Computational Methods to Advance Directed Evolution of Enzymes and Metabolomics Data Analysis

A dissertation submitted by

Neda Hassanpour, B. Sc., M.Sc.

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

May 2018

ADVISOR: Prof. Soha Hassoun

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Professor Soha Hassoun for being my true mentor. Her patience, encouragement, support and immense knowledge provided me a fertile environment to grow and learn. A special thanks to my collaborators Professors Nikhil Nair, Kyongbum Lee, and Li-ping Liu for their generosity in offering their help and knowledge to make writing this thesis possible. I would like to thank all members of my defense committee, especially Professor Anselm Blumer, for their valuable and constructive feedback to improve my thesis. I am immensely grateful to Nicholas Alden for his valuable knowledge and assistance with the experimental section of my thesis. I thank all my friends and members of the Department of Computer Science, specially Sara Amin and Ehsan Ullah, for always being there for me and encouraging me at the time of obstacles.

I would like to thank my family, whom without their support I could not be here. I thank my brothers, Navid and Saeed, for their support and being my inspiration through all these years. At last, my deepest appreciation goes to my

lovely parents, who always made me to believe in myself for pushing the boundaries. Thank you for standing by me all the way long, and for all the sacrifices you made to this day.

Abstract

The engineering of living cells promises to advance many applications including synthetic biology and personalized medicine. Experimental efforts, however, can be costly and time-consuming, requiring large efforts to interpret collected data and many iterative design-and-test cycles to achieve desired results. Computational efforts that harness the continuing growth of computing power and catalogued biological data can advance biological system design by interpreting measurements, efficiently exploring the design space and expediting biological discoveries.

This thesis advances state-of-the-art in the engineering and analysis of cellular metabolism by computationally addressing two challenges. The first challenge concerns the lack of systematic ways to design selection pathways in directed evolution of enzymes, an iterative process of creating mutant libraries and choosing desired phenotypes through screening or selection until the enzymatic activity reaches a desired goal. Identifying high-throughput screens or selections to isolate the variant(s) with the desired property is the biggest challenge in directed enzyme evolution, as there are currently no known generalized strategies or computational techniques to do so. This thesis presents a computational metabolic engineering framework, termed Selection Finder (SelFi), to construct a selection pathway from a desired enzymatic product to a cellular host and to couple the pathway with cell survival. When applied to construct selection pathways for several target enzymes and their desired enzymatic products, SelFi identifies selection pathways that were previously manually designed and experimentally validated.

The second challenge concerns the interpretation of data measured through untargeted metabolomics, where molecular masses of thousands of small molecules are measured simultaneously via mass spectrometry. Annotating the masses by assigning them a chemical identity and interpreting their biological relevance is challenging, as a particular mass may be associated with multiple chemical compounds.

This thesis contributes to solving the metabolite interpretation challenge in two ways. This thesis presents a novel computational workflow, termed Expanded Metabolic Model based Annotation (EMMA). EMMA constructs a biological filter consisting of an Expanded Metabolic Model (EMM) that includes not only the canonical substrates and products of enzymes, but also metabolites that can form due to substrate promiscuity, where an enzyme transforms other substrates in addition to its natural substrate. This expanded model is used to reduce the number of candidate chemical identities from large chemical databases that can be assigned to the measurements. EMMA is applied to two untargeted metabolomics data sets. Compared to a basic annotation workflow that analyzes every candidate compound in large chemical databases, EMMA reduces the number of calculations by 4 orders of magnitude. Additionally, EMMA increases the number of annotated masses by average of 1.71 and 2.39-fold, respectively, when compared to using the sample's metabolic model. Further, the results show that EMMA increases the number of annotated masses and biologically relevant candidate molecules by the average of 2.65 and 2.80-fold, respectively, when compared to using candidate sets from a biological database. The EMMA workflow was experimentally validated by confirming the presence of 4-hydroxyphenyllactate, a Chinese Hamster Ovary (CHO) cell metabolite in the EMM that has not been previously identified as part of CHO cell metabolism.

Further contributing to metabolite interpretation, this thesis presents a novel probabilistic approach, termed Probabilistic modeling for Untargeted Metabolomics Analysis (PUMA), for predicting the likelihood of activity of metabolic pathways by assigning measurements directly to metabolic pathways and then deriving probabilistic assignment of measurements to candidate chemical identities. This approach captures measurements and metabolic models within a probabilistic model, and uses stochastic sampling to compute posterior probability distributions. When applied to a test case, pathway activity results are biologically meaningful and distinctly different from those obtained using statistical pathway enrichment techniques. Further, annotation results are in agreement with those obtained using other tools that utilize additional information in the form of spectral signatures.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

The analysis and engineering of living cells have become central in many applications ranging from synthetic biology to personalized medicine. Experimental efforts to advance these applications however are expensive and time-consuming and can benefit significantly when coupled with computational efforts. Computational approaches can expedite biological discoveries and engineering efforts by interpreting measurements, guiding experiments, efficiently exploring design spaces, and identifying design alternatives.

This thesis addresses challenges related to the analysis and engineering of living cells. To expedite experimental efforts for improving enzymatic function, the thesis develops a computational method to engineer a cellular host to select for altered enzymatic function. To interpret measurements collected through untargeted metabolomics, the thesis develops computational methods to associate measurements with chemical identities and to interpret measurements in a biological context.

1.1 Directed Evolution of Enzymes

Directed evolution has emerged as a key technology to generate mutants of enzymes with new or improved properties, such as altered substrate specificity and enantioselectivity [1], thermal stability [2] [3] [4], and organic solvent resistance [5] [6]. A prominent example is commercially viable subtilisin, whose stability in detergent solutions was enhanced using directed evolution [7]. Several other such successful products include potent therapeutic agents [8] [9] [10], novel vaccines [11] [12], and potent antibodies [11].

The goal of directed evolution is to enhance the enzymatic activity of a target enzyme towards a desired functional goal. Once a target enzyme with engineering potential is identified, an iterative process of creating mutant libraries and choosing desired phenotypes over a synthetic fitness landscape is then initiated until the goal is achieved or the desired property cannot be further improved. Significant research efforts focused on developing methodologies to create larger mutant libraries with greater functional diversity [13] (e.g., tunable error-prone PCRs, saturation mutagenesis, indel mutagenesis, gene shuffling and homology-independent recombination).

There are two techniques to identify desirable variants in mutant libraries [14] [15]. With screening, the desirable property is linked to a visual output signal such as color to identify functional mutants. Every mutant of the enzyme is evaluated for the desired property. With selection, the desired property is linked to an essential metabolic function such as cell survival. Selection therefore automatically eliminates nonfunctional variants of enzymes and only positive variants are used for the next iteration of directed evolution. Selection allows for the assessment of large mutant libraries. Given its high throughput, selection is preferable to screen.

Currently, the biggest bottleneck in directed enzyme evolution is identifying high-throughput screens or selections to isolate the variant(s) with the desired property. Novel platforms to screen larger libraries have been aided by technologies like Fluorescence-Activated Cell Sorting (FACS) and microfluidic devices. However, these ultrahigh-throughput screening methodologies have primarily enabled engineering of non-catalytic function such as protein stability or binding affinity. Adaptation of these methods to catalytic functions has lagged far behind due to the inability to generically link any biochemical transformation to readouts like cell density or fluorescence. Hence, most directed evolution of enzymes are still largely limited by the inability to identify and implement selections or screens. This bottleneck is widely recognized in the field [16] [17], yet little has been done to address this concern as a whole.

1.2 Data Analysis for Untargeted Metabolomics

Metabolomics is an expanding field of research that involves the characterization of small molecules in cells, tissues and other biological systems. As metabolites within the cell are the results of both genetic and environmental factors, metabolomics offers great advantages over other omics in characterizing the phenotype [18]. Metabolomics now plays a significant role in many diverse scientific applications. It has been broadly adopted in the discovery of biomarkers for diseases such as pre-diabetes [19], diabetes [20], cancer [21], Parkinson's disease [22], Crohn's disease [23], and many others. In pharmacometabolomic studies, biochemical changes and pathway engagement can be associated with drug responses, paving the way to more individualized treatments (e.g., [24] [25]). In environmental metabolomics, metabolic responses of both plants and animals to temperature, water, food, and other aspects of the environment across individuals and populations can shed light on various aspects of ecophysiology, ecology, and genetic adaptation [26].

Several factors have spurred the increased use of metabolomics-based studies over the past decade. New generations of mass spectrometers that offer improved robustness, high resolution, and greater mass accuracy are now available. Importantly, the ability to measure the molecular masses of thousands of small molecule metabolites simultaneously, a technique known as *untargeted metabolomics*, allows unprecedented opportunities to characterize the phenotype of the particular biological sample under study. Coupling of mass spectrometry with gas and liquid chromatographic separation systems provide valuable additional information for elucidating the chemical identities of the measurements. Using a mass spectrometer, molecules within a biological sample are ionized and sorted based on their mass-to-charge ratio (m/z). In hyphenated mass spectrometry, where gas or liquid chromatography or MS is followed by an additional MS step, ionized molecules are fragmented by a number of disassociation techniques (e.g., collision-induced disassociation), and additional measurements are collected in the form of a spectral signature. Each spectral signature comprises a chromatographic retention time (RT) paired with mass measurements (m/z) for a particular metabolite and of its fragments.

