indicate the cofactors involved in a reaction. We transfer EC number annotation by comparing a EC annotated reaction graph with an unannotated reaction graph for common subgraph isomorphs. If isomorph preserves both the reaction center as well as the co-factors, we safely transfer the EC annotation. Each unannotated reaction is compared with the entire graph and only the best match is considered for annotation transfer as illustrated in Figure B.2. A similar approach of comparing unannotated pathways is presented in Figure B.3 wherein we compare two pathways. Subsystem classifications can be suggested or transferred using similarity scores. Furthermore, we annotate reactions with atom mapping information using CLCA. Atom mapping is further discussed in Aim 3.

C. Specific Aim 2: Development of Novel Database Designs for MetRxn

C.1. Background

One of the biggest challenges with maintaining a heterogeneous dataset is in the way the database has to be designed in order to accommodate all the possible queries that users post. Our experience with MetRxn has shown that maintenance can become a big bottleneck in performance and execution as well as adding strain on hardware if the design is not implemented carefully. We have modified the schema several times so that queries can execute fast and the hardware resource utilization is optimal. The next challenge would arise when we start including massive datasets such as whole genomes and proteomes. One of the ways heterogeneous data management is handled is by moving away from structured schemas towards unstructured schemas. In the IT infrastructure domain, this would be called the NoSql [11] database technology. Kbase (www.kbase.us) has done this using the MongoDb [12] architecture while the Bio4j(http://www.bio4j.com/) project using a graph database approach.

One of the goals related to database architecture with respect to MetRxn is to provide a design that allows for real-time execution of pathway prediction algorithms [13-15]. The first step in this direction will start with the inclusion of proteomic data. We plan to include proteomic data from UniProt [16] and this would again lead to some changes in the underlying schema of MetRxn. With the recent availability of the Kbase infrastructure we believe this effort will be easier than before. Since the overhead of hardware management is offloaded and also that we do not need to bother about any underlying design changes.

C.2. Results

Movement into Kbase:

We have been in conversation with members of the Kbase design team (i.e., Tom Brettin) in order to plan a path forward for converting the MetRxn schema as shown in Figure C.3 into the Kbase schema as shown in Figure C.1 and C.2. Kbase already incorporates pipelines from SEED and RAST allowing for rapid annotation of organisms. With the integration of MetRxn into Kbase, we will enhance their biochemistry database thereby augmenting and enhancing the quality of annotation in the SEED pipeline. Integration of MetRxn within the KBase resource will be a priority for this year. To have a smooth transition between the existing database and application, we modified our technology from two-tier architecture to three-tier architecture. In two-tier architecture, individual server is dedicated to a database and a web service. Such architecture will not allow programmatic access to the underlying data, since security considerations will prevent direct access to the database server. In contrast, a three-tier architecture provides a dedicated server called the application server to allow users to programmatically access and query data using REST or SOAP based http services in remote locations. Due to this change, developing REST/ SOAP based API's to access MetRxn data from KBASE remotely is possible.

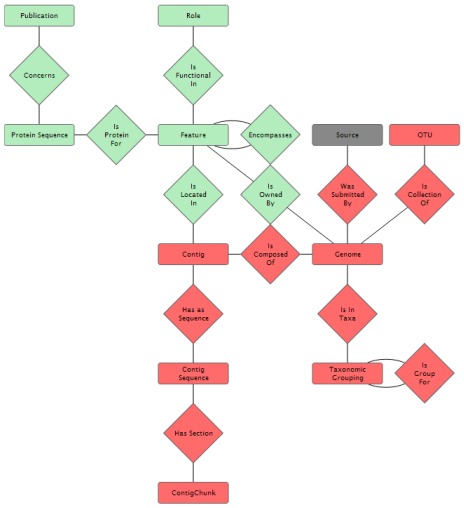


Figure C.1: Kbase schema snapshot

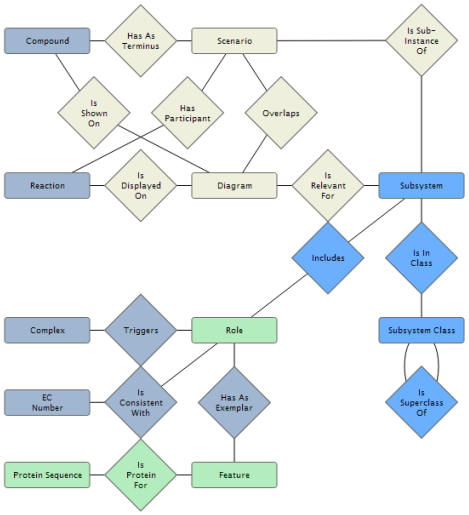


Figure C.2: Kbase Chemistry dataset spec 1

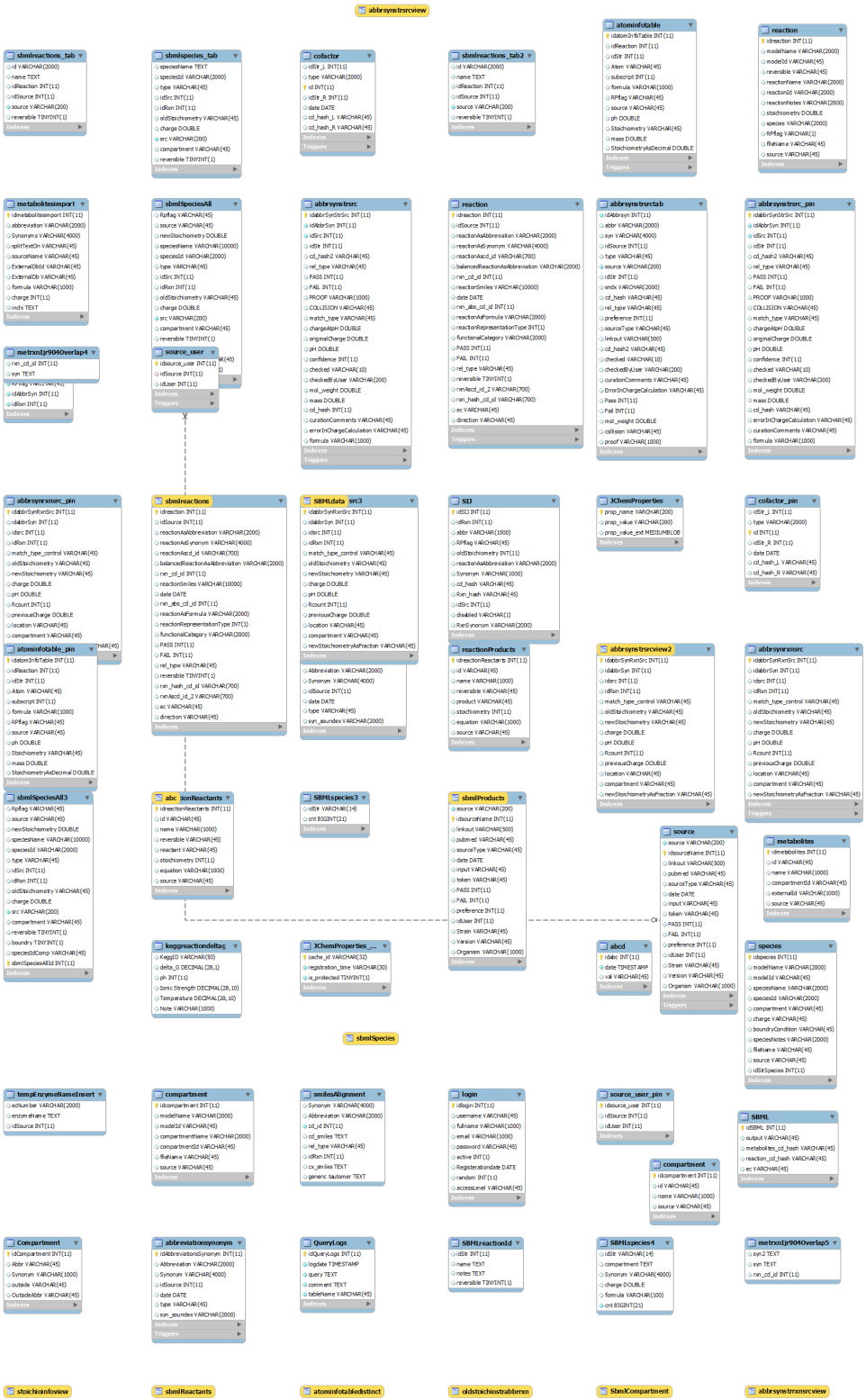


Figure C.3: MetRxn schema: The tables along with their column names are shown in blue and the views are shown in yellow.

D. Specific Aim 3: Incorporation of Atom Mapping Information for all Reaction Entries in MetRxn

D.1. Background

The cellular phenotype of an organism cannot be fully understood without gaining meaningful insight of the distribution of fluxes in its metabolic network [17, 18]. Metabolic Flux Analysis (MFA) [19] has emerged as the most powerful tool in quantifying *in vivo* fluxes in cells [20] leading towards varied applications in the fields of metabolic engineering, biotechnology and medicine [21-23]. Using stable radio isotopes (such as ^{13}C, ^{15}N etc.), labeled substrates are allowed to be ingrained in the backbone of cellular metabolites with distinct labeling patterns (i.e. isotopomers). The isotopomers are detected by mass spectroscopy (MS) or nuclear magnetic resonance (NMR). From the relative isotopic abundance in each metabolite, an indirect estimate on the flux patterns in the metabolic network could be made. Several mathematical models have been developed to correlate the flux distribution to isotopomer abundances [24-27]. Nevertheless, at the heart of each model lies an (usually non-linear) optimization protocol to estimate the flux distribution which minimizes the sum of squared residuals (SSR) in isotopomer abundances. Sensitivity analysis is usually performed to verify whether the measured flux distribution is within allowable statistical error, and confidence intervals of each flux are also calculated [28].

An important feature of the mathematical models for flux analysis is the high redundancy in isotopomer labeling measurements as compared to the number of free fluxes in the model that are required to be estimated. This redundancy is however dependent on the choice of labels used for the experiments [29]. Another potential cause for such redundancy is the relative small size of the metabolic networks used for the mathematical models. An opportunity thus arises in utilizing this data redundancy for expanding the scope of MFA to estimate fluxes for entire genome-scale models. With rationale driven optimization in choosing the correct combination of complementary labels, fluxes in genome-scale networks can measured with high fidelity. Using the Openflux algorithm [30] which uses the Elementary Mode Analysis approach [24], we will use our genome-scale metabolic model to the estimate reaction fluxes, and calculate their confidence intervals. Subsequently, we will develop an optimization framework to determine complementary labeling and/or, specific isotopomer measurement strategies to improve on the confidence scores of the metabolite fluxes.

Atom mapping of metabolic reactions finds its application in finding new biotransformation routes, synthesis of new pathways through engineering, providing the isotopomer-mapping matrix for use in MFA as well as in numerous other applications in systems biology. Atom mapping information also helps avoid the traversal of biologically infeasible and meaningless routes during identification of novel biotransformation routes through pathfinding [31, 32]. Finding correct atom maps for the whole metabolic network using automated techniques becomes the primary challenge needs to be addressed prior to performing MFA. A number of efforts have addressed this challenge [33-37] with the most recent effort being from the MetaCyc group [38] wherein they formulated this problem as a Mixed Integer Linear Problem (MILP) to calculate the minimum number of edits needed in the transformation of one graph (i.e., the reactant graph) into another (i.e., the product graph). They demonstrated this methodology on 7501 reactions of the MetaCyc database with a very low error rate of just 0.9% (22 reactions) when compared to the manually vetted 2446 reaction atom mappings from Kyoto Encyclopedia of Genes and Genomes (KEGG) RPAIR database [3]. The authors claim that their approach is extremely efficient and that 87% of the models were solved in less than 10 s. They call this formulation as the minimum weighted edit-distance (MWED) metric. This formulation is very much similar to the formulation presented by [33] where they try to maximize the number of common edges between to graphs. This formulation is called the maximum common edge subgraph (MCES) problem and has the same computational complexity as the most common subgraph (MCS) problem. We develop a novel subgraph isomorphism algorithm with is tractable in polynomial time. Our algorithm outperforms previous efforts in all aspects of accuracy, time and resource utilization. The manuscript detailing the effort mentioned in Aim 3 is under preparation and the results are available online at http://metrxn.che.psu.edu.

D.2. Results

Research was initiated towards developing complete atom-mapping information for all entries in the MetRxn database. Since MetRxn currently includes 112 metabolic models and 8 databases, such an effort would provide atom mapping information for each of the metabolic model completely, thereby allowing us to perform MFA on each of the metabolic model. To this end we developed a novel polynomial run-time algorithm with complexity $\theta(\eta^3)$. Our algorithm is based on two conjectures wherein the first states that that if two given graphs are identical when reordered, then the given graphs are identical(Köbler, Schöning, & Torán, 1994). The second conjecture is of Integer factorization wherein we use prime numbers to uniquely identify and label each node in a graph. The Algorithm proceeds as follows: (i) Identify canonical labels for each atom in all metabolites, (ii) Rank order with prime numbers only for labels common to all compared metabolites (unique labels are assigned assign '1'), (iii) Reassign labels based on product of neighboring atom labels, (iv) Repeat until atom ranks do not change (assign final integer labels), (v) Identify all non-maximal (disconnected) subgraphs by common labels, (vi) Identify and keep the largest subgraph or fragment, (vi) Extend largest subgraph to maximum common subgraph using the A* search algorithm (Heinonen, Lappalainen, Mielikäinen, & Rousu, 2011). This procedure is illustrated in Figure D.1

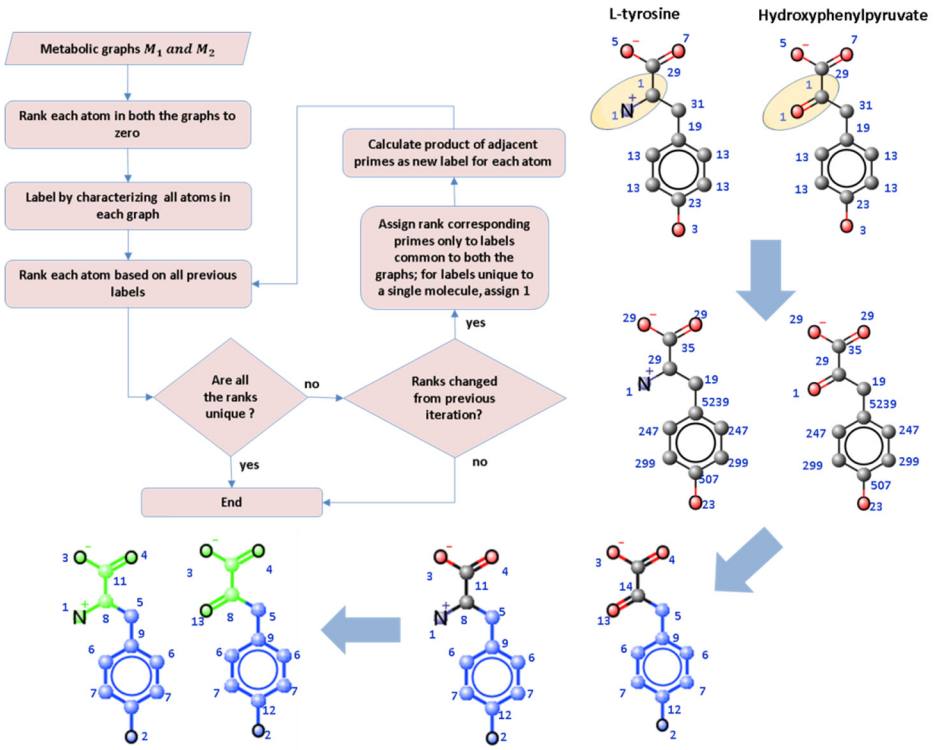


Figure D.1: Most common substructure using canonical labelling: The two subgraphs identified between L-tyrosine and Hydroxyphenylpyruvate, the larger subgraph is colored in blue. In the final step, we extend the subgraph size using the A* search methodology. The extended section is colored in green. The traversal and subgraph extension always starts from the largest fragment, and in this case, it starts from the vertex with index = 5. The numbers shown represent ranks generated from the labelling algorithm. The two non-equivalent atoms are stamped with different numbers 1 and 13.

To improve accuracy, we also combine the Floyd–Warshall algorithm with a complexity of $O(n^3)$ with the above mentioned steps. Using CLCA we identified and validated atom mappings for over 27,000 reactions in MetRxn. The average run time was around 14 milliseconds per reaction. CLCA always generates

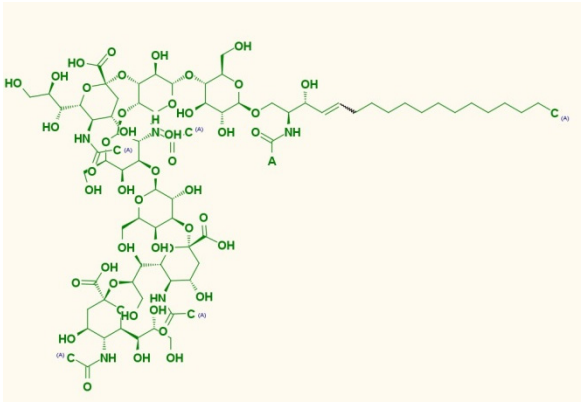


Figure D.2: Preliminary result of the mapping between two large molecules: The Figure shows the KEGG compound C06138 (Neu5Ac-alpha2->8Neu5Ac-alpha2->3Gal-beta1->3GalNAc-beta1->4(Neu5Ac-alpha2->3)LacCer) which will be mapped with C06139 in the reaction R05113.

homotopic mappings and avoids the computational overheads of identifying alternate solutions due to symmetric groups. Preliminary implementation on two extremely large molecules is shown below with the common substructures as shown in Figure D.2 and D.3, where the common vertices(atoms) and edges(bonds) between the graphs are colored green.

We plan to accelerate this effort by making atom-mapping information readily available on MetRxn. We validate the atom mappings in a two-prong effort. Firstly, we compare our results to DREAM, MetaCyc and KEGG. Secondly we incorporate atom maps into C^{13} metabolic flux analysis (MFA) models. Work on Metabolic flux analysis (MFA) has been limited in scale by the availability of atom mapping information since the non-linear equations [39] are constructed using mapping matrices [40] that trace the path of each atom and subsequently each isotopomer (isotope isomer) in a metabolic reaction. Initially the impact of scale-up of MFA models from their typical sizes of 70 reactions to 200 reactions was investigated. The generated atom mapping data was used to decompose the network into sub-networks using the EMU algorithm (Antoniewicz, 2007).

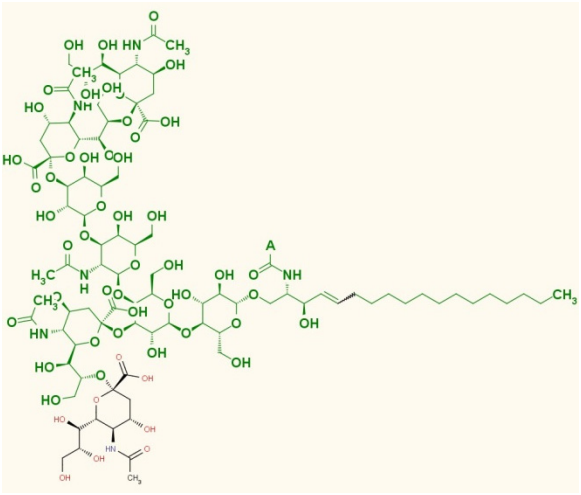


Figure D.3. Preliminary result of the mapping between two large molecules: The Figure depicts the mapped portion in green for the molecule C06139 (Neu5Ac-alpha2->8Neu5Ac-alpha2->3Gal-beta1->3GalNAc-beta1->4(Neu5Ac-alpha2->8Neu5Ac-alpha2->3)LacCer).

$$\min \phi = \sum_{i=1}^{N} \left(\frac{f_i(v) - x_i^m}{e_i} \right)^2$$

$$\text{s.t.} \quad S.v = 0$$
$$v_j^{lb} \le v_j \le v_j^{ub}$$
$$\text{MDV balances}$$

Fluxes were estimated by solving the following NLP problem: Where, $f_i(v)$ is the predicted MDV of measured fragments, x_i^m is the measured MDV of fragments, e_i is the associated experimental error, MDV balances are steady state balances on the different mass fractions of every metabolite in the network. On scaling the network up from 70 to 205 reactions we found that the flux ranges of glycolytic fluxes increased as a result of coexistence of both glycolysis and gluconeogenesis, the pentose phosphate pathway showed little change, both lower and upper bound of the TCA cycle increased, and the arginine degradation pathway, which was assumed to be inactive in the simplified model was found to have a non-zero lower bound, indicating definite activity.

E. Specific Aim 4: Creating Models using MetRxn: Metabolic Model Reconstruction of Plants

E.1. Background

Zea mays, commonly known as maize or corn, is an essential plant as a food crop, biofuel energy source, and model for studying plant genetics. Maize production is increasing at the greatest rate among all cereals [41] with a record 877 million tons produced in the 2011-2012 fiscal year [42]. Maize is a major C_4 plant that overcomes the inefficiencies of RuBisCO, to capture oxygen over the preferred carbon dioxide, by separating the carbon fixation process into two cell types: the bundle sheath and mesophyll cell. This separation allows for C_4 plants to have lower photorespiration rates [43], higher photosynthetic nitrogen

use efficiency [44], and higher net photosynthesis at high light intensities [45] than plants that do not separate the carbon fixation process. A thorough evaluation of the metabolic capabilities of maize will address challenges associated with its dual role as a food (e.g., starch storage) and biofuel crop (e.g., cell wall deconstruction), in addition to provide cues for improving nitrogen use efficiency. Integration of the recently completed maize genome [46], maize specific databases (i.e. MaizeCyc [47], MaizeGDB [48], and MetaCrop [49]), and the previous maize model (i.e *i*RS1563) [50] will allow for a high quality second generation genome-scale *in silico* model.

The development of a genome-scale model for maize is significantly challenging due to its genome size, which is approximately 14 times larger [51] than that of *Arabidopsis thaliana* [52]. The *i*RS1563 model contains 1,563 genes and 1,825 metabolites participating in 1,985 reactions from both primary and secondary metabolism of maize [53]. Due to the scarcity of information available during the first reconstruction effort, the previous model (*i*RS1563) suffers from: incomplete pathways (e.g., sterol biosynthesis, sphingolipid biosynthesis, ubiquinone biosynthesis and starch degradation), limited enzyme localization information, and approximate representation of photosynthesis reactions and electron transport chain. Finally, the *i*RS1563 model was developed as a generic maize model including all reactions known to occur in any cell or tissue within maize. The second-generation model has been developed for the leaf tissue including the distinction between the two cell types as displayed in Figure E.1. The bundle sheath cell contains seven compartments: the cytosol, mitochondrion, peroxisome, plastid, plasma membrane, thylakoid membrane, and vacuole. The mesophyll cell contains six compartments: the cytosol, mitochondrion, plastid, plasma membrane, thylakoid membrane, and vacuole. Compartmentalization for the second-generation model is based on maize specific experimental and proteomic data [54-57].

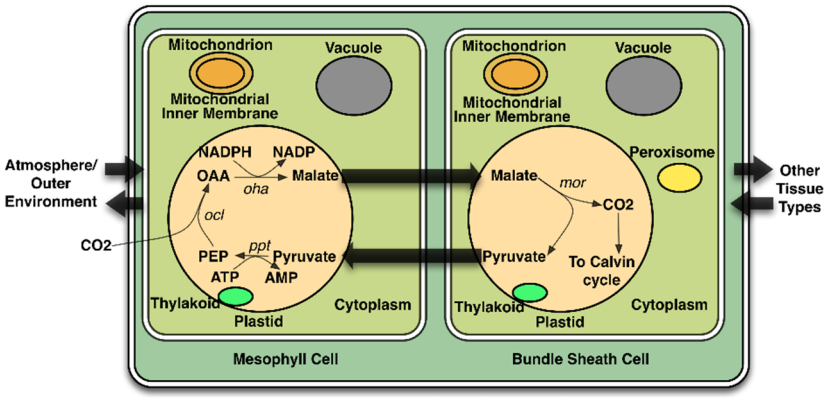


Figure E.1: Second-generation model schematic

E.2. Results

Under this aim we are collaborating with Dr. Bertrand Hirel's group from INRA Centre de Versailles-Grignon, France to reconstruct a second-generation model of maize, by creating five tissue specific tissues, namely the leaf, root, stalk, tassel, and seed. By utilizing available transcriptomic, proteomic and metabolomics data from literature and experimentally measured biomass components by Dr. Hirel's group, our goal is to reconstruct high quality tissue specific models that can be used to answer important biological questions on nitrogen and energy efficiency.

The second-generation maize leaf model was developed using gene, protein and reaction information from the *i*RS1563 model and databases, such as KEGG [58], MaizeCyc [47] and Metacrop [49]. Reactions and metabolites from different databases were compiled using MetRxn and compartmentalization was based on literature evidence [54-57]. The second-generation maize model includes 5,824 genes and 8,408 reactions, which is approximately 4 times the size of the *i*RS1563 model. The light reactions [59] and mitochondrial electron transport chain reactions [60] were updated to include the proton exchange of ATP synthase between compartments. Specific reactions were added to model glycerolipid synthesis [61-67], which to our knowledge is the first plant model to include specific glycerolipid synthesis. Thermodynamically infeasible cycles, generated due to the permissive inclusion of reactions in the model, were subsequently

identified and eliminated by first restricting directionality of reactions and then removing duplicate or generic reactions.

In order to improve the nitrogen use efficiency in maize, a comprehensive understanding of nitrogen metabolism within the organism is required. In order to simulate nitrogen conditions more accurately, gene-protein-reaction relationships are used to map the gene transcripts to proteins that are statistically expressed at a low level [68] to reactions that are turned-off in the model. The model was simulated at a wild-type, limited nitrogen, *gln1-3* mutant, and *gln1-4* mutant condition in the vegetative stage. Reaction fluxes were restricted for 90 reactions in the wild-type condition, 33 reactions in the limited nitrogen condition, 106 reactions in the *gln1-3* mutant, and 8 reactions in the *gln1-4* mutant. Reactions that are restricted in the wild-type condition mainly correspond to reactions known to occur under stress conditions. Biomass components were measured by the Hirel group to apply condition-specific biomass equations to the model.

A.) N^- Condition

In silico Metabolite Flux-Sum Change	*In vivo* Metabolite Concentration Change: Increasing from Wild-type	Decreasing from Wild-type
Increasing from Wild-type	2/13	0/58
Decreasing from Wild-type	11/13	58/58

B.) N^- Condition without Transcriptomic and Proteomic Data Included

In silico Metabolite Flux-Sum Change	*In vivo* Metabolite Concentration Change: Increasing from Wild-type	Decreasing from Wild-type
Increasing from Wild-type	0/13	4/58
Decreasing from Wild-type	4/13	33/58
No Change from Wild-type	9/13	21/58

C.) *gln1-3* Mutant

In silico Metabolite Flux-Sum Change	*In vivo* Metabolite Concentration Change: Increasing from Wild-type	Decreasing from Wild-type
Increasing from Wild-type	0/12	0/25
Decreasing from Wild-type	12/12	25/25

D.) *gln1-4* Mutant

In silico Metabolite Flux-Sum Change	*In vivo* Metabolite Concentration Change: Increasing from Wild-type	Decreasing from Wild-type
Increasing from Wild-type	0/3	0/8
Decreasing from Wild-type	3/3	8/8

Figure E.2: Comparison between the directional concentration change and simulated flux-sum change for metabolites in the: A.) limited nitrogen condition, B.) limited nitrogen condition with no transcriptomic or proteomic data included, C.) gln1-3 mutant condition, and D.) gln1-4 mutant condition.

The metabolomics data [68] was compared to the flux predictions within the model in each of the nitrogen conditions. The increasing or decreasing trend of the metabolite concentration was qualitatively compared to the flux-sum changes determined by the model. The flux-sum of a metabolite is a measure of the amount of flux through the metabolite. Overall, the model accuracy in correctly predicting the directional change in concentration levels is approximately 84% in the limited nitrogen condition, 68% in the *gln1-3* mutant condition, and 73% in the *gln1-4* mutant condition as displayed in Figure E.2. In comparison to the limited nitrogen condition when transcriptomic and proteomic data was not included, the accuracy of the model to predict the directional change of the metabolite concentration was reduced to 46%.

F. Specific Aim 5: Curating Metabolic Models using MetRxn

F.1 Background

New models are being added to MetRxn as they are published or made available to us. It is available as a web-based resource at http://metrxn.che.psu.edu. Increasingly, metabolite and reaction information is organized in the form of genome-scale metabolic reconstructions that describe the reaction stoichiometry, directionality, and gene to protein to reaction associations. A key bottleneck in the pace of reconstruction of new, high-quality metabolic models is the inability to directly make use of metabolite/reaction information from biological databases or other models due to incompatibilities in content representation (i.e., metabolites with multiple names across databases and models), stoichiometric errors such as elemental or charge imbalances, and incomplete atomistic detail (e.g., use of generic R-group or non-explicit specification of stereo-specificity).

There have already been a number of efforts aimed at addressing some of these limitations. The Rhea database, hosted by the European Bioinformatics Institute, aggregates reaction data primarily from IntEnz [69] and ENZYME [70], whereas Reactome.org is a collection of reactions primarily focused on human metabolism [5, 71]. Even though they crosslink their data to one or more popular databases such as KEGG, ChEBI, NCBI, Ensembl, Uniprot, etc., both retain their own representation formats. More recently, the BKM-react database is a non-redundant biochemical reaction database containing known enzyme-catalyzed reactions compiled from BRENDA, KEGG, and MetaCyc. The BKM-react database currently contains 20,358 reactions. In addition, the contents of five frequently used human metabolic pathway databases have been compared [72]. An important step forward for models was the BiGG database, which includes seven genome-scale models from the Palsson group in a consistent nomenclature and exportable in SBML format [73-75]. Research towards integrating genome-scale metabolic models with large databases has so far been even more limited. Notable exceptions include the partial reconciliation of the latest *E. coli* genome scale model *i*AF1260 with EcoCyc [76] and the aggregation of data from the *Arabidopsis thaliana* database and KEGG for generating genome-scale models [77] in a semi-automated fashion. Additionally, ReMatch integrates some metabolic models, although its primary focus is on carbon mappings for metabolic flux analysis [78]. Also, many metabolic models retain the KEGG identifiers of metabolites and reactions extracted during their construction [79]. An important recent development is the web resource Model SEED that can generate draft genome-scale metabolic models drawing from an internal database that integrates KEGG with 13 genome scale models (including six of the models in the BiGG database) [80]. All of the reactions in Model SEED and BiGG are charge and elementally balanced.