Realizing the full potential of untargeted metabolomics hinges on solving two problems. The problem of *metabolite annotation* concerns associating measured masses with their chemical identities. This problem is challenging, as a particular mass may be associated with multiple chemical formulas (e.g., there are 10,132 known structural formulae for molecules with the same mass as $C_{20}H_{22}N_2O_4$). Fragmentation information in spectral signatures plays a critical role in distinguishing molecules with the same mass. Spectral signatures can be looked up in spectral databases that catalogue experimentally generated fragmentation patterns (e.g., METLIN [27], HMDB [28], MassBank [29], or NIST [30]). The coverage of spectral libraries, however, is limited due to the burden of experimentally generating spectral signatures. Alternatively, computational methods that either mimic the ionization and fragmentation process or utilize machine learning techniques (e.g. MetFrag [31], Fragment Identificator (FiD) [32], CFM-ID [33] and CSI:FingerID [34]) score the measured spectra against those in a *candidate set*. The user specifies this set, as either molecules within a particular database (e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG) [35]), or molecules within a metabolic model that corresponds to the biological sample. The quality of the results depends on the candidate set. If the compound that corresponds to the measured spectral signature is not in the candidate set, then the spectral signature cannot be annotated. Setting the candidate set to molecules from a large database such as PubChem [36] increases the chance of discovery. However, not all molecules in large databases are biologically relevant. Further, the computational cost can be prohibitive as the runtime of annotation tools is a function of the number of molecules in the candidate set. Selecting metabolites within a metabolic model as the candidate set is also problematic. Metabolic models assembled through genome reconstruction are incomplete. Further, enzymes are generally promiscuous, where an enzyme transforms other substrates in addition to its natural substrate. This form of enzyme promiscuity is referred to as *substrate promiscuity* [37]. Recently developed databases (e.g., MINEs database [38], MyCompoundID [39]) aim to catalogue novel chemical structures due to promiscuous enzymatic activities. These databases however are large and not specific to the biological sample under study. This thesis investigates a systemic method to create an Extended Metabolic Model (EMM) that includes putative metabolites due to substrate promiscuity. Further, this thesis explores the use of metabolites within this extended model as the candidate set to increase annotation beyond what is possible with a metabolic model without incurring large computational costs associated with exploring large databases.

The second problem concerns interpreting data collected through untargeted metabolomics to determine their biological role. Interpreting measurements in the context of metabolic pathways, a problem referred to as *pathway enrichment analysis*, provides a framework to study coordinated changes arising in response to cellular responses to genetic and environmental perturbations. Statistical tests, such as Fisher's exact test, determine pathways that are statistically enriched with measured metabolites compared to other pathways in the sample. Current enrichment methods for metabolomics, however, do not account for uncertainty in metabolite annotation as it is assumed that measurements are properly annotated with the correct chemical identities. This thesis explores the use of hierarchical graphical modeling and Bayesian inference to determine the likelihood of pathway activities. Instead of using the annotated measurements to determine metabolite annotations, the measurements are directly used to interpret pathway activities. Further, the pathway activities are used to determine likely metabolite annotations.

1.3 Thesis Contributions

This thesis presents three computational methods. Selection Finder (SelFi) is the first computational method that synthesizes and integrates a selection pathway within a cellular host to advance the directed evolution of enzymes. Expanded Metabolic Model Annotation (EMMA) is a novel computational workflow to enhance metabolite annotation. Probabilistic Modeling for untargeted Metabolomics Analysis (PUMA) predicts the likelihood of activity of metabolic pathways and derives probabilistic assignment of measurements to candidate metabolites. Collectively, these methods advance the engineering and analysis of biological systems. The key contributions of the thesis are as follows:

- Designing SelFi to combine synthesis pathway construction with metabolic engineering knockout strategies to ensure coupling of selection pathways with cell survival.
- Demonstrating that SelFi identifies high-quality selection pathways for several enzymes with desired reaction products, where some identified pathways are confirmed as valid selection schemes based on published literature, while others present potential valuable alternate strategies to demonstrated selection schemes.
- Developing the concept of an Expanded Metabolic Model (EMM), a metabolic model that includes metabolites that can form due to substrate promiscuity.
- Using EMMs within the EMMA framework to identify biologically relevant candidate metabolites for annotation and to allow metabolite annotation of novel molecules associated with the biological sample.
- Showing that EMMA increases annotation beyond what is possible with a metabolic model without incurring large computational costs associated with exploring large databases.
- Using EMMA to guide the experimental verification of 4-hydroxyphenyllactate, a Chinese Hamster Ovary (CHO) cell metabolite that has not been previously identified as part of CHO cell metabolism.
- Developing PUMA to utilize hierarchical graphical modeling and inference to approximate posterior probabilities of pathway activities and metabolite annotations.

- Demonstrating through PUMA the capabilities of Bayesian reasoning in evaluating pathway activities and contrasting them with those obtained using more traditional statistical pathway enrichment methods.
- Applying PUMA to metabolomics datasets and showing: (a) high level of agreement in annotation between PUMA and other annotation approaches that utilize additional information in the form of spectral signatures, and (b) an increase in the number of measurements that can be annotated over those obtained using other tools.

1.4 Thesis Organization

This thesis consists of 6 chapters. Chapter 2 provides background for synthesis pathway construction and gene modification techniques as they relate to creating selections to isolate desired enzymatic mutants. Chapter 2 also provides background on metabolomics analysis including metabolite annotation and pathway enrichment techniques.

Chapter 3 describes how SelFi constructs selection pathways and identifies knockout targets required to link cell survival with the selection pathway. SelFi is evaluated by applying it to engineer selections for four enzymatic reactions.

Chapter 4 presents a computational workflow, EMMA, to enhance the chance of biological discovery while speeding metabolite annotation. Using data collected through untargeted metabolomics, the results of EMMA are compared to those obtained using alternate annotation workflows in terms of computational cost and size of the candidate sets. Some of the computational results are experimentally evaluated.

Chapter 5 presents PUMA to explore the use of graphical hierarchical modeling and inference to predict the likelihood of pathway activities and metabolite annotation. Applied to untargeted metabolomics datasets, PUMA is evaluated and compared to other pathway enrichment and metabolite annotation techniques.

Chapter 6 summarizes the thesis and outlines directions for future research.

Chapter 2
Background

The contributions of this thesis are in two areas. One contribution, in the area of directed evolution of enzymes, involves the construction of a selection pathway from a desired enzymatic product to a cellular host then engineering the cellular host to couple the pathway with cell survival. The second contribution advances the annotation and interpretation of measurements collected through untargeted metabolomics.

2.1 Metabolic Engineering for Directed Evolution of Enzymes

Despite the adoption of computational tools such as synthesis pathway construction and gene modifications in synthetic biology and metabolic engineering, there are no computational tools that automatically design and integrate selection pathways for the directed evolution of enzymes. This thesis adapts and integrates synthesis pathway construction and genetic modification tools to engineer selection pathways that isolate mutants of an enzyme with a desired functionality.

2.1.1 *Synthesis Pathway Construction*

Designing novel synthesis pathways to generate desirable compounds that are not naturally produced by a cellular organism allows the production of useful compounds such as high-valued industrial chemicals. A synthesis pathway includes a sequence of reactions steps that convert a source compound within the cellular host to a target compound. Computational tools for the construction of synthesis pathways can be classified as either graph-based or rule-based [40]. Graph-based approaches exploit compound and reaction data in databases such as KEGG [35] for synthesis pathway construction. By analyzing the database as a graph, with compounds as nodes and reactions as edges, graph-based approaches seek to find a path in the graph from a starting compound to a desired target product. For example, PathMiner [41] [42] seeks to build pathways that minimize the biochemical transformation cost. This approach favors reactions involving the addition of smaller functional groups, which can select against canonical modifications such as phosphorylation. OptStrain [43] is a pathway synthesis approach where a mixed integer linear programming framework is utilized to identify high-yielding stoichiometrically balanced synthesis pathways by adding or deleting reactions from a curated database to the host metabolic network. *ProPath* [44] identifies non-naive synthesis pathways between two compounds by probabilistically sampling available reactions within a database as the search space and employment of a backtracking algorithm to explore the search space recursively to find a solution.

Rule-based approaches have the advantage of constructing synthesis pathways for metabolites with no known reactions catalogued in databases. Using either known transformations or compound-reaction data stored in a database, rule-based approaches generalize an existing reaction into a transformation rule between compounds. One approach, integrated Computational Explorer (BNICE) [45] [46] utilizes the EC classification number of enzymes to generate the transformation rules. Another approach, PathPred [47], utilizes transformational patterns derived from KEGG RDM patterns [48] to generate general transformation rules between each reactant and product pair. University of Minnesota Pathway Prediction System (UM-PPS) [49] is another rule-based approach that was developed to identify novel biodegradation pathways. Metabolic rules based on organic functional groups are derived from The University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD).

2.1.2 *Gene Modifications*

Increasing the rate of producing a target molecule within a cellular host requires modifying the host through either deletions or up/down regulation of a selected set of genes. Computational tools can be used for the identification of gene modifications. For example, OptKnock [50] uses bi-level programming to identify gene deletions that satisfy the coupled objectives of target overproduction and biomass formation. There are several improvements to OptKnock (e.g., OptReg [51], OptORF [52], MOMAKnock [53], and RobustKnock [54]). In addition to bi-level optimization based techniques, OptGene [55] employs the principle of evolution to identify gene knockout targets. CCOpt formulation [56] is the first work to incorporate uncertainties when computing gene modifications, where constraints in CCOpt are probabilistically met at a user-specified confidence level. OptForce [57] and CosMos [58] mathematically identify required coordinated changes among reactions.