F.2. Results

Using Gene Essentiality and Synthetic Lethality Information to Curate Existing Metabolic Reconstructions

Essentiality (ES) and Synthetic Lethality (SL) information identify combination of genes whose deletion inhibits cell growth. Essentiality and SL analyses refer to identifying sets of gene deletions (single, double and higher order thereof) that render the strain nonviable. Essentiality analysis identifies the list of genes, each of which when deleted *in silico*, limits the biomass flux to lower than 10% of its theoretical maximum. SL analysis identifies the list of *in silico* gene pairs (and higher order) whose removal constrains the biomass flux to lower than the aforesaid essentiality criterion. These analyses serve the dual purpose of model refinement (by comparing with available *in vivo* knockout information) and prediction for identifying genes (or combination of genes) whose knockouts could potentially be lethal. This information is important for both identifying drug targets for tumor and pathogenic bacteria suppression and for flagging and avoiding gene deletions that are non-viable in biotechnology, such as during strain design. In this study, we performed a comprehensive ES and SL analysis of two important eukaryotic models: *S. cerevisiae* and CHO cells so as to propose model changes that remedy inconsistencies with data model predictions [81]. ES and SL analyses are supplemented by auxotrophy information to help elucidate the cause (*i.e.*, nutrient or biomass precursor deficiency) for lethality.

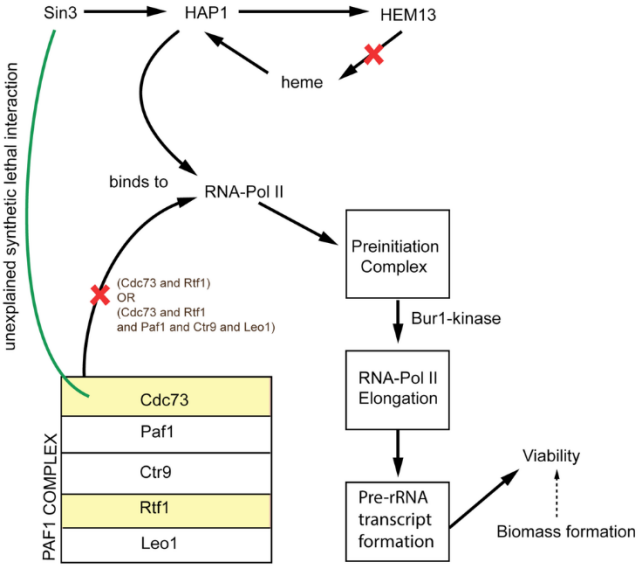


Figure F.1: Schematic showing the non-metabolic lethal interaction between *Cdc73* and *Hem13* gene in yeast. The red crosses represent the loss of function upon deletion of *Hem13* and *Cdc73* genes.

For CHO 1.2, we identified eight instances where model and experimental data do not match. Upon supplementing this mismatched set with another 11 cases of model and experiment discrepancies from the mouse model [82], we suggested 14 additional (single, double and higher) gene deletion experiments for

maximally resolving mutant growth phenotypes in CHO cell lines. The absence of a comprehensive single-gene knockout database for Chinese Hamster Ovary (CHO) cells (unlike yeast) makes the assessment CHO 1.2 genome-scale model [83] more difficult. Therefore, we supplemented limited experimental data with predicted lethal gene deletions based on the most recent mouse model [83] and gene knockout studies in mouse embryonic stem cells [84] that exhibited high degree of sequence similarity (functionality of the encoded protein is at least 70% conserved across all mammalian systems [84]) with the CHO cell genome. Any inconsistencies between mouse and CHO cell lethality was used as an opportunity to correct the CHO model. Eight GPR modifications were proposed for CHO 1.2 in order to address and reconcile five ESG cases to GG, three GES cases to ESES, three SL2ES cases to ESES and one ESSL2 case to SL2SL2. In addition, we proposed a number of gene deletion experiments to verify non-intuitive synthetic lethal gene combinations. Reaction level essentiality analysis *in silico* revealed 90 essential reactions. Utilizing the GPR associations for these reactions, 57 essential genes were identified for growth under aerobic minimal essential media.

The proposed model modifications on Yeast 7.11 involve 50 literature-supported changes that improve the sensitivity and specificity of Yeast 7.11 by 2.66% and 20.4% respectively and decrease the false viable rate (FVR) by 8.42%. Overall, we reconciled 50 growth discrepancies between model and experiment. Twelve ESG cases were identified that form ESSL2 inconsistencies in combination with other non-metabolic genes. For example, gene *HEM13* whose deletion causes an ESG discrepancy has a non-metabolic function in chromatin assembly and interacts with RNA-polymerase II in transcription. It forms a synthetic lethal with *CDC73* [85] (cell division cycle gene) due to the inability to form the pre-rRNA transcript upon simultaneous deletion of the two. We propose a possible interaction schematic (see Figure F.1) explaining the cause for the lethal interaction based on information from existing experimental studies [86-88].

Metabolic flux analysis for genome-scale reconstructions

Metabolic flux analysis (MFA) helps estimate intracellular fluxes, thereby elucidating flux divisions at key metabolic branch points, such as that between glycolysis and the pentose phosphate pathway or that between fermentation and respiration. Usually metabolic models used in 13C MFA employ a limited number of reactions primarily from central carbon metabolism, choosing to omit degradation pathways, complete cofactor balances, and atom transition contributions for reactions outside central metabolism. We sought to assess the impact of scaling up 13C MFA mapping models to a genome-scale model, using the yeast iAF 1260 model as a chassis. Labeling data for 17 amino acid fragments obtained from cells fed with glucose labeled at the second carbon was used to obtain fluxes and ranges. Metabolic fluxes and confidence intervals are estimated by minimizing the sum of square of differences between predicted and experimentally measured labeling patterns using the EMU decomposition algorithm.

We find that both topology and estimated values of the metabolic fluxes remain largely consistent between core and GSM model. Stepping up to a genome-scale mapping model leads to wider flux inference ranges for 20 key reactions present in the core model. The glycolysis flux range doubles due to the possibility of active gluconeogenesis, the TCA flux range expanded by 80% due to the availability of a bypass through arginine consistent with labeling data, and the transhydrogenase reaction flux was essentially unresolved due to the presence of as many as five routes for the inter-conversion of NADPH to NADH afforded by the genome-scale model. A non-zero flux for the arginine degradation pathway

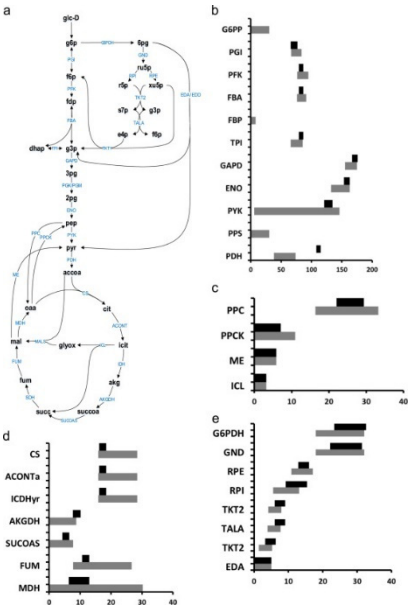


Figure F.2: Flux distribution comparison for core model and GSM model. (a) Schematic representation of all reactions and metabolites involved in central metabolism of *E. coli*. Comparison of flux ranges (in mmol/dmol-glucose) using (■) core model and (Fig. 2) GSM model for (b) glycolysis and gluconeogenesis, (c) anaplerotic reactions and glyoxylate shunt, (d) TCA cycle, (e) PPP and ED pathway.

was identified to meet biomass precursor demands as detailed in the iAF1260 model. Inferred ranges for 81% of the reactions in the genome-scale metabolic (GSM) model varied less than one-tenth of the basis glucose uptake rate (95% confidence test). This is because as many as 411 reactions in the GSM are growth coupled meaning that the single measurement of biomass formation rate locks the reaction flux values. This implies that accurate biomass formation rate and composition are critical for resolving metabolic fluxes away from central metabolism and suggests the importance of biomass composition (re)assessment under different genetic and environmental backgrounds. In addition, the loss of information associated with mapping fluxes from MFA on a core model to a GSM model manifested in the TCA cycle and associated fluxes, due to the presence of alternate pathways between metabolites. For instance, the lower bound of AKGDH decreased to zero due to the presence of multiple alternate pathways between glutamate and succinate. The conversion of glutamate to succinate via γ-aminobutyrate and γ-glutamylsuccinate showed similar flux ranges as AKGDH indicating the inability of 13C MFA to resolve between these alternative pathways. Thus, this study proved that the application of GSM models to 13C MFA will allow the use of closed cofactor balances without the risk of altering the actual flux distribution predicted using the flux estimation procedure. This will enable identifying metabolic bottlenecks leading to more informed metabolic engineering interventions that improve the yield of target products.

Elucidation of Photoautotrophic Carbon Flux Topology using Genome-scale Carbon Mapping Models

We have also applied genome-scale genome-scale isotopic instationary ^{13}C-Metabolic Flux Analysis (INST-MFA) to elucidate photoautotrophic metabolism in *Synechocystis*. Reactions capable of carrying flux in *i*Syn731 [89] are identified via FVA using extracellular flux measurement data [90]. The corresponding GSMM model *i*mSyn617 includes all carbon-balanced reactions Atom mapping for reactions shared with *E. coli* is derived from *i*mEco726 [91] and the remaining reactions are mapped using the CLCA algorithm or based on reaction mechanism when available. A customized algorithm is developed with improved scalability and memory efficiency leading to a 48% reduction per iteration in the computational time required to simulate of metabolite labeling dynamics in larger networks. INST-MFA is performed to identify a suitable flux distribution accurately recapitulating the labeling distribution and dynamics of 15 central metabolites obtained during photoautotrophic growth of *Synechocystis* with 50% ^{13}C-labeled bicarbonate as the tracer [90]. In response to degeneracy in the metabolic network and experimental errors, 95% confidence intervals were also determined using the established procedure [91, 92] to identify flux ranges for all reactions.

Upon evaluating the significance of the improved recapitulation afforded by *i*mSyn617 using the F-test, the F-statistic is 1.335 ($p = 0.012$). In comparison, the corresponding F-statistic for scale-up in *E. coli* was 0.152 ($p = 0.999$) indicating that the core model accounts for the carbon paths necessary to recapitulate the labeling data used in that study [91]. The increased uncertainty of flux estimation was attributed to the inclusion of alternate paths with identical atom mapping information. In contrast, the statistical significance associated with model scale-up in this study implies that unique and often surprising insights into the carbon flows under phototrophic growth are obtained by the re-analysis of an existing dataset using a detailed description of the entirety of metabolism in *Synechocystis*. Flux elucidation of photoautotrophic growth of *Synechocystis* using *i*mSyn617 reveals that *Synechocystis* deploys a carbon efficient metabolism enabling maximal conversion of fixed carbons to biomass precursors with minimal production of organic acids and glycogen. This is in contrast to heterotrophic bacteria such as *E. coli* where 35% of the taken-up glucose is secreted as acetate [93] resulting in a 30% biomass yield loss from the theoretical maximum biomass yield [94]. The flux ranges estimated in this

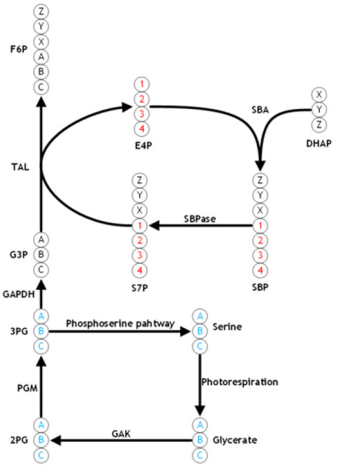


Figure F.3: Recycling of conserved moieties within central metabolism. The conserved E4P moiety generated due to the interaction between TAL from the non-oxidative PP pathway and SBA and SBPase from the regeneration phase of the CBB cycle is indicated in red whereas the conserved triose phosphate moiety recycled between the serine biosynthetic pathway, photorespiration, and lower glycolysis is indicated in blue.

study provide a comprehensive set of essential and dispensable metabolic reactions in *Synechocystis* under photoautotrophic growth conditions to serve as a guideline for editing photosynthetic prokaryotic genomes. The estimated flux ranges reveal that net carbon fixation accounts for only 88% of the assimilated bicarbonate. The remaining 12% is fixed by PPC, but is subsequently oxidized to CO_2 via malic enzyme, TCA cycle, and peripheral metabolic reactions. These carbons are not recycled by the CBB cycle and are therefore off-gassed. This inability to recycle these carbons via the CBB cycle is identified as a target to improve upon in photosynthetic carbon fixation. It is unclear from this analysis whether this is caused by a rate-limiting enzyme in the CBB cycle or a paucity of available NADPH and ATP as the fluxes through the photosynthetic light reactions and oxidative phosphorylation are poorly resolved by INST-MFA.

Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models

In flux balance analysis (FBA), one of the fundamental constraints is the steady-state mass balance equation, which quantifies the conservation of component balance. As a prerequisite for quantitative predictions, all reactions must be component and charge balanced. This principle of mass balance also applies to the biomass reaction, which expresses biomass as a defined ratio of macromolecules synthesized from metabolites [95]. A GSM model describes in quantitative terms the substrate-to-biomass conversion, from mmol of substrates to gram dry cell weight of cells. By definition the biomass produced must have a molecular weight (MW) of 1 g $mmol^{-1}$ in order to quantitatively compare biomass formation with the observed growth yields or specific growth rates. FBA identifies optimal solutions that strike a balance between biomass formation and ATP production, thus any discrepancy in biomass weight may tilt the balance to have a disproportionate influence on FBA derived flux predictions. However, the standard is rarely verified in the current practice and the chemical formulae of biomass components such as proteins, nucleic acids and lipids are often represented by undefined side groups (e.g. X, R). We developed a systematic MILP-based procedure called 'minimum inconsistency under parsimony (MIP)' for checking the biomass weight and ensuring complete mass balance of a model [96].

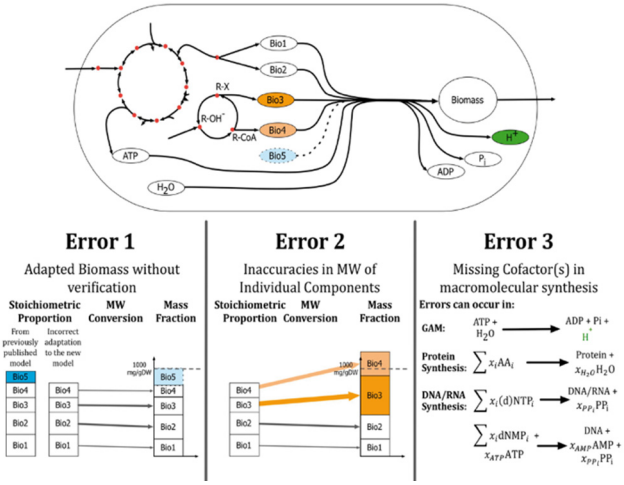


Figure F.4: Three sources of errors in the biomass reactions: (i) biomass reactions generated by automated platforms or adapted from other models with biomass components deleted ('Bio5') or newly added; (ii) inaccurate stoichiometric coefficients in the biomass reaction ('Bio3', 'Bio4') partially due to the existence of undefined side-groups (e.g., 'R' and 'X'); and (iii) missing cofactors in macromolecular synthetic reactions, such as proton in GAM ('H+'), water in protein synthesis and pyrophosphate in DNA and RNA syntheses.

MIP formulates the elemental balance as an optimization problem that solves for the chemical formulae of generic metabolites by minimizing the inconsistencies using the information of known metabolites. The MIP solution provides guidelines for resolving all imbalances in a model. We identified significant departures after examining 64 published models. The biomass weights of 34 models differed by 5–50%, while 8 models have discrepancies >50%. In total 20 models were manually curated. By maximizing the original versus corrected biomass reactions, flux balance analysis revealed >10% differences in growth yields for 12 of the curated models. We identified three primary sources leading to inaccuracies in the biomass MW (see Figure F.4). First, a subset of models uses biomass reactions that were made by automated platforms or adapted from other models (e.g. the models for *Bacteroidesthetaiotaomicron*, *Faecalibacteriumprausnitzii*, and Yeast 7). In the absence of experimental data, the mass fractions for the biomass reaction were simply obtained by uniformly normalizing over all biomass components. Second, for some models we found inconsistent stoichiometric coefficients in the biomass reaction because the MWs of macromolecules used for calculating the coefficients were not the same as the actual MWs implied by their elemental balance. For example, in

the *Yarrowialipolytica* model, our MIP procedure calculated MWs for phospholipids that were ~100× larger than the MWs used in the original model construction (Pan and Hua, 2012) yielding a biomass MW of 30 g $mmol^{-1}$. The reason for this was that the model lipid building blocks such as 1-acyl-sn-glycerol 3-phosphate were synthesized as polymers with 100-mers (e.g. in the reaction for glycerol 3-phosphate acyltransferase), instead of monomers as in other models. Inconsistent stoichiometric coefficients were also found in the models for *Corynebacteriumglutamicum*, *Clostridiumacetobutylicum*, and *Eubacteriumrectale*. A probable reason for the errors is the lack of the application of a procedure to ensure complete mass balance and verify the biomass MW. Some GSM models included metabolites with undefined side-groups (e.g. acyl groups in lipids) that complicate the estimation of the MWs of macromolecules. Finally, small molecules in macromolecular synthesis reactions were sometimes missing, (e.g. missing proton in the growth-associated maintenance (GAM), H_2O in protein synthesis, and pyrophosphate in DNA or RNA synthesis). This was observed in the models for *B. subtilis*, *C. acetobutylicum*, *C. glutamicum*, *E. rectale* and *Pichiapastoris*. Figure F.4 pictorially illustrates the sources of error in the calculation of biomass composition.

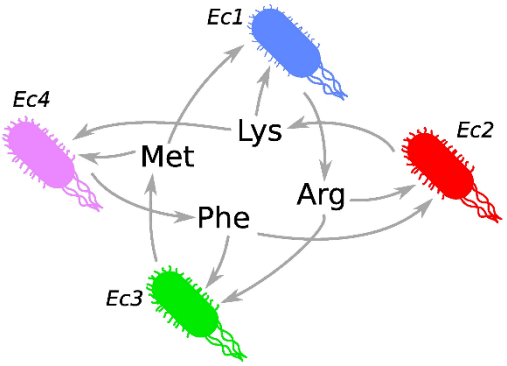
G. Specific Aim 2: MetRxn-driven metabolic reconstruction and analysis of microbial communities

G.1. Background

Most microorganisms exist in nature as integrative and interactive communities, and are responsible for driving biochemical cycles of nitrogen and carbon [97], and playing central roles in human health and disease [98, 99]. Members of such microbial communities can interact by unidirectional or bidirectional exchange of metabolites, giving rise to interactions such as mutualism, commensalism, parasitism, or competition [100, 101]. A classic example of these examples would be day-night or seasonal variations, where inter-species interactions and their temporal changes play pivotal roles in shaping community composition and function [102, 103]. However, the nature of these interactions and the dynamic variations therein are not well understood, necessitating model frameworks that help describe the lesser understood aspects of metabolism in microbial communities. Existing frameworks employ optimization frameworks that maximize a single objective function related to an individual species, which cannot always capture the multi-level nature of decision making. Furthermore, simply adapting dynamic single-species modeling frameworks such as d-FBA [104] is not trivial owing to the increased complexities and missing information about interspecies interactions in a changing environment.

G.2. Results

We developed the SteadyCom [105] optimization framework for predicting metabolic flux distributions consistent with the steady-state requirement. As opposed to earlier developed community modeling algorithms such as joint FBA, OptCom, d-OptCom and CASINO, SteadyCom directly imposes time-averaged equality of growth rates and apportions ATP maintenance (ATPM) requirements across different microbes in accordance with specific growth. The potential of SteadyCom to predict species abundance and perform constraint-based analysis in community models with community steady-state implemented is demonstrated here using the hypothetical case of the co-growth of four *E. coli* triple mutants using the genome-scale metabolic reconstruction *E. coli* *i*AF1260 [106]. The community consists of four *E. coli* mutants (*Ec1*, *Ec2*, *Ec3* and *Ec4*), each auxotrophic for two amino acids and devoid of the exporter of one amino acid (see Figure G.1). Each mutant competes with another for the amino acids produced by the



Strain	Genotype	Synthesis				Export			
		Arg	Lys	Met	Phe	Arg	Lys	Met	Phe
Ec1	*ΔlysAΔmetAΔyddG*	✓			✓	✓			
Ec2	*ΔargHΔpheAΔyjeH*		✓	✓			✓		
Ec3	*ΔargHΔlysOΔpheA*		✓	✓				✓	
Ec4	*ΔargOΔlysAΔmetA*	✓			✓				✓

Figure G.1: A hypothetical microbial community of four *E. coli* mutants. Each *E. coli* mutant is auxotrophic to two amino acids and produces one amino acid that is essential to the community. The genotype and ability to synthesize and export the focus amino acids are displayed.

other two mutants. Thus, co-growth is theoretically possible and every mutant is essential for community survival and growth. The maximum growth rate predicted by joint FBA was 0.572 h^{-1} while the prediction by SteadyCom was 0.736 h^{-1}. This significant deviation was found to be a result of the non-growth-associated ATPM requirement in the model. In joint FBA, the predicted flux distribution needed to fulfill the ATPM requirement for four units of biomass, leading to the underestimation of the maximum growth rate. In contrast, the flux distribution predicted by SteadyCom satisfied the ATPM requirement for one unit of biomass in total. The allowable ranges of the relative abundance of the mutants at ≥ 90% of the maximum community growth rate computed by flux variability analysis (FVA) indicate the essentiality of each mutant for growth using SteadyCom. The ranges converge to a unique community composition as the community growth rate increases to its maximum. In contrast, joint FBA optimizing for an unweighted sum of biomass predicts that each of the mutants can have abundances ranging from 0 to 100% for ≤ 99% maximum community growth and only the growth of *Ec2* and *Ec3* are necessary at 100% maximum community growth

SteadyCom was also applied to a gut microbiota model consisting of nine species to predict the composition of gut microbiota given the dietary information. A community model consisting of nine microbes present in the human gut with available genome-scale metabolic reconstructions was compiled. The organisms include one species in the phylum Bacteroidetes, five species in Firmicutes (two Clostridia and three lactic acid bacteria), two species in Proteobacteria and one species in Actinobacteria (*B. adolescentis*). In the assembled community model, *B. thetaiotaomicron* and *F. prausnitzii*are the only organisms able to digest dietary fiber. Using the nine proxy models, SteadyCom was able to predict the universal dominance of Bacteroidetes and Firmicutes with non-zero abundances for Actinobacteria and Proteobacteria given a typical diet [27]. With randomizing the uptake rates of microbes, an abundance profile of the phylum proxies similar to the experimental phylum distribution was predicted. A recent study comparing vegans and omnivores from an urban USA area found surprisingly similar gut microbiota compositions between the two groups [28]. Both the community growth rates predicted by SteadyCom (see Figure G.1) and the maximum growth rates for each species predicted by joint FBA (see Figure G.2) given community uptake rates based on the consumption and chemical composition of the average American diet, lie in the range of the intestinal microbial growth rates reported (i.e. 0.02–0.25 h^{-1}) [29]. This consistency supports the validity of constraint-based modeling frameworks based on the mass balance of biochemical conversion and the potential for qualitative and quantitative predictions of gut microbiota metabolism.

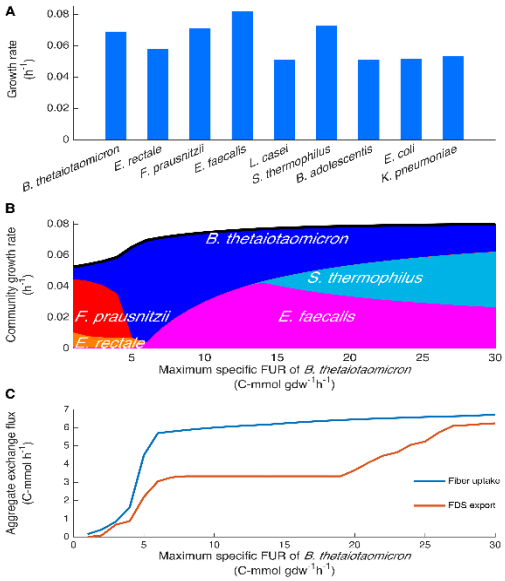


Figure G.2: Simulation of the gut microbiota model subject to the estimated average American diet. (A) The maximum possible growth rates were predicted for each species using joint FBA by maximizing the biomass reaction of each species individually. (B) The maximum community growth rate (the black curve) and species composition (filled area) were predicted by SteadyCom at varying maximum specific fiber uptake rate (FUR) of *B. thetaiotaomicron*. (C) Aggregate fiber uptake and fiber-derived substrate (FDS) export by *B. thetaiotaomicron* that are required for maximum community growth were calculated using FVA.

d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities [107]
To capture the multi-level nature of decision making in microbial communities, we had previously developed OptCom [108] that uses a multi-level and multi-objective optimization formulation capable of capturing both species- and community-level fitness criteria. Recently we developed d-OptCom (dynamic OptCom) for the multi-objective dynamic analysis of the microbial communities. To this end, new time-

dependent constraints representing the conservation of mass for the biomass of each species and shared metabolites with available uptake kinetics are added to the outer problem (see Figure G.3). The upper bound on the uptake rate of each shared metabolite is determined by using the uptake kinetic expressions incorporated as additional constraints in the outer problem. The inter-organism flow constraint (from the original OptCom procedure) [108] is used instead of conservation of mass equations for the shared metabolite without any uptake kinetics. The uptake/export rates of the shared metabolites are determined by the outer objective function, however, they act as parameters for the inner problems of the respective community members. This multi-level optimization problem can be recast as a nonlinear problem or a mixed-integer nonlinear problem by using the strong duality or KKT conditions for the inner problems, respectively. In both cases the problem is, in general, nonconvex due to the presence of uptake kinetic expressions and conservation of mass equations.

d-OptCom incorporates the dynamic mass balance equations and substrate uptake kinetics and enables the direct assessment of the shared metabolites and biomass concentrations in a given community. For example, it is possible to maximize the total biomass concentration of the community instead of maximizing the combined biomass flux of the community as in the original OptCom procedure [108]. Alternatively, one can maximize (minimize) the concentration of a desired (undesired) shared metabolite, or minimize deviation from a target time-dependent concentration pattern as the engineering objective. Furthermore, this extends the concept of Descriptive OptCom [108] to a dynamic context (i.e., Descriptive d-OptCom) where constraints on actual extracellular concentrations (e.g., the biomass composition of the community) can be added to the outer problem in order to determine the dynamic changes in optimality levels of each community member.

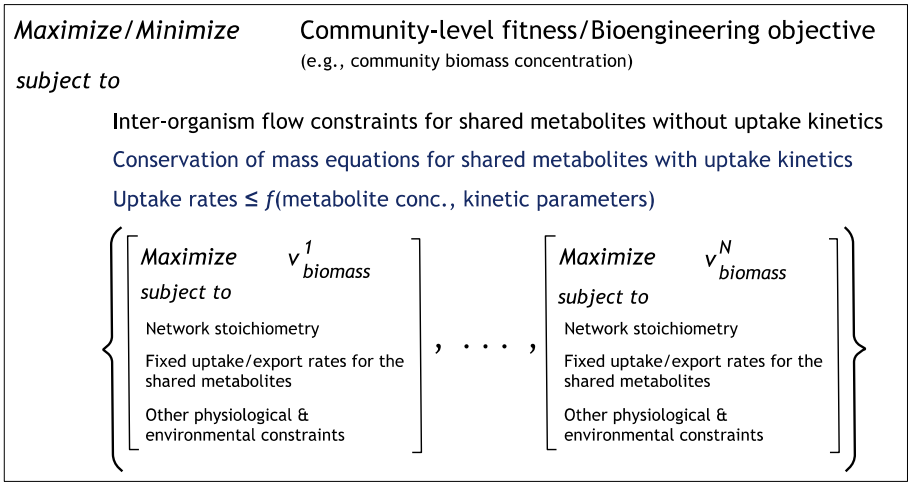


Figure G.3. Optimization structure of d-OptCom. Dynamic equations representing the conservation of mass for each shared metabolite with available uptake kinetics are added as new constraints to the outer problem. The upper bounds on the uptake rates are determined by using the uptake kinetic expressions.[107]

We model the competition between *Rhodoferax ferrireducens* and *Geobacter sulfurreducens* in subsurface anaerobic environments [109]. In particular, given the time course *G. sulfurreducens* biomass fractions data under different conditions, we are using Descriptive OptCom within each time interval to gain insights into how the optimality levels for the participating community members change over time. Now we have examined the impact of the addition of an acetate producer[110-112] (*Shewanella oneidensis*) to the *G. sulfurreducens*-*R. ferrireducens* community by using the d-OptCom procedure. The combined uranium reduction capability of *S. oneidensis* [113], and *G. sulfurreducens* promise a more effective bioremediation strategy. The dynamic analysis of the uranium-reducing communities in the Rifle site with an additional member showed that the incorporation of kinetic information can significantly sharpen the inference of inter-organism metabolite trafficking due to the concentration limits of the shared metabolite and/or the relative differences in the uptake efficiencies of community members. In addition, this analysis revealed

that addition of a new member to an existing community can significantly affect the behavior and composition of the community exemplified by the dominance of *S. oneidensis* in the long run.