2.2 Interpreting Measurements Collected through Untargeted Metabolomics

The current gold standard for assigning a chemical identity to a measurement collected through untargeted metabolomics is to verify the measured spectral signature against that of an authentic standard using the same equipment and settings. This method is impractical, as it requires costly and enormous experimental efforts. Importantly, candidate matches for testing must be identified *before* any experimental authentication. There are now spectral databases that catalogue chemical structures and their spectral signatures and there are *in silico* annotation tools to computationally determine the likelihood of matching the query spectra to a specified set of candidate metabolites. Further interpretation of metabolomics data, mostly in the form of computing pathway activities, is enabled by statistical and topological pathway enrichment techniques.

2.2.1 *Spectral Databases for Annotation*

Due to the impracticality of establishing in-house libraries, analysis of MS/MS data often relies on reference databases such as METLIN [27], HMDB [28], MassBank [59], and NIST [30]. However, the coverage of compounds in these databases remains limited. For example, NIST contains ~652,000 spectra for 15,243 unique compounds [60], roughly 0.025% of the more than 60,000,000 catalogued compounds in PubChem [36]. Other spectral databases contain fewer spectra, with METLIN currently cataloging ~16,000 spectra [61] and MassBank cataloguing 74,447 unique compounds across 211,545 spectra [62]. The number of covered unique compounds is just over to 2,256 for HMDB [63]. These spectra sometimes correspond to both peptides and small molecules (e.g., as in NIST and METLIN), and across multiple platforms (e.g., LC-MS, GC-MS, etc.). As a result, only a small percentage of total number of measurements detected in a sample can be annotated using spectral libraries.

2.2.2 *In silico Metabolite Annotation*

Earlier software packages for annotation such as Mass Frontier [64], ACD/MS Fragmenter [65], and Hammer [66] are tools that have been developed to predict fragmentation patterns for an input chemical structure. MetFrag [67], Fragment Identificator (FiD) [68], Fragment Formula Calculator [69], and Mass Spectrum Interpreter [60] are other approaches that have been introduced for *in silico* fragmentation pattern prediction. More recently introduced *in silico* annotation tools, such as CFM-ID [70], and CSI:FingerID [71] employ machine learning algorithms. CFM-ID provides estimated fragmentation based upon a probabilistic, generative model. The algorithm of CFM-ID enumerates the possibilities of bond breaking in a molecule structure in a breadth-first manner, computing the probability that each bond breaks. The algorithm is then applied recursively on fragments of the original molecule structure that can be generated due to high probability bond breakings. Once the probabilities are generated, the algorithm predicts a spectral signature for the input molecule structure. Having spectral signatures predicted for a list of known metabolites, CFM-ID uses them to compare against an unknown spectral signature for annotation. The algorithm in CSI:FingerID consists of two phases, learning and prediction. In the learning phase, the algorithm uses a database of reference compounds with known molecular structure, computing a molecular fingerprint for each compound. The algorithm then trains a Support Vector Machine (SVM) on each molecular property in the fingerprint. In the prediction phase, the algorithm predicts a fingerprint for an unknown compound by using the trained SVMs to predict the probability of absence or presence of each molecular property in the unknown compound. The runtime of *in silico* metabolite annotation techniques can be computationally prohibitive as the computational cost is a function of the number candidate metabolites under consideration.

Exploiting biological context can enhance metabolite annotation by focusing on biologically relevant candidate metabolites. A method is described for identifying potential substrate-product pairs based on the mass change in manually curated well-known metabolic conversions and the mass differences between pairs of detected mass spectrometry features [72]. Another method, iMet [73], exploits the fact that neighboring metabolites within a metabolic network have similar MS/MS spectra and trains a classifier to predict the closest neighbor in databases for an unknown query spectra. BioCAN utilizes annotation evidence that is collected through spectral databases and *in silico* annotation tools in the neighborhood of a measured mass to determine the most likely annotation [74].

2.2.3 *Pathway Enrichment Analysis*

Pathways represent a connected set of reactions and metabolites that are involved in performing a particular function such as glycolysis or the tricarboxylic acid (TCA) cycle. Pathways are curated based on the literature and domain knowledge and catalogued in databases. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) manually curated metabolic pathways for thousands of organisms [75]. MetaCyc contains 2,609 pathways from 2,914 different organisms [76]. The Small Molecular Pathway Database (SMPDB) catalogues 30,000 small molecule pathways found in humans [77]. There are now several computational techniques to perform pathway enrichment analysis based on metabolomics data. These techniques can be broadly classified in two categories: Overrepresentation Analysis (ORA) and Topological Analysis (TA). ORA employs statistical testing (e.g., Fisher’s exact test) to determine if a dataset is enriched in a particular set of metabolites to a degree greater than expected by chance, given a set of pathways assumed to be expressed in the biological system of interest. Pathway Enrichment Analysis (PEA) performs a similar test based on the measured concentrations of metabolites. Metabolomics data have been also analyzed using tools originally developed for gene expression analysis (e.g., globaltest [78]). For example, MSEA, a web-based pathway analysis tool, employs gobaltest to implement ORA [79]. TA estimates the observed metabolites' centrality and connectivity, which measure the importance of a metabolite in the flow of material through a pathway or network. MetaboAnalyst [80] is a web-based platform featuring a number of pathway analysis tools, which afford integration of metabolite and gene expression data to explore enriched pathways based on the joint evidence from these two types of data. A similar capability is available through Integrated Molecular Pathway-Level Analysis (IMPaLA) [81], which performs overrepresentation and enrichment analysis with user-specified lists of metabolites and genes by referencing a large number of pathways cataloged in multiple databases. A recent comparison has shown that current ORA techniques provide consistent results regardless of their approach [82]. ORA and TA techniques, however, do not address issues related to uncertainty in metabolite annotation.

The two problems, metabolite annotation and pathway enrichment analysis, have traditionally been solved as two independent problems, where path enrichment assumes that the chemical identity of each measured mass is known *a priori*. One exception is Mummichog, a set of robust statistical algorithms that predicts functional activity directly from measurements, circumventing annotation [83]. The biological context encoded in pathways/modules aids in reducing and in some cases eliminating ambiguity in metabolite annotation. Mummichog produces quality results in agreement with validated annotation in experimental studies.

Chapter 3
Selection Finder (SelFi): A Computational Selection Finder for Directed Evolution of Enzymes

We present in this chapter a computational metabolic engineering framework, SelFi, to identify high-throughput selections to isolate active mutant enzyme with a desired catalytic function. Such an enzyme catalyzes a reaction that transforms the enzyme’s precursor to an enzymatic product desired for its beneficial industrial use [84]. Selection is a technique that automatically identifies functional mutants by linking the desired catalytic functionality to cell survival. In this work, the link is established by first constructing a pathway from the desired enzymatic product to a molecule within the host, which is then engineered to make consumption of the desired product essential for cellular growth. The pathway includes a series of reactions from the desired enzymatic product to a molecule within the cellular host. The pathway is referred to as a *consumption pathway*, as it provides a mechanism for the cellular host to consume the desired enzymatic product. The pathway is constructed using an adaptation of *ProPath* [44], an algorithm for constructing synthesis pathways that transform a molecule in the host to a desirable target molecule typically not produced by the host. Given a desired enzymatic product, our framework identifies several candidate selection pathways and corresponding genetic engineering strategies for the host. The candidate pathways are then ranked based on predicted consumption flux and required cellular engineering efforts. An ideal selection provides maximum dynamic range with minimal cellular engineering effort.

SelFi identifies a selection pathway in four steps. In the first step, SelFi constructs traversal pathways to consume the desired enzymatic reaction product and convert it to a native metabolite within the cellular host. Utilizing *ProPath* [44], reactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [35] are used to construct candidate consumption pathways. In the second and third steps, SelFi identifies the minimal set of carbon sources and knockout targets required to link cell survival with the consumption pathway and to guarantee a minimum flux. In the last step, SelFi ranks the resulting pathways based on flux, pathway length, and number of required knockouts.

Modifications identified by SelFi can be experimentally utilized as follows. The mutagenized library of the enzyme to be engineered and the identified consumption pathway along with supporting pathways are co-expressed (using, for example, plasmids) in the selection strain with the identified knockouts. This will create a high-throughput pooled library system from which the desired enzymatic reaction will be selected.

3.1 Methods

3.1.1 *Construction of Consumption Pathways*

To identify selection pathways, we utilize a modified version of *ProPath* [44], a probabilistic algorithm for constructing synthesis pathways that start from a metabolite within the host and end with a desirable target. Using reactions in the KEGG database, *ProPath* recursively explores a tree representing all possible synthesis pathways that start from the target metabolite (Figure 1).

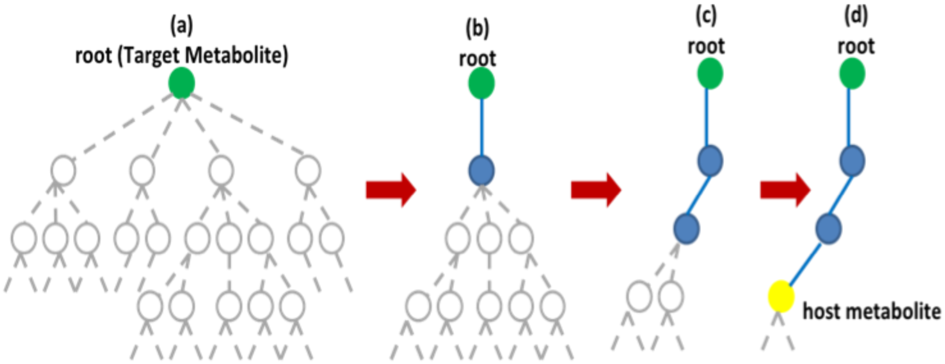


Figure 1. Probabilistic pathway construction using the *ProPath* algorithm. The dashed and solid lines show the possible routes and selected reactions, respectively. (a) Tree representing all possible synthesis pathways for a target metabolite. The root of the tree is the target metabolite. (b) and (c) only one reaction is selected at a time, in a depth-first fashion and (d) recursive exploration terminates at a metabolite within the host network.