We also used d-OptCom to model and analyze the dynamics of a synthetic mutualistic relationship between pairs of auxotrophic E. coli mutants. Wintermute and Silver[114] previously examined the co-growth of several combinations of 46 mutant strains, where the deletion(s) in each strain blocks the biosynthesis of a biomass precursor such as an amino acids, nucleotides or co-factors, thereby making them unable to grow in minimal medium. We examined whether d-OptCom is capable of recapitulating the co-growth of cooperating partners. To this end, we selected three such mutant pairs comprised of four genes involved in the production of different amino acids with available uptake kinetics. These pairs include (*ΔargH, ΔlysA*), (*ΔlysA, ΔtrpC*) and (*ΔmetA, ΔilvE*) where the deletion of *argH, lysA, trpC, metA* and *ilvE* block the production of L-arginine, L-lysine, L-tryptophan, L-methionine and L-isoleucine, respectively. The selected mutant pairs expand their own pool of required amino acids by aiding the growth of their conjugate partners, thereby enhances the co-growth. This cooperative behavior was captured by d-OptCom as it simultaneously takes into account species and community-level fitness functions enabling it to identify the impact of inter-species interactions on the shared metabolite and biomass concentrations.

H. Specific Aim 3: Extracting Knowledge using MetRxn: Pathway Prospecting and Synthetic Biology

H.1. Kinetic Modeling

H.1.1. Background

The primary attraction behind constraint-based models is the minimal amount of biochemical knowledge required to make quantitatively predictive inferences about network behavior. Despite their many successes in metabolic system characterization and biological applications, their major limitation is the inability to comment on the metabolite/enzyme concentrations and interactions, and to describe the transient nature of metabolism. These caveats can be addresses by using kinetic metabolic models, which use ordinary differential equations to obtain time-dependent metabolite concentrations and reaction fluxes. These models require knowledge of stoichiometry, enzyme kinetics and efficient parameter estimation. The major challenge in their construction lies in deducing reaction mechanisms and elucidating kinetic enzyme parameters. Current methods use lumped kinetic expressions to describe the relationship between metabolite concentrations, enzyme activities and reaction fluxes, and to estimate the kinetic parameters (those not been measured experimentally) by minimizing the error between model prediction and experimental measurements (usually of metabolite concentrations and/or steady-state metabolic fluxes) [115, 116]. Furthermore, as these models are usually constructed using data taken from a single strain, they do not accurately capture the behavior of perturbed strains.

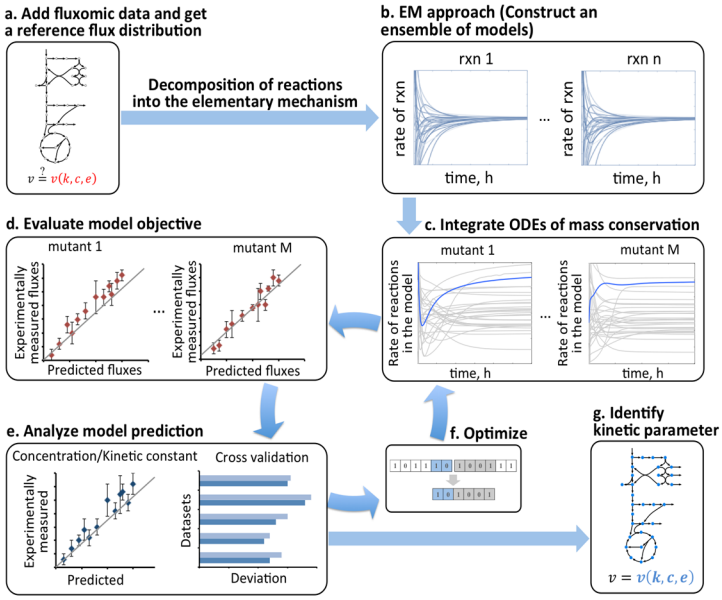


Figure H.1.1: A schematic representation of the optimization-driven parameter identification method.

H.1.2. Results

Optimization-driven parameter estimation

In the context of this aim, we introduced a stepwise optimization procedure for kinetic parameter estimation of a given metabolic network by making use of multiple fluxomic data. The main limitation in the construction of kinetic models is the absence of available kinetic parameter values with reliable rate expressions. The recently developed Ensemble Modeling (EM) [117] approach holds promise to address some of these challenges by decomposing metabolic reactions into elementary reaction steps thus bypassing the need for identifying the lumped form of kinetic rate laws. To this end, recently the EM approach has been combined with an optimization-driven parameter identification method [118] to elucidate kinetic parameters upon integration of multiple omic (i.e., fluxomic and metabolomic) data, describing metabolic fluxes as a function of metabolite concentration and enzyme activity, as shown in Figure H.1.1.1. First, a steady-state flux distribution is obtained by imposing the available fluxomic data and refining the flux ranges for a reference strain (Figure H.1.1.1a). Next, in accordance with EM procedure, each reaction is decomposed into its elementary mechanistic steps and the model parameters (i.e., reactions reversibilities and enzyme fractions) are uniformly sampled within identified feasible ranges. Sampling of model parameters provides an ensemble of models all of which are able to predict the same reference steady-state flux distribution (Figure H.1.1.1b). For a given set of kinetic parameters from the sampled models in the ensemble, the ODEs representing the conservation of mass are integrated until reaching steady-state (Figure H.1.1.1c). The model integration allows for the evaluation of the objective function of the optimization problem which is deviation from experimental flux measurements (Figure H.1.1.1d). The model predictions are validated by a comparison between the available metabolomics, kinetic constants and performing cross-validation tests (Figure H.1.1.1e). In order to improve model fitness, the optimization procedure provides a new set of model parameters, based on the feedback receiving from predictive performance of the model (Figure H.1.1.1f). Ultimately, a set of kinetic models that is tested and validated along different fluxomics and metabolomics is identified (Figure H.1.1.1g). This procedure is implemented in a metabolic model of *E. coli* core metabolism [118] that consists of 138 reactions, 93 metabolites and 60 substrate-level regulatory interactions [119, 120] by making use of the fluxomic data for wild-type and seven mutant strains [121]. The predicted fluxes by the model are within the uncertainty range of experimental flux data for 78% of the reactions (with measured fluxes) for both the reference (wild-type) and seven mutant strains. The predicted metabolite concentrations by the model are also within uncertainty ranges of metabolomic data for 68% of the metabolites. In addition, 80% of K_m and k_{cat} parameters are within one order of magnitude of literature available values.

H.2. k-OptForce: Integrating kinetics with FBA for strain design

H.2.1. Background

There has been rapid progress in recent years in the development of computational strain design protocols for system-wide identification of intervention strategies for the overproduction of biochemicals in microorganisms [122-129]. However, existing approaches relying solely on stoichiometry and rudimentary on-off regulation overlook the effects of metabolite concentrations and substrate-level enzyme regulation while identifying metabolic interventions. The k-OptForce protocol [130] was developed to extend the previously developed OptForce procedure [131] by bridging this gap

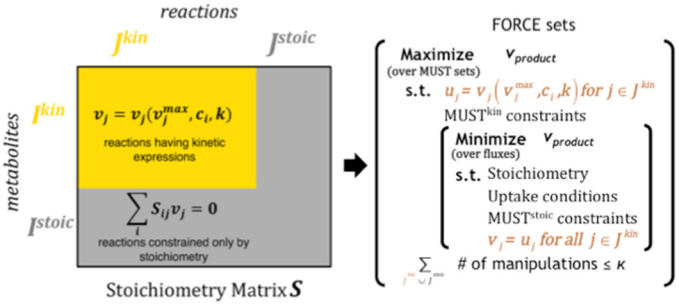


Figure H.2.1: Incorporation of kinetic information within the stoichiometry matrix in k-OptForce and bilevel formulation for identification of minimal set of interventions

between stoichiometry-only and kinetics-based descriptions of metabolism.

H.2.2. Results

k-OptForce protocol [130] seamlessly integrates the mechanistic detail afforded by kinetic models within a constraint-based optimization framework tractable for genome-scale models. Instead of relying on surrogate fitness functions such as biomass maximization or worst-case simulation for predicting flux re-directions, k-OptForce uses kinetic rate expressions to (re)apportion fluxes in the metabolic network. Using mechanistic models available in literature the allowable phenotype of both the reference and the engineered strain are characterized to be consistent with the allowable kinetic space. Subsequently, alternative genetic intervention strategies consistent with the restrictions imposed by maximum enzyme activity and kinetic regulations, as well as with the worst-case scenario of production of the desired chemical are identified using a bilevel optimization framework (Figure H.2.1).

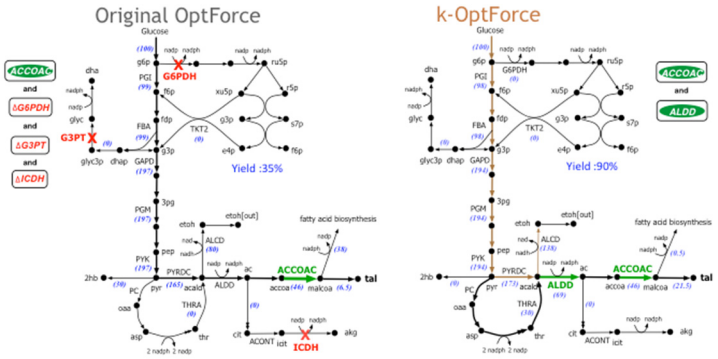


Figure H.2.2: Comparison of intervention strategies predicted by regular OptForce and k-OptForce for overproduction of TAL in *S. cerevisiae*

Application of the k-OptForce for the microbial overproduction of TAL in *S. cerevisiae* revealed the impact of additional kinetic constraints in alleviating a severe worst-case simulation of regular OptForce, resulting in a higher prediction of TAL yield (90% vs 35% of theoretical maximum) from fewer interventions (2 vs 4) as compared to regular OptForce predictions (Figure H.2.2). In general, both procedures suggest strategies that increase the availability of precursors accoa and malonyl-CoA (malcoa), up-regulating glycolysis, down-regulating Pentose Phosphate pathway, and reducing nadph production. However, while regular OptForce suggests a number of knockouts to prevent leaking flux away from acetyl-coA carboxylase, k-OptForce identifies that the kinetic expressions work in concert with the overproduction goal (given the imposed concentration ranges) without the need for any direct enzymatic interventions. In addition, the incorporation of kinetic information pushes metabolic flux in the direction that is needed for overproduction and away from the "worst-case" behavior, resulting in higher predicted TAL yield.

H.3. Integration of computational strain design and synthetic biology techniques for metabolic engineering applications

H.3.1. Background

OptForce procedure [131] has been employed to identify metabolic intervention strategies to increase Neurosporene production in *E. coli*. Neurosporene is produced through 1-deoxy-D-xylulose 5-phosphate pathway (DXP pathway). The DXP pathway is also used for the biosynthesis of isoprenoid, a class of organic compounds that are potentially useful as biofuels, pharmaceuticals, nutraceuticals, flavors and cosmetics products [132-134]. The DXP pathway requires the cofactor NADPH [135]. Here, we proposed a synthetic pathway for efficient NADPH regeneration and developed a systematic approach to rationally control NADPH regeneration from the synthetic pathway.

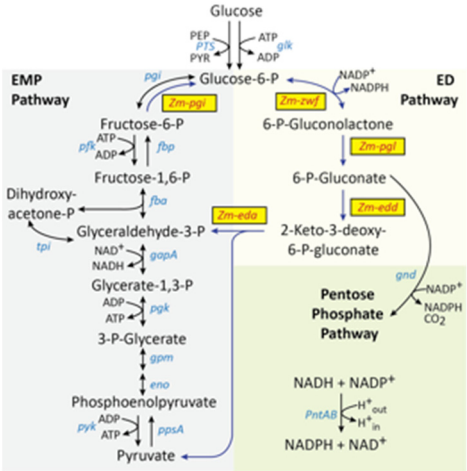


Figure H.3.1: Glucose metabolism in *E. coli*. The genes selected for overexpression are highlighted in yellow boxes.

H.3.2. Results

Five enzymes from *Zymomonas mobilis* Entner-Doudoroff (ED) pathway were selected for overexpression (See Figure H.3.1). Amino acid sequences of all five enzymes are back-translated using Operon Calculator

developed in the previous year. The resulting plasmid, R6K-LacI-ED-tetAR, containing all five enzymes is assembled using Gibson Assembly method [136]. ED-tetAR cassette is amplified from the plasmid and integrated into the genome of EcNR2 (*E. coli* MG1655) strain [137] to produce ED 1.0 strain. We then designed ribosome binding site (RBS) libraries for each enzyme using RBS Library Calculator (https://salis.psu.edu/software). Each library contains 16 different RBS sequences that span a large range of translation initiation rates. 40 cycles of multiplex automated genome engineering (MAGE) [137] was performed with oligonucleotides mixture containing all these RBS libraries. We then transformed the resulting pool of combinatorial variants with pMG3-mBFP plasmid harboring a NADPH-dependent metagenomic blue fluorescent protein[138]. Wild-type EcNR2 strain and ED 1.0 strain, both transformed with pMG3-mBFP, were used as reference strains. Screening was performed by characterizing specific blue fluorescence production rate of the variants using microplate reader. In each round of screening, the variants were found to span a large range specific blue fluorescence (See Figure H.3.2). The wild-type EcNR2 strain harboring the mBFP plasmid exhibited a relatively lower specific blue fluorescent when compared with ED 1.0 and other ED variants. The variants within the highest range and lowest range of specific blue fluorescence were then sent for sequencing.

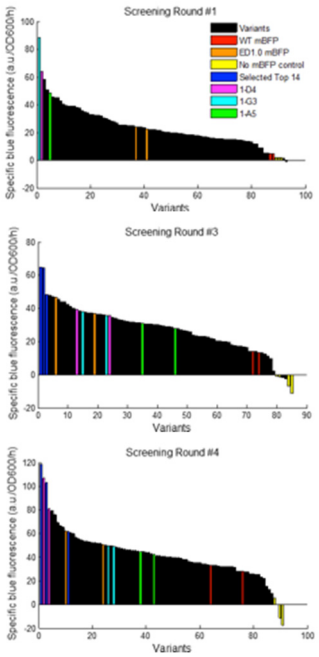


Figure H.3.2: Rank-ordered specific blue fluorescence in three different rounds of screening.

ED-tetAR cassette from two of the variants, 1-D4 and 1-G3, were amplified and integrated into the chromosome of EcHW2 DXS-15 strain. The strains were then transformed with pBad-crtEBI plasmid. Expression of *crtEBI* operon allows the cells to produce Neurosporene. EcHW2 DXS-15 *crtEBI* strain without the synthetic ED operons was used as reference. Neurosporene titer of all the strains were characterized in M9 minimal media supplemented with 0.4% w/v of glucose. The control strain accumulated 2212.1 µg/g DCW of Neurosporene. Expression of 1-D4's ED pathway significantly improved the production titer up to 3802.5 µg/g DCW (71.9% higher than reference strain). Another ED pathway variant 1-G3 produced 53.2% more Neurosporene than the reference strain (3389.0 µg/g DCW of Neurosporene). The DXP pathway required glyceraldehyde-3-phosphate and pyruvate as the pre-cursor molecules and also NADPH as the source of reducing equivalents. Synthetic ED pathway is able to supply both the pre-cursors and NADPH. This result demonstrates that the synthetic ED pathway is a promising approach for enhancing production titer of isoprenoid.

We used kinetic model approaches for strain design by applying the k-OptForce [139] procedure for the recently published large-scale kinetic model of *E. coli* core metabolism [140]. The kinetic model includes 138 reactions, 93 metabolites, and 60 substrate-level regulatory interactions and accounts for glycolysis/gluconeogenesis, pentose phosphate (PP) pathway, TCA cycle, major pyruvate metabolism, anaplerotic reactions, glyoxylate shunt, Entner–Doudoroff (ED) pathway, and a number of reactions in other parts of the metabolism. The model was parameterized using the ensemble modeling (EM) formalism [141] by simultaneously satisfying normalized flux data per 100 mmol of glucose uptake (for approximately 25 reactions per mutant) for the wild-type and seven single gene deletion mutants, under aerobic condition [121]. The EM approach decomposes all reactions into elementary steps bypassing the need of detail kinetic expressions. First, an ensemble of kinetic models is generated by uniformly sampling reaction reversibilities and enzyme fractions following different time trajectories but all reaching the same steady-state flux values. Next, a Genetic Algorithm (GA) implementation is used to "swap" kinetic parameterizations between models in the ensemble so as to minimize the deviations from all set of mutant network fluxes. Models constructed using flux data for a single strain do not always perform well in predicting deletion strain metabolic phenotypes [142].

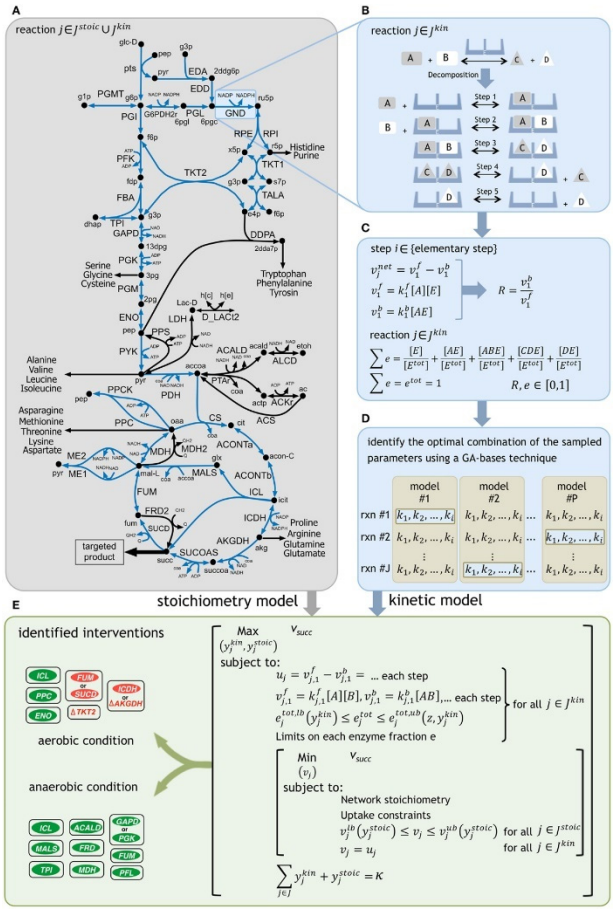


Figure H.3.3: A schematic representation of the framework. (A) The reactions with kinetic descriptions are shown in blue. (B) The reactions are first decomposed into their elementary steps. (C) Elementary kinetic parameters are expressed as a function of reaction reversibilities and enzyme fractions. Reaction reversibilities and enzyme fractions are sampled to construct an ensemble of models, for any given reaction. (D) A genetic algorithm (GA) implementation identifies the optimal combination of the sampled parameters by minimizing the deviation from experimentally measured flux data for multiple mutant strains. (E) The k-OptForce procedure identifies a minimal set of interventions that maximizes the yield of targeted product

The k-OptForce procedure [143] was used to identify the minimal interventions that maximize the yield of succinate production using a hybrid kinetic [140] and stoichiometric *i*AF1260 [118] description of *E. coli* metabolism. Succinate was chosen as the target bioproduct as there exists numerous experimental strain-engineering studies to compare the suggestions of k-OptForce procedure [144-146]. This study was carried out under both aerobic and anaerobic conditions to assess the fidelity of the kinetic model when used to make predictions for a different environmental condition (i.e., anaerobic) than the one parameterized for (i.e., aerobic). The goal was to quantify the reduction in prediction quality moving from aerobic to anaerobic under glucose minimal condition and suggest model modifications that remedy these shortcomings. k-OptForce recapitulated existing strategies while also pointing at promising but currently unexplored interventions. In addition, results under anaerobic condition indicate that the kinetic model needs to be re-parameterized with mutant flux information involving a reversed TCA cycle routing flux towards succinate. A number of regulatory modifications of the kinetic model are also found to be necessary to better reflect metabolic fluxes associated with anaerobic succinate production. These include activation of fermentation pathways and pyruvate formate lyase (PFL) by key regulatory proteins FNR (fumarate and nitrate reductase regulation) and ArcA (aerobic respiratory control).

H.4. Synthetic Pathway Design

H.4.1. Background

The introduction of the concept of 'total synthesis' by Wöhler [147] was one of the milestones in chemistry [148]. The possibility to create synthetic compounds from simple chemical building blocks has been a driving force of our modern world. Hence, it has been one of the ultimate goals in biology to achieve the same conceptual and synthetic level as reached in chemistry, ever since the principle of 'metabolic engineering' was developed in the early 1990s. Yet, cells are still far from being 'little chemical factories' and metabolic engineering has so-far been limited in its synthetic capabilities, relying mainly on the transplanting known pathways to a new host followed by optimization. Nevertheless, in silico pathway design has been realized for the biosynthesis of the value-added chemical 2,4-Dihydroxybutyric acid (DHB) [149], for producing therapeutics[150], and for designing synthetic carbon fixation pathways so as to increase the net carbon assimilation [151].

H.4.2. Results

Synthetic Entner-Doudoroff Pathway Design

Neurosporene in *E. coli* is produced through the 1-deoxy-D-xylulose 5-phosphate pathway (DXP pathway). The DXP pathway is also used for the biosynthesis of isoprenoid, a class of organic compounds that are potentially useful as biofuels, pharmaceuticals, nutraceuticals, flavors and cosmetics products [152-154]. The DXP pathway requires the cofactor NADPH [155]. We rationally engineered a synthetic version of the Entner–Doudoroff pathway from *Zymomonas mobilis* that increased the NADPH regeneration rate in *Escherichia coli* MG1655 by 25-fold. We combined systematic design rules, biophysical models, and computational optimization to design synthetic bacterial operons expressing the 5-enzyme pathway as a drop-in module, while eliminating undesired genetic elements for maximum expression control. Starting from the first version of the pathway, we carried out systematic optimization of the enzymes' expression levels to improve the pathway's activity, first employing a NADPH-dependent fluorescent protein reporter to measure NADPH regeneration rates, followed by measuring the ED pathway's effect on an NADPH-dependent terpenoid biosynthesis pathway. By combining MAGE genome mutagenesis with our RBS Library Calculator algorithm, we introduced targeted genome modifications to greatly vary the ED pathway's individual enzyme expression levels and to efficiently search its 5-dimensional expression space. We screened 624 ED pathway-genome variants for high NADPH regeneration rates, and then extensively characterized 22 re-integrated pathways by measuring in vivo NADPH regeneration rates and NADPH-dependent biosynthesis rates. The best variant exhibited 25-fold higher normalized mBFP levels when compared to wild-type strain. Combining the synthetic Entner–Doudoroff pathway with an optimized terpenoid pathway further increased the terpenoid titer by 97%

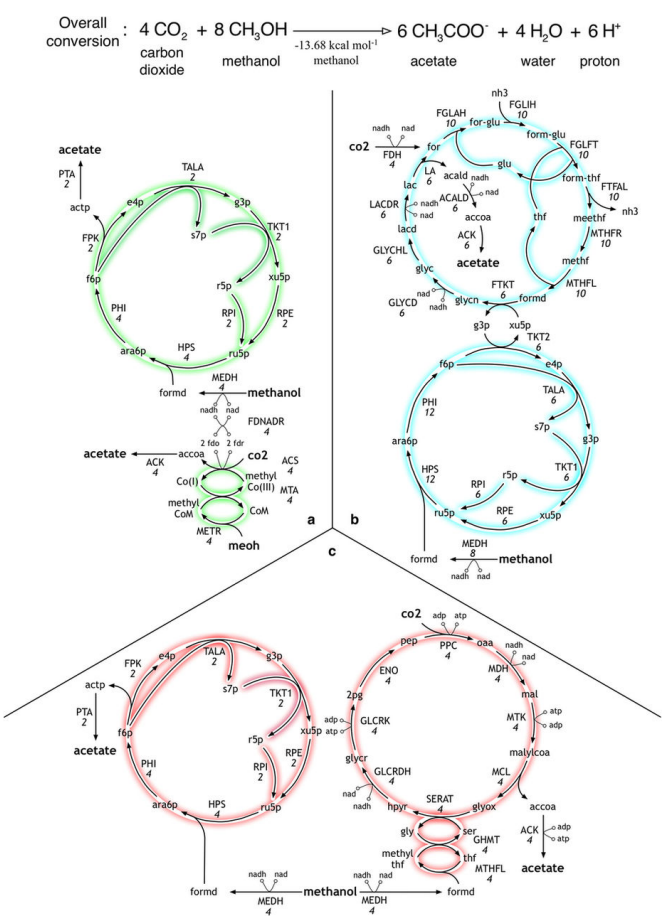


Figure H.4.1: Network designs for the co-utilization of methanol and carbon dioxide towards acetate.

Computational strain design and hence the ability to catalyze any tailor-made stoichiometry-balanced metabolic conversion with high specificity and control lies at the very heart of metabolic engineering. Existing computational procedures for the *de novo* pathway design rely on either optimization techniques or graph-search approaches. Linear Programming (LP) and Mixed Integer Linear Programming (MILP) approaches for pathway design, in general, extract a minimal stoichiometry-balanced sub-network that converts a source metabolite to a target chemical with high yield [156]. However, these procedures do not necessarily conform to a previously identified optimal conversion stoichiometry thereby missing the opportunity to optimally recycle intermediates to reach a maximum yield. Hence, we developed optStoic [157] which is a two-step algorithm that first identifies the optimal overall stoichiometry by exploring exhaustively co-reactant/co-product combinations, then adds intervening reactions from a database to link the chosen reactants to the selected products. We demonstrated the effectiveness of the two-step procedure for three separate case studies of increasing complexity. The first one exhaustively identifies networks that convert glucose to acetate while conserving all carbon atoms with no provision for any additional co-reactants or co-products in the spirit of the study by Bogorad *et al.* [158]. The second study explores the reverse problem of identifying new ways of forming carbon-carbon bonds from the combined use of methanol and CO_2 to stoichiometry-feasible C_{2+} products such as acetate. Finally, the third study identifies suitable co-reactant and co-product pairs to drive forward the thermodynamically unfavorable methane to acetate conversion. We compared the optimal pathways identified here with two existing pathway design tools (Chou *et al.* [159] and Bar-Even *et al.* [151]) for all three case studies. The results showed that the depth-first graph search algorithm for Chou *et al.* [159] identified only linear paths connecting the primary substrate to the primary product for all the case studies, while failing to identify any of the carbon-conserving cyclic networks. The Bar-Even *et al.* [151] approach successfully recapitulated the NOG cycle (Case Study 1), however, cofactor imbalances were introduced for the last two studies.

de novo Pathway Design

We have developed two novel procedures rePrime and novoStoic for the de novo pathway design. rePrime is a reaction rule based algorithm that encodes reaction centers as elementally balanced operators. These reaction rules capture moiety changes in the reaction centers by using the changes in the counts of prime numbers (i.e., canonical label for moieties) between substrates and products. Other than metabolites currently present in the database, our approach can be extended to novel metabolites as long as the structure can be codified as counts of moieties (i.e., molecular signature). rePrime allows for different moiety sizes. In the current implementation of rePrime, we trace moieties of up to a size of λ=3. In principle, one could expand the size of moieties traced or customize the size of the moiety traced based on the underlying reaction chemistry. By combining metabolite balance and moiety balance constraints, novoStoic

simultaneously integrates reaction rules with known reactions. It thus enables homing in first to the most desirable designs avoiding costly enumeration of alternatives that either include too many novel steps, are redox imbalanced, or fail to meet cost/yield requirements. The MILP based computational framework allows for straightforward control of cofactor regeneration, the number of novel reactions and the imposition of carbon yield or profit margin requirements.

novoStoic allows us to exploit enzyme plasticity by suggesting homologs to perform the hypothesized conversion when natural options are not available. In typical industrial bioprocesses, the number of novel reactions must be carefully controlled (or minimized) as each novel reaction implies an additional enzyme-substrate engineering challenge. In the event that the homolog is not promiscuous, protein engineering steps have to be recruited to enhance non-natural substrate binding (e.g., by tuning the binding pocket structure to accommodate the non-natural substrate [160]) and subsequently to increase catalytic rate [161]. For example, Cargill, Inc. engineered a multi-step 3-hydroxypropionic acid biosynthesis pathway, which employed a single non-natural enzyme (i.e. alanine 2,3-aminomutase), to bypass an ATP consuming step. The team had to engineer a homolog lysine 2,3-aminomutase to confer it the desirable activity and at the same time select a variant with the least negative effect on the host cell [162]. With the capability to blend known reactions and non-natural ones, novoStoic could invoke novel steps only when necessary.

A number of chemical manufacturing processes are increasingly exploiting the chemoselectivity and catalytic rate boost potential of enzymes [163] for the synthesis of pharmaceuticals and precursors. Studies have demonstrated that multi-enzyme cascades of non-natural enzymes can be implemented in both in vivo and in vitro fashion as well as in combination [164]. rePrime/novoStoic address the timely challenge of integrating recent advancements for the rapid identification of complete pathways for bio-based chemosynthesis and the elucidation of intermediates of ill-defined xenobiotic degradative pathways. The detailed degradation map can therefore assist in evaluating the toxicity and potential side effects of new drugs, and even enable the assessment of synergistic, antagonist or toxic drug interactions.

novoStoic sometimes predicts pathways where the rules invoked to fill in intermediate steps could map to multiple possible reactions. The degree of specificity of the reaction rules can be controlled by preferentially using moieties of size 3 or 2 and only size 1 if no solutions were recovered. Note that novoStoic does not allow for mixing of moieties of different size during the pathway design phase. As anticipated, larger moiety sizes generally yield novel steps “closer” to a known reaction and thus more likely to involve an existing (promiscuous) enzyme with some level of this activity. However, larger moiety sizes (distance of 2 or 3) severely restrict the number of possibilities for novel steps. Generally, we start the pathway design using moieties of distance 3 and then reduce to 2 or even 1 depending on the efficacy of the search so far. In addition, the requirement of elementally balanced reaction rules with proper cofactor utilization and stereo-chemical changes necessitate a high quality biochemical database as an input for rePrime/novoStoic. Incomplete or incorrect reaction annotation (e.g., molecular structure, stereochemistry, stoichiometry, cofactor and reaction mechanism) could significantly affect the quality of the rules identified and the reliability of a pathway. A number of automated algorithms [165] and procedures [166] have been developed to reduce annotation inconsistencies and unify discrepancies across different databases. However, expert curation is often necessary to include updated discoveries (e.g., fixing cofactor utilization of a stoichiometrically balanced reaction) as well as to evaluate and resolve contradicting information [167]. Furthermore, we generally treat all rules as reversible. Therefore, additional scrutiny may be needed to ensure that the reaction rule ultimately maps to a reaction that is thermodynamically feasible.