ProPath selects a single reaction from a list of candidate reactions in the KEGG database that involve the target metabolite as a product. Reaction selection occurs with equal likelihood of selecting a candidate reaction. The selected reaction, represented by an edge, is added to the tree. This edge expands the tree by attaching new nodes representing the product metabolites and cofactors of the selected reaction. Further pathway construction proceeds in a depth-first fashion. Each added node becomes a new root for the construction, unless the corresponding metabolite is already present in the host organism or previously added to the tree. A limit is set on the number of reactions that can be used to construct a pathway. When the addition of a reaction to the tree violates this limit, the search algorithm backtracks. The algorithm then proceeds by adding to the tree another reaction that has not been previously explored, effectively exploring an alternative pathway. If none of these alternative routes satisfy the pathway length limit, the algorithm further backtracks and continues from there. The algorithm finishes when all permitted-length branches of the tree terminate in a metabolite that is native to the host organism. Due to the probabilistic nature of selecting the reactions, the completed tree does not exhaustively enumerate all possible pathways. Rather, each tree represents a single pathway from the target metabolite to one or more metabolites that are native to the host. The search is iterated many times to explore a diverse number of possible pathways. Our route construction is based on *ProPath*, as it was shown effective in generating synthesis pathways with fluxes comparable with those reported for limited-in-depth exhaustive search methods. Additionally, *ProPath* was able to reproduce experimentally obtained pathways published in the literature. We reverse *ProPath*'s search direction to identify a pathway starting from a compound of interest to an endogenous metabolite within the host. We refer to the use of the algorithm in this reversed manner as *retroProPath.* In *retroProPath,* the root of the tree is also the desired product of an engineered enzymatic reaction. The first set of edges added to the tree represents KEGG reactions that *consume* the desired enzymatic product. A single reaction among them is selected with equal likelihood to expand the tree. *retroProPath* continues the probabilistic search using product-side non-cofactor metabolites of the selected reaction. The in-depth path construction terminates when a metabolite within the host is reached or the path length limit is reached. The algorithm backtracks to explore a different pathway, as needed. To ensure that reactant-side cofactors associated with the identified pathway are available to the cell, SelFi utilizes *ProPath* to construct supporting synthesis pathways that start from a metabolite within the host and terminate at each such cofactor. The pathway from the enzymatic product to a metabolite within the host along with these supporting synthesis pathways are referred to as a *consumption* pathway. Other pathway synthesis methods (e.g., PathPred [47], PathMiner [42]) can be utilized in place of *ProPath*.

3.1.2 *Evaluating Consumption Pathway Flux and Consumption Demand*

Flux Balance Analysis (FBA) [85] is a constraint-based approach for calculating the flow (flux) in a metabolic network under steady-state conditions. A metabolic network consists of m metabolites and n reactions. An $m \times n$ matrix, $\mathbf{S}$, represents the network where each row corresponds a metabolite and each column corresponds to a stoichiometrically balanced reaction. Matrix entry $s_{i,j}$ represents the stoichiometric coefficient of metabolite i in reaction j. A vector, $\mathbf{v}$, of length n, represents the flux in all network reactions. FBA uses linear programming to solve a particular cellular objective such as maximizing biomass production, or minimizing or maximizing the flux for a particular reaction. A set of equations, $\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$, constrain the system to operate in steady state. Additional constraints imposed by physiological conditions and metabolite exchange fluxes are represented as upper and lower bounds on each reaction flux.

In this work, FBA is utilized to evaluate the maximum and minimum consumption flux through the engineered enzyme, and hence through the consumption pathway. In addition, FBA is utilized to assess the consumption demand for metabolites within the host. A consumption pathway terminating at a high-demand metabolite within the host suggests the possibility of constructing a high-flux consumption pathway. A consumption pathway terminating at a low-demand metabolite explains the low yield associated with such a consumption flux.

We define the *consumption demand* of a metabolite as the maximum sum of all outgoing fluxes consuming the metabolite. The consumption flux of metabolite *i* with *k* consuming reactions connected to the metabolite can be calculated using FBA by solving the following optimization problem:

$$\begin{gathered}
\text{Maximize } \sum_{j=1}^{k} v_{ij} \\
\text{subject to} \\
\mathbf{S} \cdot \mathbf{v} = \mathbf{0} \\
v_{bio} \geq v_{bio}^{min} \\
v_{ij}^{lb} \leq v_{ij} \leq v_{ij}^{ub} \ \text{ for } j = 1 \cdots k
\end{gathered}$$

Where v_{ij} represents the consumption flux of metabolite *i* through its *j*th consuming reaction, $\mathrm{v}_{\mathrm{ij}}^{\mathrm{lb}}$ and $\mathrm{v}_{\mathrm{ij}}^{\mathrm{ub}}$ are respectively the associated lower and upper bounds for v_{ij}. $\mathrm{v}_{\mathrm{bio}}$ represents the rate of biomass production, and $\mathrm{v}_{\mathrm{bio}}^{\mathrm{min}}$ represents the minimum desired biomass production rate. To identify the maximum demand while allowing for reaction reversibility, several instance problems are considered. With *m* reversible reactions connected to metabolite *i*, we consider that each reversible reaction can operate in forward and reverse directions. We then generate 2^m possible objective functions for the optimization problem, representing all possible reaction directionalities associated with the consumption flux. Using FBA, we assess consumption flux of metabolite *i* based on each of the objective functions for a desired biomass production rate. Among all feasible solutions obtained by FBA, we select the maximum as the consumption demand associated with metabolite *i*. In this work, we evaluate the consumption demand for metabolites within the cellular host. Consumption pathways that terminate on host metabolites with low consumption demand may not be suited for high-throughput selection.

3.1.3 *SelFi Framework*

Given an engineered enzymatic reaction and its reactant and product, SelFi constructs pathways that consume the enzymatic product, and then engineers the cell to couple the consumption pathway with cell survival. SelFi has four steps - the outcomes of which are illustrated in Figure 2.

Step 1. Constructing consumption pathways

Using *retroProPath*, SelFi constructs a set of possible consumption pathways from the desired product to a metabolite within the cellular host. Based on the practicality of simultaneous gene insertions, the pathway length limit is set to 20 [86]. While most cofactors required by the consumption pathways (e.g. H^+, O_2, CO_2, $NAD(P)^+$, NAD(P)H) are likely native to the host, SelFi constructs synthesis pathways from the host to the cofactors if needed. To do so, SelFi first determines if all reactant-side cofactors are native to the host. If a cofactor is not native, SelFi uses *ProPath* to construct synthesis pathways starting with a host metabolite to the cofactor. Figure 2 shows an example pathway construction. The blue pathway is constructed using *retroProPath*, and provides a consumption pathway from the desired enzymatic product (green number 2) to a metabolite within the host (yellow number 1). A supporting pathway (orange) from a metabolite native to the host (yellow number 2), to a non-native cofactor on the reactant side of the reaction along the selection pathway is constructed using *ProPath.* While this step identifies consumption pathways, these are not yet high flux selection pathways since a link to cellular viability is not yet established. The following steps address these requirements.

Step 2. Eliminating alternate carbon sources

Selection requires linking the enzymatic product to a metabolic function essential for cell viability. This can be accomplished by forcing the consumption pathway to be the only cellular carbon source. SelFi therefore eliminates all organic carbon uptakes except for the uptake provided through the consumption pathway, and inorganic CO_2, an essential waste product. As a cellular host, *E. coli* has many carbon uptakes including D-glucose, D-fructose, D-galactose, D-mannose, D-xylose, L-arabinose, D-ribose, D-glyceraldehyde, and glycerol. Typically, the cell utilizes only one such carbon source at a time for growth. Mathematically, the elimination of carbon uptake by a specific reaction can be modeled by setting its minimum and maximum operating flux to zero. In the example provided in Figure 2, the precursor of the desired enzymatic product (green number 1) is not native to the host. Eliminating external carbon sources, marked by red “×”, couples the consumption pathway (blue) with cell survival.

In some cases, the reactant of the specified enzymatic reaction is native to the host, and limiting carbon uptake to be only through the consumption pathway is not possible. Here, an external carbon source must be provided to keep the cell alive. FBA is used to determine the external source that maximizes flux through the consumption pathway. For each carbon source, the consumption pathway flux is maximized while constraining the biomass production rate to be at least 10% production of the wild-type maximum biomass rate. Selecting a carbon source that maximizes the consumption flux does not result in coupling a consumption pathway with cell survival as the cell is not reliant on the consumption pathway. This issue is addressed in Step 3.

Step 3. **Identifying knockout targets**

SelFi seeks one of two goals in this step: coupling the consumption flux with cell survival, if that is not accomplished in Step 2, and improving guaranteed non-zero minimum consumption flux. SelFi can successfully couple the consumption pathway to the host survival in Step 2 except when the reactant of the given enzymatic reaction is native to the host. If the reactant is not native, survival and consumption are automatically coupled in the absence of alternate carbon sources and a minimum non-zero consumption flux is guaranteed in Step 2. Knockouts can improve this guaranteed minimum consumption flux. In presence of native reactant for the enzymatic reaction, survival-consumption coupling is not guaranteed in Step 2. In this case, SelFi searches for knockout targets in the host to guarantee non-zero minimum consumption flux to ensure the consumption pathway is linked to cell survival.

To identify possible knockout targets, SelFi utilizes a sequential greedy strategy. SelFi explores knocking out one reaction at a time using FBA to calculate the minimum flux through the consumption pathway. Among knockout targets that improve the minimum consumption flux, SelFi selects the one that improve the minimum flux the most. SelFi continues to find an additional knockout target that improves the flux, repeating this process until the maximum allowed number of knockouts, as specified by the user, is reached. In this work, we set the number of knockouts to three to limit the computational cost associated with evaluating each consumption pathway. Alper and *et al*. used a similar greedy knockout strategy to maximize the production of lycopene in *E. coli* [87].

Step 4. Ranking selection pathways

For each pathway identified by Steps 1-3, SelFi generates a listing of reactions in the selection pathway and their corresponding supporting pathways needed to generate co-factors. SelFi reports the total number of steps in both the selection and support pathways, and the computed guaranteed minimum and maximum consumption fluxes before and after knockouts. SelFi provides all information such that the end user can explore various options. Ideally, candidate pathways are chosen based on the guaranteed minimum consumption flux, pathway length, and number of required knockouts. Shorter, higher-consumption flux pathways with the smallest number of knockouts are preferable over others.

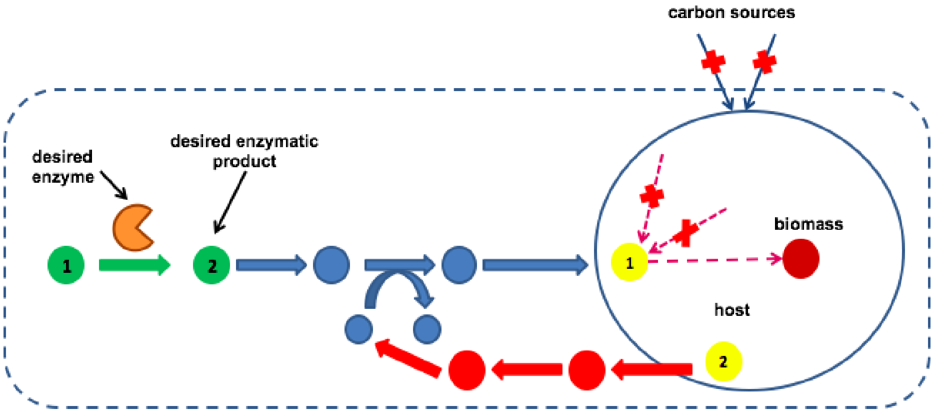


Figure 2. Illustration of SelFi implementation. The large round circle indicates boundaries of the wild-type *E. coli*, and the dotted box indicates boundaries of the cell after co-expression of the selection system. The desired enzyme catalyzes a reactant (green number 1) to a desired product (green number 2). A consumption pathway (blue) from the desired product to a metabolite (yellow number 1) within the wild-type *E. coli* is constructed using *retroProPath* to consume the desired product. A supporting pathway (orange) from a native metabolite (yellow number 2) in the wild type to a cofactor on the reactant side of a consumption pathway is constructed using *ProPath*, if needed. An “x” within the cell indicates a knockout, and an “x” outside the cell indicates eliminating carbon sources.

3.2 Results

To analyze the effectiveness of our algorithm, we applied SelFi to several test cases including desired products xylitol, D-ribulose-1,5-bisphosphate, methanol, and aniline that can be potentially synthesized through the action of engineered enzymes Xylose Reductase (XR), Phosphoribulokinase (PRK), Methane Monooxygenase (MMO), and Aromatic Amino acid Decarboxylase (AAD), respectively. We utilized the genome-scale model of *E. coli* metabolism (*i*AF1260) [88] as the host organism. The *i*AF1260 model constraints were modified as follows. A constraint is added to ensure that the biomass flux is equal to or greater than 10% of the maximum biomass flux rate of the wild type. The lower and upper bounds on oxygen uptake were set to −1000 and 1000 (mmol/gDCW/hr) respectively to allow for aerobic growth conditions. The lower and upper flux bounds for the engineered enzymatic reaction were set to 0 and 1000 (mmol/gDCW/hr) respectively. Lower and upper flux bounds of reactions along the added selection pathways were set to −1000 and 1000 (mmol/gDCW/hr), respectively.

3.2.1 *Summary of Results*

We executed *retroProPath* for 1000 iterations. Selection pathways identified by SelFi (after Step 1) are summarized in Table 1. The first column lists the product of the enzymatic reaction. The second column lists a label we assigned to each selection pathway. The first letter of the subscript indicates the product (X for xylitol, D for D-ribulose-1,5-bisphosphate, M for methanol, and A for aniline), while the second letter indicates the selection pathway number. The third column lists the reactions along identified selection pathways. All cofactors along the pathways are native to the host, thus eliminating the need for adding synthesis pathways for reactant-side cofactors.

Table 2 summarizes flux characterization results after restricting carbon uptakes (after Step 2), and after knockouts (after Step 3). The first column lists the selection pathways by their labels as designated in Table 1. The second column lists the length of the pathways. The third and fourth columns list minimum and maximum consumption fluxes before applying any knockouts. In many cases, the minimum consumption flux is zero, indicating that the added consumption pathway is not essential for growth, and that the host must be engineered through knockouts to couple the consumption pathway with cell survival. The fifth column lists the number of knockouts identified to improve minimum consumption fluxes. The two last columns show minimum and maximum consumption fluxes after applying knockouts. In this work, the knockouts were selected to provide a minimal guaranteed flux. In each case, after knockouts, the minimum consumption flux increases whereas the maximum consumption flux decreases. A higher minimum uptake rate will enable selection for mutants with higher activity. Conversely, a lower minimum guaranteed uptake rate will provide a less stringent selection, and for identification of mutants with lower activity. Thus, the knockout process provides a mechanism to place a threshold on minimum desired enzymatic activity. Table 3 summarizes the consumption demand for metabolites terminating the selection pathways identified by SelFi.

The first column lists the pathway label, while the second column lists the terminating metabolite. The following columns report the consumption fluxes calculated using FBA assuming various desired lower bounds on biomass production, expressed as a percentage of the maximum biomass production in the wild type. For each end metabolite except for L-arabinose, the consumption demand remains constant assuming 10%−70% minimal biomass production. L-arabinose drops to 1750 mmol/gDCW/hr when assuming 70% or higher minimal biomass production, while 3-dehydro-L-gulonate, D-ribose 1,5-bisphosphate, formaldehyde and 4-aminobenzoate show no change across the 10%-90% minimal biomass production range. All consumption demands are relatively high except for two end metabolites. 3-dehydro-L-gulonate is produced from 2-3-dioxo-L-gulonate, a metabolite that is not produced by any other reaction in the model. The consumption demand for 3-dehydro-L-gulonate is thus zero. If 2-3-dioxo-L-gulonate is supplied to the cell with an uptake rate of 1000 mmol/gDCW/hr, the consumption demand for 3-dehydro-L-gulonate becomes 1000 mmol/gDCW/hr assuming 10%−70% minimal biomass production, and 333 mmol/gDCW/hr assuming 90% minimal biomass production. In contrast, metabolite 4-aminobenzoate has low demand and can only be utilized in relatively small quantities for biomass production.

Table 1. Description of identified selection pathways for enzymatic products xylitol, D-ribulose-1,5-bisphosphate, methanol, and aniline.

Desired product	**Selection pathway label**	**Consumption pathway**
For engineering Xylose Reductase		
xylitol	**SP_{X1}**	xylitol + NAD^+ ↔ D-xylulose + NADH + H^+
xylitol	**SP_{X2}**	xylitol + $NAD(P)^+$ ↔ L-xylulose + NAD(P)H + H^+ L-xylulose + NADH + H^+ ↔ L-arabitol + NAD^+ L-arabitol + $NAD(P)^+$ ↔ L-arabinose + NAD(P)H + H^+
xylitol	**SP_{X3}**	xylitol + $NAD(P)^+$ ↔ L-xylulose + NAD(P)H + H^+ ATP + L-xylulose ↔ ADP + L-xylulose 5-phosphate
xylitol	**SP_{X4}**	xylitol + $NAD(P)^+$ ↔ L-xylulose + NAD(P)H + H^+ L-xylulose + NADH + H^+ ↔ L-arabitol + NAD^+ L-arabitol + NAD^+ ↔ L-ribulose + NADH + H^+
xylitol	**SP_{X5}**	xylitol + $NAD(P)^+$ ↔ L-xylulose + NAD(P)H + H^+ L-xylulose ↔ L-lyxose
xylitol	**SP_{X6}**	xylitol + $NAD(P)^+$ ↔ L-xylulose + NAD(P)H + H^+ L-xylulose + CO_2 ↔ 3-dehydro-L-gulonate
xylitol	**SP_{X7}**	xylitol + $NAD(P)^+$ ↔ L-xylulose + NAD(P)H + H^+ ATP + L-xylulose ↔ ADP + L-xylulose 1-phosphate L-xylulose 1-phosphate ↔ glycerone phosphate + glycolaldehyde
For engineering Phosphoribulokinase		
D-ribulose-1,5-bisphosphate	**SP_{D1}**	D-ribulose-1,5-bisphosphate ↔ D-ribose 1,5-bisphosphate
D-ribulose-1,5-bisphosphate	**SP_{D2}**	D-ribulose-1,5-bisphosphate + CO_2 + H_2O ↔ 2 3-phospho-D-glycerate