References

1. Kumar A, Suthers PF, Maranas CD: **MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases.** *BMC bioinformatics* 2012, **13**:6.

2. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Söhngen C, Stelzer M, Thiele J, Schomburg D: **BRENDA, the enzyme information system in 2011.** *Nucleic acids research* 2011, **39**:D670-676.
3. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic acids research* 1999, **27**:29-34.
4. Caspi R, Altman T, Dreher K, Fulcher Ca, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller La *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic acids research* 2012, **40**:D742-753.
5. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B *et al*: **Reactome knowledgebase of human biological pathways and processes.** *Nucleic acids research* 2009, **37**:D619-622.
6. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Molecular systems biology* 2007, **3**:121.
7. Kjeldsen KR, Nielsen J: **In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network.** *Biotechnology and bioengineering* 2009, **102**:583-597.
8. Durot M, Le Fèvre F, de Berardinis V, Kreimeyer A, Vallenet D, Combe C, Smidtas S, Salanoubat M, Weissenbach J, Schachter V: **Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data.** *BMC systems biology* 2008, **2**:85.
9. Henry CS, Zinner JF, Cohoon MP, Stevens RL: **iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.** *Genome biology* 2009, **10**:R69.
10. David H, Ozçelik IS, Hofmann G, Nielsen J: **Analysis of Aspergillus nidulans metabolism at the genome-scale.** *BMC genomics* 2008, **9**:163.
11. Cattell R: **Scalable SQL and NoSQL Data Stores**. 2010, **39**.
12. Chodorow K, Dirolf M: **MongoDB: the definitive guide**. 2010.
13. Cho A, Yun H, Park JH, Lee SY, Park S: **Prediction of novel synthetic pathways for the production of desired chemicals.** *BMC systems biology* 2010, **4**:35.
14. Blum T, Kohlbacher O: **MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization.** *Bioinformatics (Oxford, England)* 2008, **24**:2108-2109.
15. Ranganathan S, Maranas CD: **Microbial 1-butanol production: Identification of non-native production routes and in silico engineering interventions.** *Biotechnology journal* 2010, **5**:716-725.
16. Wu CH, Apweiler R, Bairoch A, Natale Da, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R *et al*: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic acids research* 2006, **34**:D187-191.
17. Stephanopoulos G: **Metabolic fluxes and metabolic engineering.** *Metabolic engineering* 1999, **1**:1-11.
18. Hellerstein MK: **In vivo measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research.** *Annual review of nutrition* 2003, **23**:379-402.
19. Wiechert W: **13C metabolic flux analysis.** *Metabolic engineering* 2001, **3**:195-206.
20. Crown SB, Antoniewicz MR: **Parallel labeling experiments and metabolic flux analysis: Past, present and future methodologies.** *Metabolic engineering* 2012, **16C**:21-32.
21. Ahn WS, Antoniewicz MR: **Parallel labeling experiments with [1,2-(13)C]glucose and [U-(13)C]glutamine provide new insights into CHO cell metabolism.** *Metabolic engineering* 2013, **15**:34-47.

22. Hiller K, Metallo CM: **Profiling metabolic networks to study cancer metabolism.** *Current opinion in biotechnology* 2013, **24**:60-68.
23. Boghigian Ba, Seth G, Kiss R, Pfeifer Ba: **Metabolic flux analysis and pharmaceutical production.** *Metabolic engineering* 2010, **12**:81-95.
24. Young JD, Walther JL, Antoniewicz MR, Yoo H: **An Elementary Metabolite Unit (EMU) Based Method of Isotopically Nonstationary Flux Analysis**. 2008, **99**:686-699.
25. Weitzel M, Nöh K, Dalman T, Niedenführ S, Stute B, Wiechert W: **13CFLUX2--high-performance software suite for (13)C-metabolic flux analysis.** *Bioinformatics (Oxford, England)* 2013, **29**:143-145.
26. Srour O, Young JD, Eldar YC: **Fluxomers: a new approach for 13C metabolic flux analysis.** *BMC systems biology* 2011, **5**:129.
27. Chang Y, Suthers PF, Maranas CD: **Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments.** *Biotechnology and bioengineering* 2008, **100**:1039-1049.
28. Antoniewicz MR, Kelleher JK, Stephanopoulos G: **Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements.** *Metabolic engineering* 2006, **8**:324-337.
29. Antoniewicz MR: **13C metabolic flux analysis: optimal design of isotopic labeling experiments**. *Current Opinion in Biotechnology* 2013:1-6.
30. Quek L-E, Wittmann C, Nielsen LK, Krömer JO: **OpenFLUX: efficient modelling software for 13C-based metabolic flux analysis.** *Microbial cell factories* 2009, **8**:25.
31. Blum T, Kohlbacher O: **Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks.** *Journal of computational biology : a journal of computational molecular cell biology* 2008, **15**:565-576.
32. Heath AP, Bennett GN, Kavraki LE: **Finding metabolic pathways using atom tracking.** *Bioinformatics (Oxford, England)* 2010, **26**:1548-1555.
33. Bahiense L, Manić G, Piva B, de Souza CC: **The maximum common edge subgraph problem: A polyhedral investigation**. *Discrete Applied Mathematics* 2012, **160**:2523-2541.
34. Schellewald C: **A Convex Relaxation Bound for Subgraph Isomorphism**. *International Journal of Combinatorics* 2012, **2012**:1-18.
35. Manić G, Bahiense L, de Souza C: **A branch&cut algorithm for the maximum common edge subgraph problem**. *Electronic Notes in Discrete Mathematics* 2009, **35**:47-52.
36. Heinonen M, Lappalainen S, Mielikäinen T, Rousu J: **Computing atom mappings for biochemical reactions without subgraph isomorphism.** *Journal of computational biology : a journal of computational molecular cell biology* 2011, **18**:43-58.
37. Mann M, Ekker H, Stadler PF, Flamm C: **Atom Mapping with Constraint Programming**.
38. Latendresse M, Malerich JP, Travers M, Karp PD, International SRI, Ave R, Park M, States U: **Accurate Atom-Mapping Computation for Biochemical Reactions**. 2012.
39. Schmidt K, Nielsen J, Villadsen J: **Quantitative analysis of metabolic fluxes in Escherichia coli, using two-dimensional NMR spectroscopy and complete isotopomer models.** *Journal of biotechnology* 1999, **71**:175-189.
40. Zupke C, Stephanopoulos G: **Modeling of Isotope Distributions and Intracellular Fluxes in Metabolic Networks Using Atom Mapping Matrices**. 1994:489-498.
41. Leveau V, Lorgeou J, Prioul J-L: **Maize in the world economy: a challenge for scientific research - how to produce more cheaper!** In: *Advances in Maize.* Edited by Prioul JLT, C.; Molnar, T., vol. 3. UK: Society for Experimental Biology; 2011.
42. International Grains Council: **International Grains Council: Report for Fiscal Year 2011/12**. In.: International Grains Council; 2013.
43. Kennedy RA: **Photorespiration in c(3) and c(4) plant tissue cultures: significance of kranz anatomy to low photorespiration in c(4) plants**. *Plant Physiol* 1976, **58**(4):573-575.

44. Brown RH: **A Difference in N Use Efficiency in C3 and C4 Plants and its Implications in Adaptation and Evolution1**. *Crop Sci* 1978, **18**(1):93-98.
45. Zelitch I: **Pathways of Carbon Fixation in Green Plants**. *Annual Review of Biochemistry* 1975, **44**(1):123-145.
46. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
47. Monaco MK, Sen TZ, Dharmawardhana PD, Ren L, Schaeffer M, Naithani S, Amarasinghe V, Thomason J, Harper L, Gardiner J *et al*: **Maize Metabolic Network Construction and Transcriptome Analysis**. *Plant Gen* 2013, **6**(1):-.
48. Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Sen TZ, Lawrence CJ: **MaizeGDB: curation and outreach go hand-in-hand**. *Database : the journal of biological databases and curation* 2011, **2011**:bar022.
49. Schreiber F, Colmsee C, Czauderna T, Grafahrend-Belau E, Hartmann A, Junker A, Junker BH, Klapperstuck M, Scholz U, Weise S: **MetaCrop 2.0: managing and exploring information about crop plant metabolism**. *Nucleic Acids Res* 2012, **40**(Database issue):D1173-1177.
50. Saha R, Chowdhury A, Maranas CD: **Recent advances in the reconstruction of metabolic models and integration of omics data**. *Current opinion in biotechnology* 2014, **29**(0):39-45.
51. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
52. Bennett MD, Leitch IJ, Price HJ, Johnston JS: **Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb**. *Ann Bot* 2003, **91**(5):547-557.
53. Saha R, Suthers PF, Maranas CD: **Zea mays iRS1563: A Comprehensive Genome-Scale Metabolic Reconstruction of Maize Metabolism**. *Plos One* 2011, **6**(7).
54. Majeran W, Cai Y, Sun Q, van Wijk KJ: **Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics**. *The Plant cell* 2005, **17**(11):3111-3140.
55. Friso G, Majeran W, Huang MS, Sun Q, van Wijk KJ: **Reconstruction of Metabolic Pathways, Protein Expression, and Homeostasis Machineries across Maize Bundle Sheath and Mesophyll Chloroplasts: Large-Scale Quantitative Proteomics Using the First Maize Genome Assembly**. *Plant Physiol* 2010, **152**(3):1219-1250.
56. Salimi F, Zhuang K, Mahadevan R: **Genome-scale metabolic modeling of a clostridial co-culture for consolidated bioprocessing**. *Biotechnol J* 2010, **5**(7):726-738.
57. Chang YM, Liu WY, Shih ACC, Shen MN, Lu CH, Lu MYJ, Yang HW, Wang TY, Chen SCC, Chen SM *et al*: **Characterizing Regulatory and Functional Differentiation between Maize Mesophyll and Bundle Sheath Cells by Transcriptomic Analysis**. *Plant Physiol* 2012, **160**(1):165-177.
58. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Res* 2012, **40**(Database issue):D109-114.
59. Nelson DL, Cox MM: **Oxidative Phosphorylation and Photophosphorylation**. In: *Lehninger Principles of Biochemsitry.* Fifth edn. New York: W.H.Freeman & Co.; 2009: 707-772.
60. Taiz LaZ, E.: **Plant Physiology**, Fifth edn. Sunderland, Massachusetts: Sinauer Associates Inc.; 2010.
61. Bachlava E, Dewey R, Burton J, Cardinal AJ: **Mapping candidate genes for oleate biosynthesis and their association with unsaturated fatty acid seed content in soybean**. *Mol Breeding* 2009, **23**(2):337-347.

62. Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, Bates PD, Baud S, Bird D, Debono A, Durrett TP *et al*: **Acyl-lipid metabolism**. *The Arabidopsis book / American Society of Plant Biologists* 2010, **8**:e0133.
63. Mekhedov S, de Ilarduya OM, Ohlrogge J: **Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis**. *Plant Physiol* 2000, **122**(2):389-402.
64. Murata N: **Molecular-Species Composition of Phosphatidylglycerols from Chilling-Sensitive and Chilling-Resistant Plants**. *Plant and Cell Physiology* 1983, **24**(1):81-86.
65. Moore TS: **Phospholipid Biosynthesis**. *Annu Rev Plant Phys* 1982, **33**:235-259.
66. Rolland N, Curien G, Finazzi G, Kuntz M, Marechal E, Matringe M, Ravanel S, Seigneurin-Berny D: **The biosynthetic capacities of the plastids and integration between cytoplasmic and chloroplast processes**. *Annual review of genetics* 2012, **46**:233-264.
67. Murata N, Tasaka Y: **Glycerol-3-phosphate acyltransferase in plants**. *Bba-Lipid Lipid Met* 1997, **1348**(1-2):10-16.
68. Martin A, Lee J, Kichey T, Gerentes D, Zivy M, Tatout C, Dubois F, Balliau T, Valot B, Davanture M *et al*: **Two cytosolic glutamine synthetase isoforms of maize are specifically involved in the control of grain production**. *The Plant cell* 2006, **18**(11):3252-3274.
69. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R: **IntEnz, the integrated relational enzyme database.** *Nucleic acids research* 2004, **32**:D434-437.
70. Bairoch a: **The ENZYME database in 2000.** *Nucleic acids research* 2000, **28**:304-305.
71. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S *et al*: **Reactome: a knowledge base of biologic pathways and processes.** *Genome biology* 2007, **8**:R39.
72. Stobbe MD, Houten SM, Jansen Ga, van Kampen AHC, Moerland PD: **Critical assessment of human metabolic pathway databases: a stepping stone for future integration.** *BMC systems biology* 2011, **5**:165.
73. Bornstein BJ, Keating SM, Jouraku A, Hucka M: **LibSBML: an API library for SBML.** *Bioinformatics (Oxford, England)* 2008, **24**:880-881.
74. Hucka M, Finney a, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin aP, Bornstein BJ, Bray D, Cornish-Bowden a *et al*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics* 2003, **19**:524-531.
75. Strömbäck L, Lambrix P: **Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX.** *Bioinformatics (Oxford, England)* 2005, **21**:4401-4407.
76. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for Escherichia coli.** *Nucleic acids research* 2005, **33**:D334-337.
77. Radrich K, Tsuruoka Y, Dobson P, Gevorgyan A, Swainston N, Baart G, Schwartz J-M: **Integration of metabolic databases for the reconstruction of genome-scale metabolic networks.** *BMC systems biology* 2010, **4**:114.
78. Pitkänen E, Akerlund A, Rantanen A, Jouhten P, Ukkonen E: **ReMatch: a web-based tool to construct, store and share stoichiometric metabolic models with carbon maps for metabolic flux analysis.** *Journal of integrative bioinformatics* 2008, **5**:1-13.
79. Gonzalez O, Gronau S, Falb M, Pfeiffer F, Mendoza E, Zimmer R, Oesterhelt D: **Reconstruction, modeling & analysis of Halobacterium salinarum R-1 metabolism.** *Molecular bioSystems* 2008, **4**:148-159.
80. Henry CS, DeJongh M, Best Aa, Frybarger PM, Linsay B, Stevens RL: **High-throughput generation, optimization and analysis of genome-scale metabolic models.** *Nature biotechnology* 2010, **28**:977-982.
81. Chowdhury R, Chowdhury A, Maranas C: **Using Gene Essentiality and Synthetic Lethality Information to Correct Yeast and CHO Cell Genome-Scale Models**. *Metabolites* 2015, **5**(4):536.