D-ribulose-1,5-bisphosphate	**SP_{D3}**	D-ribulose-1,5-bisphosphate + O_2 ↔ 3-phospho-D-glycerate + 2-phosphoglycolate
For engineering Methane Monooxygenase		
methanol	**SP_{M1}**	methanol + NAD^+ ↔ formaldehyde + NADH + H^+
methanol	**SP_{M2}**	methanol + formate ↔ 2 formaldehyde + H_2O
methanol	**SP_{M3}**	methanol + O_2 ↔ formaldehyde + H_2O_2
methanol	**SP_{M4}**	methanol + H_2O_2 ↔ formaldehyde + 2 H_2O
For engineering Aromatic Amino Acid Decarboxylase		
aniline	**SP_A**	aniline + CO_2 ↔ 4-aminobenzoate

Table 2. Characterizing consumption pathways, showing length of the pathways, minimum and maximum consumption fluxes before applying knockouts, number of identified knockouts and minimum and maximum consumption fluxes after knockouts, for each selection pathway

Selection pathway label	**Pathway length**	**Minimum consumption flux before knockouts (mmol/gDCW/hr)**	**Maximum consumption flux before knockouts (mmol/gDCW/hr)**	**Number of identified knockouts**	**Minimum consumption flux after knockouts (mmol/gDCW/hr)**	**Maximum consumption flux after knockouts (mmol/gDCW/hr)**
SP_{X1}	1	0	1000	3	111.66	184.09
SP_{X2}	3	0	325	3	111.66	184.09
SP_{X3}	2	0	325	3	111.66	184.09
SP_{X4}	3	0	325	3	111.66	184.09
SP_{X5}	2	0	325	3	111.66	184.09
SP_{X6}	2	0	325	3	111.66	184.09
SP_{X7}	3	0	325	3	110.27	184.46
SP_{D1}	1	0	1000	2	4.23	623.34
SP_{D2}	1	0	847.04	1	3.78	689.08

SP$_{D3}$	1	0	615.66	1	3.55	500.58
SP$_{M1}$	1	44.53	726.79	3	117.50	673.76
SP$_{M2}$	1	47.01	500	3	127.29	500
SP$_{M3}$	1	60.28	510.44	1	196.18	502.40
SP$_{M4}$	1	60.38	515.84	2	278.59	502.70
SP$_{A}$	1	0	0.01	1	0.001	0.01

Table 3. Consumption demand flux of end metabolites in mmol/gDCW/hr for each selection pathway, assuming a minimum of 10%, 30%, 50%, 70%, and 90% biomass production compared to the maximum biomass production of the wild type.

Pathway	**End metabolite in host**	**Minimum biomass production rate**				
		10%	**30%**	**50%**	**70%**	**90%**
SP$_{X1}$	D-xylulose	1000.00	1000.00	1000.00	1000.00	500.00
SP$_{X2}$	L-arabinose	2000.00	2000.00	2000.00	1750.00	1250.00
SP$_{X3}$	L-xylulose 5-phosphate	1000.00	1000.00	1000.00	1000.00	500.00
SP$_{X4}$	L-ribulose	1000.00	1000.00	1000.00	1000.00	500.00
SP$_{X5}$	L-lyxose	1000.00	1000.00	1000.00	1000.00	500.00
SP$_{X6}$	3-dehydro-L-gulonate	0.00	0.00	0.00	0.00	0.00
SP$_{X7}$	glycolaldehyde	1000.00	1000.00	1000.00	1000.00	500.33
SP$_{D1}$	D-ribose1,5 bisphosphate	1000.00	1000.00	1000.00	1000.00	1000.00
SP$_{D2-3}$	3-phospho-D-glycerate	2000.00	2000.00	2000.00	2000.00	1500.00
SP$_{M1-4}$	formaldehyde	1000.00	1000.00	1000.00	1000.00	1000.00
SP$_{A}$	4-aminobenzoate	0.05	0.05	0.05	0.05	0.05

3.2.2 *Xylose Reductase (XR) and Xylitol*

Xylitol is used as a low-calorie sweetener or platform chemical for the production of industrially important chemicals such as glycols [84]. Xylitol can be overproduced through an engineered XR enzyme (Figure 3) with D-xylose as a reactant and desired enzymatic product xylitol [89]. The purpose of engineered XR, as described by Nair and Zhao [89], is to engineer substrate specificity of XR while maintaining its activity toward the natural substrate, D-xylose.

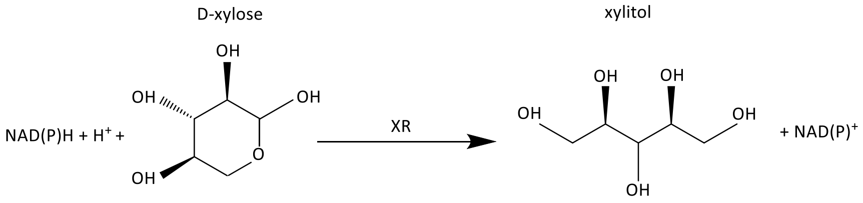


Figure 3. Reduction of D-xylose to xylitol by Xylose Reductase (XR)

SelFi identified seven consumption pathways for xylitol as specified in Tables 1 and 2. The pathways end with D-xylulose, L-arabitol, L-xylulose 5-phosphate, L-ribulose, L-lyxose, 3-dehydro-L-gulonate, and glycolaldehyde. D-xylose, the reactant of the enzymatic reaction, is native to *E. coli*. SelFi limited all external carbon sources except for D-xylose. All terminating metabolites have relatively high demand, as per Table 3, except for 3-dehydro-L-gulonate. However, with D-xylose uptake, xylitol is converted to L-xylulose, which in turn is converted to 3-dehydro-L-gulonate, and a maximum consumption flux of 325 mmol/gDCW/hr can be achieved prior to knockouts.

Table 4. Knockout targets for xylitol test case

Knockout target reaction	**KEGG ID**
D-xylose[c]* ↔ D-xylulose[c]	R01432
ADP[c] + 4.0 H^+[p]** + Pi[c] ↔ ATP[c] + H_2O[c] + 3.0 H^+[c]	R00086
D-glycerate-2-phosphate[c] ↔ 3-phospho-D-glycerate[c]	R01518

* cytoplasmic localization
** periplasmic localization

Table 5. Effect of knockouts to improve guaranteed minimum consumption fluxes. For each pathway listed in column 1, we identify knockout targets, and the corresponding minimum guaranteed flux in parenthesis.

Selection pathway label	**1st Round**	**2nd Round**	**3rd Round**
SP_{X1-X6}	Δ R01432 (21.80)	Δ R01432, Δ R00086 (57.66)	Δ R01432, Δ R00086, Δ R01518 (111.66)
SP_{X7}	Δ R01432 (21.80)	Δ R01432, Δ R00086 (57.66)	Δ R01432, Δ R00086, Δ R01518 (110.27)

SelFi identified the same set of three knockout targets for selection pathways SP_{X1} - SP_{X7}. As shown in Table 2, the minimum flux through the consumption pathways, which is zero mmol/gDCW/hr prior to applying any knockouts, is 110.27 mmol/gDCW/hr or higher after knockouts. The knockouts thus result in coupling the consumption pathway with cellular growth. Table 4 summarizes reactions identified by SelFi as knockout targets and their corresponding KEGG identification numbers (KEGG IDs). Table 5 presents the order of reactions to be knocked out and shows the extent to which each knockout improves the guaranteed minimum consumption flux. The knockout targets identified by SelFi are shown in Figure 4, where a solid arrow illustrates a single reaction, while a dashed line represents multiple reaction steps. Metabolites and reactions native to the host are enclosed in a box, while non-native ones are placed outside the box. The remaining figures utilize the same drawing convention. The complete names of abbreviations used in Figure 4 (and Figures 6−7) are listed in the Abbreviations Table in Appendix A. The first knockout target R01432 is a reaction that consumes D-xylose (xyl-D) as its reactant. Knocking out this reaction provides larger amount of D-xylose to be catalyzed by the engineered enzyme, XR, providing more xylitol to the host. The second identified knockout target, R00086, is adenosine triphosphate (ATP) synthesis reaction. With this knockout, the host is unable to efficiently synthesize ATP via oxidative phosphorylation and is consequently forced to use more substrate xylitol for generating ATP via the less-efficient substrate-level phosphorylation, increasing total xylitol demand for equivalent biomass production. The third identified knockout target, R01518, is a reaction with reactant 3-phospho-D-glycerate (3pg) and product D-glycerate-2-phosphate (2pg). As shown in Figure 4, knocking out this reaction diverts flux towards D-ribulose-5-phosphate (ru5p-D), which is involved in the production of biomass precursors. Higher production rates of ru5p-D places demand for more substrate D-xylulose-5-phosphate (xu5p-D), which results in higher demand for xylitol, and thus a higher consumption flux.