82. Selvarasu S, Karimi IA, Ghim GH, Lee DY: **Genome-scale modeling and in silico analysis of mouse cell metabolic network**. *Mol Biosyst* 2010, **6**(1):152-161.
83. Selvarasu S, Ho YS, Chong WP, Wong NS, Yusufi FN, Lee YY, Yap MG, Lee DY: **Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture**. *Biotechnol Bioeng* 2012, **109**(6):1415-1429.
84. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A: **OMIM.org: Online Mendelian Inheritance in Man (OMIM (R)), an online catalog of human genes and genetic disorders**. *Nucleic Acids Research* 2015, **43**(D1):D789-D798.
85. van Pel DM, Stirling PC, Minaker SW, Sipahimalani P, Hieter P: **Saccharomyces cerevisiae Genetics Predicts Candidate Therapeutic Genetic Interactions at the Mammalian Replication Fork**. *G3-Genes Genom Genet* 2013, **3**(2):273-282.
86. Keng T: **Hap1 and Rox1 Form a Regulatory Pathway in the Repression of Hem13 Transcription in Saccharomyces-Cerevisiae**. *Mol Cell Biol* 1992, **12**(6):2616-2623.
87. Keogh MC, Podolny V, Buratowski S: **Bur1 kinase is required for efficient transcription elongation by RNA polymerase II**. *Mol Cell Biol* 2003, **23**(19):7005-7018.
88. Chu XL, Qin XH, Xu HS, Li L, Wang Z, Li FZ, Xie XQ, Zhou H, Shen YQ, Long JF: **Structural insights into Paf1 complex assembly and histone binding**. *Nucleic Acids Research* 2013, **41**(22):10619-10629.
89. Saha R, Verseput AT, Berla BM, Mueller TJ, Pakrasi HB, Maranas CD: **Reconstruction and comparison of the metabolic potential of cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803**. *PLoS One* 2012, **7**(10):e48285.
90. Young JD, Shastri AA, Stephanopoulos G, Morgan JA: **Mapping photoautotrophic metabolism with isotopically nonstationary (13)C flux analysis**. *Metab Eng* 2011, **13**(6):656-665.
91. Gopalakrishnan S, Maranas CD: **13C metabolic flux analysis at a genome-scale**. *Metab Eng* 2015, **32**:12-22.
92. Antoniewicz MR, Kelleher JK, Stephanopoulos G: **Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements**. *Metab Eng* 2006, **8**(4):324-337.
93. Sandberg TE, Long CP, Gonzalez JE, Feist AM, Antoniewicz MR, Palsson BO: **Evolution of E. coli on [U-13C]Glucose Reveals a Negligible Isotopic Influence on Metabolism and Physiology**. *PLoS One* 2016, **11**(3):e0151130.
94. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**:121.
95. Feist AM, Palsson BO: **The biomass objective function**. *Current Opinion in Microbiology* 2010, **13**(3):344-349.
96. Chan SHJ, Cai JY, Wang L, Simons-Senftle MN, Maranas CD: **Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models**. *Bioinformatics* 2017, **33**(22):3603-3609.
97. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR *et al*: **Community genomics among stratified microbial assemblages in the ocean's interior**. *Science* 2006, **311**(5760):496-503.
98. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS *et al*: **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207-214.
99. Karlsson FH, Nookaew I, Petranovic D, Nielsen J: **Prospects for systems biology and modeling of the gut microbiome**. *Trends Biotechnol* 2011, **29**(6):251-258.
100. Xavier JB: **Social interaction in synthetic and natural microbial communities**. *Molecular Systems Biology* 2011, **7**.
101. Fuhrman JA: **Microbial community structure and its functional implications**. *Nature* 2009, **459**(7244):193-199.

102. Wintermute EH, Silver PA: **Dynamics in the mixed microbial concourse**. *Gene Dev* 2010, **24**(23):2603-2614.
103. Wintermute EH, Silver PA: **Emergent cooperation in microbial metabolism**. *Molecular Systems Biology* 2010, **6**.
104. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophys J* 2002, **83**(3):1331-1340.
105. Chan SHJ, Simons MN, Maranas CD: **SteadyCom: Predicting microbial abundances while ensuring community stability**. *Plos Computational Biology* 2017, **13**(5).
106. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Molecular Systems Biology* 2007, **3**.
107. Zomorrodi AR, Islam MM, Maranas CD: **d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities**. *ACS synthetic biology* 2014, **3**(4):247-257.
108. Zomorrodi AR, Maranas CD: **OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities**. *PLoS computational biology* 2012, **8**(2):e1002363.
109. Zhuang K, Izallalen M, Mouser P, Richter H, Risso C, Mahadevan R, Lovley DR: **Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments**. *ISME J* 2011, **5**(2):305-316.
110. Pinchuk GE, Hill EA, Geydebrekht OV, De Ingeniis J, Zhang X, Osterman A, Scott JH, Reed SB, Romine MF, Konopka AE *et al*: **Constraint-based model of Shewanella oneidensis MR-1 metabolism: a tool for data analysis and hypothesis generation**. *PLoS computational biology* 2010, **6**(6):e1000822.
111. Pinchuk GE, Geydebrekht OV, Hill EA, Reed JL, Konopka AE, Beliaev AS, Fredrickson JK: **Pyruvate and lactate metabolism by Shewanella oneidensis MR-1 under fermentation, oxygen limitation, and fumarate respiration conditions**. *Applied and environmental microbiology* 2011, **77**(23):8234-8240.
112. Tang YJ, Meadows AL, Kirby J, Keasling JD: **Anaerobic central metabolic pathways in Shewanella oneidensis MR-1 reinterpreted in the light of isotopic metabolite labeling**. *Journal of bacteriology* 2007, **189**(3):894-901.
113. Wall JD, Krumholz LR: **Uranium reduction**. *Annual review of microbiology* 2006, **60**:149-166.
114. Wintermute EH, Silver PA: **Emergent cooperation in microbial metabolism**. *Molecular systems biology* 2010, **6**:407.
115. Almquist J, Cvijovic M, Hatzimanikatis V, Nielsen J: **Kinetic models in industrial biotechnology – Improving cell factory performance**. *Metabolic engineering* 2014.
116. Chou IC, Voit EO: **Recent developments in parameter estimation and structure identification of biochemical and genomic systems**. *Mathematical biosciences* 2009, **219**(2):57-83.
117. Tan Y, Liao JC: **Metabolic ensemble modeling for strain engineers**. *Biotechnol J* 2012, **7**(3):343-353.
118. Khodayari A, Zomorrodi AR, Liao JC, Maranas CD: **A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data**. submitted.
119. Schomburg I, Chang A, Placzek S, Sohngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A *et al*: **BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA**. *Nucleic Acids Res* 2013, **41**(Database issue):D764-772.
120. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A: **The EcoCyc and MetaCyc databases**. *Nucleic acids research* 2000, **28**(1):56-59.
121. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A *et al*: **Multiple high-throughput analyses monitor the response of E. coli to perturbations**. *Science* 2007, **316**(5824):593-597.

122. Burgard AP, Pharkya P, Maranas CD: **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization**. *Biotechnology and bioengineering* 2003, **84**(6):647-657.
123. Cotten C, Reed JL: **Constraint-based strain design using continuous modifications (CosMos) of flux bounds finds new strategies for metabolic engineering**. *Biotechnology journal* 2013, **8**(5):595-604.
124. Kim J, Reed JL: **OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains**. *BMC systems biology* 2010, **4**:53.
125. Kim J, Reed JL, Maravelias CT: **Large-scale bi-level strain design approaches and mixed-integer programming solution techniques**. *PloS one* 2011, **6**(9):e24162.
126. Patil KR, Rocha I, Forster J, Nielsen J: **Evolutionary programming as a platform for in silico metabolic engineering**. *BMC bioinformatics* 2005, **6**:308.
127. Pharkya P, Burgard AP, Maranas CD: **OptStrain: a computational framework for redesign of microbial production systems**. *Genome research* 2004, **14**(11):2367-2376.
128. Pharkya P, Maranas CD: **An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems**. *Metab Eng* 2006, **8**(1):1-13.
129. Yang L, Cluett WR, Mahadevan R: **EMILiO: a fast algorithm for genome-scale strain design**. *Metabolic engineering* 2011, **13**(3):272-281.
130. Chowdhury A, Zomorrodi AR, Maranas CD: **k-OptForce: Integrating kinetics with flux balance analysis for strain design**. *PLoS computational biology* accepted.
131. Ranganathan S, Suthers PF, Maranas CD: **OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions**. *PLoS Comput Biol* 2010, **6**(4):e1000744.
132. Ajikumar PK, Xiao WH, Tyo KE, Wang Y, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G: **Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli**. *Science* 2010, **330**(6000):70-74.
133. Bro C, Regenberg B, Forster J, Nielsen J: **In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production**. *Metabolic engineering* 2006, **8**(2):102-111.
134. Martin VJ, Pitera DJ, Withers ST, Newman JD, Keasling JD: **Engineering a mevalonate pathway in Escherichia coli for production of terpenoids**. *Nature biotechnology* 2003, **21**(7):796-802.
135. Liu H, Sun Y, Ramos KR, Nisola GM, Valdehuesa KN, Lee WK, Park SJ, Chung WJ: **Combination of entner-doudoroff pathway with MEP increases isoprene production in engineered Escherichia coli**. *PloS one* 2013, **8**(12):e83290.
136. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, 3rd, Smith HO: **Enzymatic assembly of DNA molecules up to several hundred kilobases**. *Nature methods* 2009, **6**(5):343-345.
137. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM: **Programming cells by multiplex genome engineering and accelerated evolution**. *Nature* 2009, **460**(7257):894-898.
138. Hwang CS, Choi ES, Han SS, Kim GJ: **Screening of a highly soluble and oxygen-independent blue fluorescent protein from metagenome**. *Biochemical and biophysical research communications* 2012, **419**(4):676-681.
139. Khodayari A, Chowdhury A, Maranas CD: **Succinate Overproduction: A Case Study of Computational Strain Design Using a Comprehensive Escherichia coli Kinetic Model**. *Front Bioeng Biotechnol* 2014, **2**:76.
140. Khodayari A, Zomorrodi AR, Liao JC, Maranas CD: **A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data**. *Metab Eng* 2014, **25**:50-62.
141. Tran LM, Rizk ML, Liao JC: **Ensemble modeling of metabolic networks**. *Biophysical journal* 2008, **95**(12):5606-5617.
142. Jouhten P: **Metabolic modelling in the development of cell factories by synthetic biology**. *Comput Struct Biotechnol J* 2012, **3**:e201210009.

143. Chowdhury A, Zomorrodi AR, Maranas CD: **k-OptForce: integrating kinetics with flux balance analysis for strain design**. *PLoS Comput Biol* 2014, **10**(2):e1003487.
144. Lee SJ, Lee DY, Kim TY, Kim BH, Lee J, Lee SY: **Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation**. *Appl Environ Microbiol* 2005, **71**(12):7880-7887.
145. Cao Y, Cao Y, Lin X: **Metabolically engineered Escherichia coli for biotechnological production of four-carbon 1,4-dicarboxylic acids**. *J Ind Microbiol Biotechnol* 2011, **38**(6):649-656.
146. Tan Y, Rivera JG, Contador CA, Asenjo JA, Liao JC: **Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux**. *Metab Eng* 2011, **13**(1):60-75.
147. Wöhler F: **Ueber künstliche bildung des harnstoffs**. *Annalen der Physik* 1828, **88**(2):253-256.
148. Nicolaou KC, Vourloumis D, Winssinger N, Baran PS: **The art and science of total synthesis at the dawn of the twenty-first century**. *Angew Chem Int Edit* 2000, **39**(1):44-122.
149. Walther T, Topham CM, Irague R, Auriol C, Baylac A, Cordier H, Dressaire C, Lozano-Huguet L, Tarrat N, Martineau N *et al*: **Construction of a synthetic metabolic pathway for biosynthesis of the non-natural methionine precursor 2,4-dihydroxybutyric acid**. *Nat Commun* 2017, **8**.
150. Carbonell P, Planson AG, Fichera D, Faulon JL: **A retrosynthetic biology approach to metabolic pathway design for therapeutic production**. *Bmc Systems Biology* 2011, **5**.
151. Bar-Even A, Noor E, Lewis NE, Milo R: **Design and analysis of synthetic carbon fixation pathways**. *P Natl Acad Sci USA* 2010, **107**(19):8889-8894.
152. Martin VJJ, Pitera DJ, Withers ST, Newman JD, Keasling JD: **Engineering a mevalonate pathway in Escherichia coli for production of terpenoids**. *Nat Biotechnol* 2003, **21**(7):796-802.
153. Bro C, Regenberg B, Forster J, Nielsen J: **In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production**. *Metab Eng* 2006, **8**(2):102-111.
154. Ajikumar PK, Xiao WH, Tyo KEJ, Wang Y, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G: **Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in Escherichia coli**. *Science* 2010, **330**(6000):70-74.
155. Liu H, Sun Y, Ramos KRM, Nisola GM, Valdehuesa KNG, Lee WK, Park SJ, Chung WJ: **Combination of Entner-Doudoroff Pathway with MEP Increases Isoprene Production in Engineered Escherichia coli**. *Plos One* 2013, **8**(12).
156. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C: **Elementary flux modes in a nutshell: Properties, calculation and applications**. *Biotechnol J* 2013, **8**(9):1009-U1061.
157. Chowdhury A, Maranas CD: **Designing overall stoichiometric conversions and intervening metabolic reactions**. *Sci Rep-Uk* 2015, **5**.
158. Bogorad IW, Lin TS, Liao JC: **Synthetic non-oxidative glycolysis enables complete carbon conservation**. *Nature* 2013, **502**(7473):693-+.
159. Chou CH, Chang WC, Chiu CM, Huang CC, Huang HD: **FMM: a web server for metabolic pathway reconstruction and comparative analysis**. *Nucleic Acids Research* 2009, **37**:W129-W134.
160. Zhang KC, Sawaya MR, Eisenberg DS, Liao JC: **Expanding metabolism for biosynthesis of nonnatural alcohols**. *P Natl Acad Sci USA* 2008, **105**(52):20653-20658.
161. Currin A, Swainston N, Day PJ, Kell DB: **Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently**. *Chem Soc Rev* 2015, **44**(5):1172-1239.
162. Brazeau BJ, Gort SJ, Jessen HJ, Andrew AJ, Liao HH: **Enzymatic activation of lysine 2,3-aminomutase from Porphyromonas gingivalis**. *Appl Environ Microbiol* 2006, **72**(9):6402-6404.
163. Renata H, Wang ZJ, Arnold FH: **Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution**. *Angew Chem Int Ed Engl* 2015, **54**(11):3351-3367.
164. France SP, Hepworth LJ, Turner NJ, Flitsch SL: **Constructing Biocatalytic Cascades: In Vitro and in Vivo Approaches to de Novo Multi-Enzyme Pathways**. *Acs Catal* 2017, **7**(1):710-724.

165. Moretti S, Martin O, Tran TV, Bridge A, Morgat A, Pagni M: **MetaNetX/MNXref - reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks**. *Nucleic Acids Research* 2016, **44**(D1):D523-D526.
166. Lang MR, Stelzer M, Schomburg D: **BKM-react, an integrated biochemical reaction database**. *Bmc Biochem* 2011, **12**.
167. Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu ZY, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B *et al*: **On expert curation and scalability: UniProtKB/Swiss-Prot as a case study**. *Bioinformatics* 2017, **33**(21):3454-3460.