Among the identified selection pathways, SP_{X1} is the shortest pathway with length one. This pathway provides a guaranteed minimum consumption flux equal to 111.66 mmol/gDCW/hr after applying the identified knockouts. Nair and Zhao [89] used the same selection pathway to engineer XR but only applied the first identified knockout target, R01432, to guarantee a minimum consumption flux. SelFi identified two additional knockout targets to improve the selection and more strongly link growth rate to engineered XR activity.

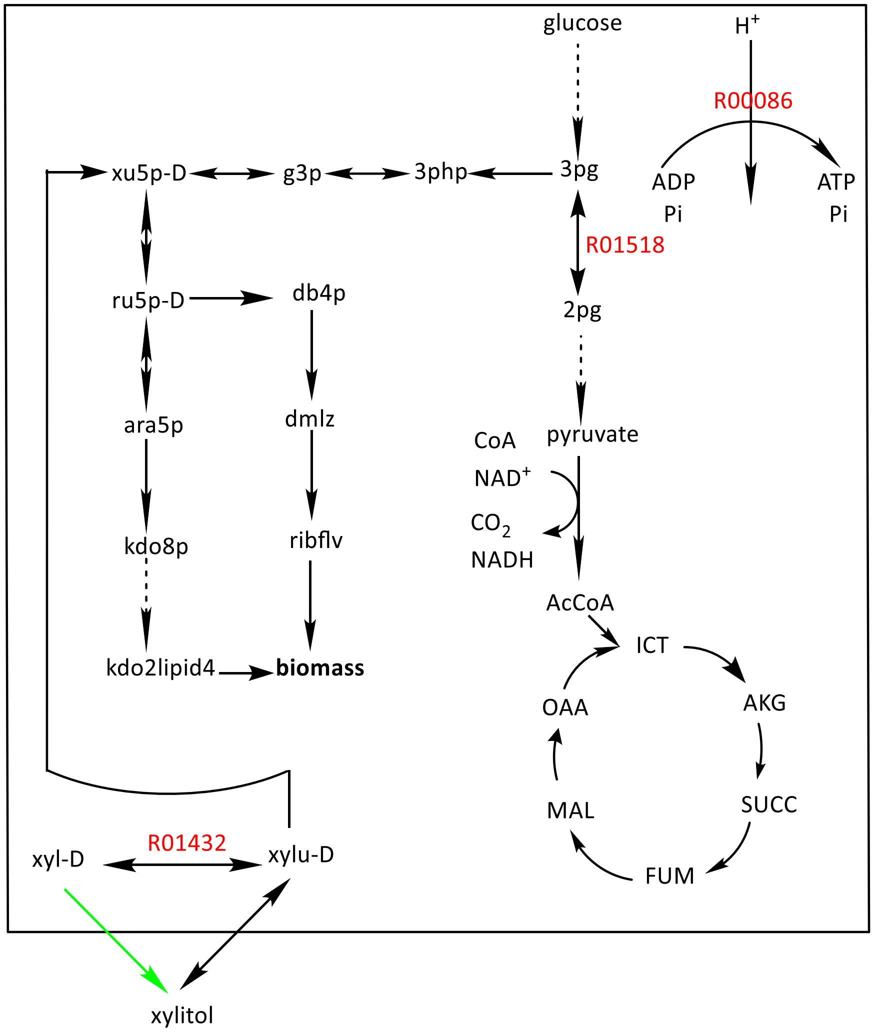


Figure 4. R01432, R00086 and R01518 are knockout targets (shown in red). The engineered enzymatic reaction is colored in green. Knockouts improve the production of xylitol from xyl-D through the engineered enzymatic reaction.

3.2.3 *Phosphoribulokinase (PRK) and D-Ribulose-1,5-bisphosphate*

Photosynthetic CO_2 fixation is a source of organic carbon and necessary for carbon sequestration. PRK catalyzes a reaction to establish the photosynthetic CO_2 fixation with D-ribulose-5-phosphate as a reactant and D-ribulose-1,5-bisphosphate as a product [90]. The reactant of this enzymatic reaction, D-ribulose-5-phosphate, is native to the host.

SelFi identified three consumption pathways as shown in Tables 1 and 2 for D-ribulose-1,5-bisphosphate. The consumption pathways end at D-ribose 1,5-bisphosphate, 3-phospho-D-glycerate, or 2-phosphoglycolate. SelFi restricts all external carbon sources except glycerol. Knockout targets that improve the minimum flux through the consumption pathways are summarized in Table 6. Table 7 presents the order of reactions to knock out and shows the extent to which each knockout improves the guaranteed minimum consumption flux of D-ribulose-1,5-bisphosphate. For selection pathway SP_{D1}, SelFi identified two knockout targets, while only one knockout target was identified for pathways SP_{D2} and SP_{D3}. Additional knockouts did not further improve the flux through the consumption pathways.

Table 6. Knockout targets for D-ribulose-1,5-bisphosphate test case

Knockout target reaction	**KEGG ID**
D-ribulose-5-phosphate[c] ↔ D-ribose-5-phosphate[c]	R01056
D-erythrose-4-phosphate[c] + D-xylulose-5-phosphate[c] ↔ D-fructose-6-phosphate[c] + glyceraldehyde-3-phosphate[c]	R01067

Table 7. Effect of knockouts to improve guaranteed minimum consumption fluxes. For each pathway listed in column 1, we identify knockout targets, and the corresponding minimum guaranteed flux in parenthesis.

Selection pathway label	1st Round	2nd Round	3rd Round
SP_{D1}	Δ R01056(2.37)	ΔR01056, Δ R01067(4.23)	---*
SP_{D2}	Δ R01056(3.78)	---*	---*
SP_{D3}	Δ R01056(3.55)	---*	---*

* No further flux-improving knockouts are identified

Red-labeled reactions in Figure 5 indicate knockout targets for the production of D-ribulose-1,5-bisphosphate through the engineered enzymatic reaction, which is colored in green. The first knockout target, R01056, is a reaction that consumes D-ribulose-5-phosphate (ru5p-D), the reactant of the enzymatic reaction. Knocking out this reaction provides larger amounts of ru5p-D for the engineered PRK leading to higher production of D-ribulose-1,5-bisphosphate. The second knockout target, R01067, similarly affects the host as R01067 is part of a pathway consuming ru5p-D.

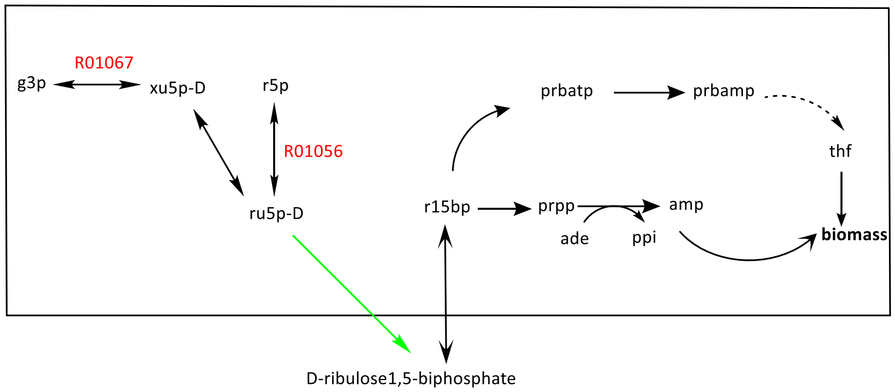


Figure 5. Knockout targets R01056 and R01067 are shown in red. Identified knockout targets are connected to D-ribulose-1,5-bisphosphate production. The engineered enzymatic reaction is colored in green. Knockouts improve the production of D-ribulose1,5-biphosphate from ru5p-D through the engineered enzymatic reaction.

The selection pathway SP_{D1} provides the highest guaranteed minimum flux, 4.23 mmol/gDCW/hr, after applying two knockouts. Among the identified pathways, the pathway from D-ribulose-1,5-bisphosphate to 3-phospho-D-glycerate was previously experimentally validated and confirmed in literature by Cai *et al.* [90], but used for selection of a different enzyme − RuBisCO. Since PRK and RuBisCO catalyze sequential reactions in CO_2 fixation, Cai *et al.* coupled both reactions to implement their selection.

3.2.4 *Methane Monooxygenase (MMO) and Methanol*

Conversion of methane to a liquid fuel such as methanol is highly desirable. While this conversion can be catalyzed by the enzyme MMO [91], Chen *et al.*

have shown that a cytochrome P450 oxidase can also be engineered to catalyze this reaction [92].

SelFi identified four consumption pathways for the methanol test case as shown in Tables 1 and 2. All identified pathways terminate at formaldehyde, but utilize different co-substrates and cofactors. Selection pathways SP_{M1} and SP_{M2} use NAD^+/NADH cofactors, while selection pathways SP_{M3} and SP_{M4} utilize H_2O_2 and O_2, respectively. The reactant of the enzymatic reaction, methane, is not native to the host. Upon restricting all external carbon sources except for methane, the identified consumption pathways became coupled with host survival enabling non-zero minimum consumption fluxes shown in Table 2. In this case, knockouts function solely to improve the non-zero minimum consumption fluxes. Table 8 shows reactions and their corresponding KEGG IDs as knockout targets. Table 9 shows the order of reactions to knock out as determined by SelFi, and the effect of each knockout has in improving the guaranteed minimum consumption fluxes.

Table 8. Knockout targets for methanol test case

Knockout Target Reaction	**KEGG ID**
CO_2[c] + H_2O[c] + phosphoenolpyruvate[c] → H^+[c] + oxaloacetate[c] + Pi[c]	R00345
4.0 H^+[c] + 0.5 O_2[c] + ubiquinol-8[c] → H_2O[c] + 4.0 H^+[p] + ubiquinone-8[c]	R09504
2.0 H^+[p] + NADH[c] + $NADP^+$[c] → 2.0 H^+[c] + NAD^+[c] + NADPH[c]	R00112

Table 9. Effect of knockouts to improve guaranteed minimum consumption fluxes. For each pathway listed in column 1, we identify knockout targets, and the corresponding minimum guaranteed flux in parenthesis.

Selection pathway label	**1st Round**	**2nd Round**	**3rd Round**
SP_{M1}	Δ R00345 (72.16)	Δ R00345, Δ R09504 (98.79)	Δ R00345, Δ R09504, Δ R00112 (117.50)
SP_{M2}	Δ R00345 (76.17)	Δ R00345, Δ R09504 (107.03)	Δ R00345, Δ R09504, Δ R00112 (127.29)
SP_{M3}	Δ R00112(196.18)	---*	---*
SP_{M4}	Δ R00112 (200.52)	Δ R00112, Δ R00345 (278.59)	---*

* No further flux-improving knockouts are identified

Figure 6 illustrates the knockout targets in red. The first knockout target, R00345, produces oxaloacetate, a metabolite involved in TCA cycle. The next identified knockout, R09504, is the cytochrome oxidase reaction, which is coupled with ATP synthesis. Knocking out cytochrome oxidase affects the functionality of ATP synthesis in the cell. The third identified knockout target, R00112, is $NAD(P)^+$ transhydrogenase reaction, recycling NAD^+, one of the cofactors involved in TCA cycle. These knockout targets increase the inefficiency in the TCA cycle and cause decreased ATP generation. Consequently, the host uses more substrate (methanol) through substrate-level phosphorylation to compensate for degraded ATP generation and meeting the minimal biomass production constraint. Among the selection pathways, SP_{M4} has the highest guaranteed minimum consumption flux, 278.59 mmol/gDCW/hr, after applying two knockouts.

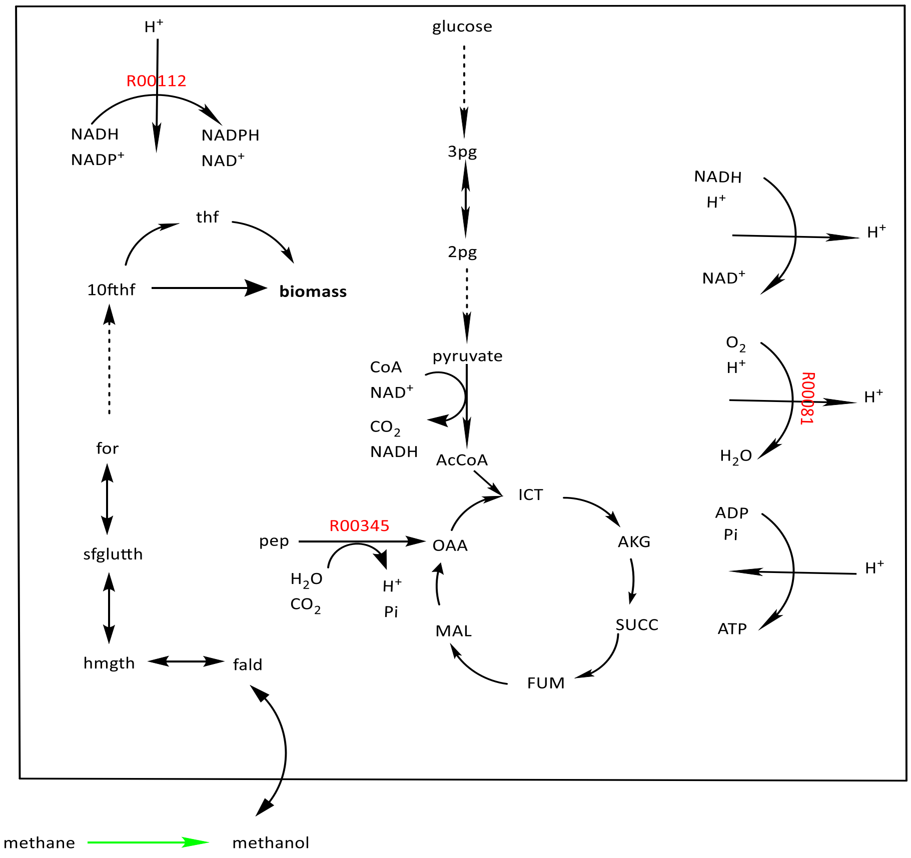


Figure 6. Knockout targets R00345, R09504, R00112 are shown in red. The engineered enzymatic reaction is colored in green. Identified knockouts improve the minimum production and consequently consumption of methanol through the engineered enzymatic reaction with methane as a precursor. Methane is not native to the host. All knockout targets affect the efficiency of the host in ATP generation.

3.2.5 *Aromatic Amino Acid Decarboxyloase (AADC) and Aniline*

Aniline is an important precursor for the production of industrial chemicals such as urethane polymers [93], for which there is currently no renewable source. We hypothesize that aniline can potentially be derived via a biosynthetic route from anthranilate, a native metabolite, using an engineered AADC.

For this test case, SelFi identified one consumption pathway ending at 4-aminobenzoate, as shown in Tables 1 and 2. SelFi restricted external carbon sources, leaving D-glucose as the only carbon source for the host. Before knockouts and as shown in Table 2, the maximum consumption flux through the identified pathway was low (0.01 mmol/gDCW/hr), while the minimum consumption flux is zero. For this selection pathway, SelFi identified one knockout target reaction, which is shown in Table 10 along with the corresponding KEGG ID. The effect of the identified knockout on improving minimum consumption flux through SP_A is shown in Table 11. The amount of guaranteed minimum consumption flux after applying the knockout is slightly improved (0.001 mmol/gDCW/hr), while the maximum consumption remains the same (0.01 mmol/gDCW/hr).

The results in Table 3 show the maximum demand of 4-aminobenzoate to produce biomass is equal to 0.05 mmol/gDCW/hr under all conditions. The low maximum demand for 4-aminobenzoate illustrates the minimal need for this compound for growth.

In Figure 7, the knockout target, R05553, is shown in red. R05553 is a reaction for the synthesis of 4-aminobenzoate (4abz), the end metabolite of the consumption pathway. Knocking out this reaction eliminates the only alternative pathway to produce 4-aminobenzoate, forcing the host to rely on the consumption pathway to produce aniline from anthranilate (anth).

Table 10. Knockout target for aniline test case

Knockout target reaction	KEGG ID
4-amino-4-deoxychorismate[c] → 4-aminobenzoate[c] + H^+[c] + pyruvate[c]	R05553

Table 11. Effect of knockout to improve guaranteed minimum consumption flux. For the pathway listed in column 1, we identify knockout targets, and the corresponding minimum guaranteed flux in parenthesis.

Selection pathway label	1st Round	2nd Round	3rd Round
SP_A	Δ R05553(0.001)	---*	---*

* No further flux-improving knockouts are identified

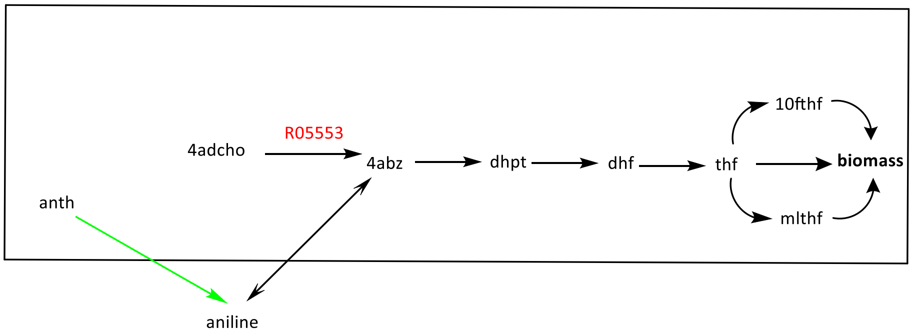


Figure 7. Knockout target R05553 is shown in red. The engineered enzymatic reaction is colored in green. The knockout target guarantees production of aniline from anth through the engineered enzymatic reaction.

3.3 Discussion and Conclusion

The framework presented in this chapter streamlines the process of identifying a cell-based high-throughput selection strategy for a desirable enzymatic reaction product. SelFi first identifies biochemical consumption pathways from the desired product towards the host. Next, SelFi links the consumption pathway with the cell growth and enhances the consumption flux by restricting carbon sources as well as identifying knockout targets in the host. In this work, SelFi identified up to three knockout targets using a greedy strategy to increase minimum selection flux.

We used SelFi to construct selection pathways for four enzymatic products. In the case of XR and xylitol, SelFi identified seven selection pathways, each with length ranging from one to three steps. The knockout targets were similar in each case, and all seven pathways attain comparable minimal yield after knockouts (110.27 mmol/gDCW/hr to 111.66 mmol/gDCW/hr). The single-step selection pathway and one of the identified knockouts were previously validated in the literature [89]. In the case of PRK and D-ribulose-1,5-bisphosphate, SelFi identified three single-step selection pathways. SelFi identified a common first knockout target amongst the three pathways, and one additional knockout target for one of the pathways. The pathway with two knockouts provided the highest guaranteed minimum flux (4.23 mmol/gDCW/hr compared to 3.78 mmol/gDCW/hr and 3.55 mmol/gDCW/hr), and was previously experimentally verified in the literature as a selection pathway for engineering RuBisCO, the enzyme catalyzing the rate-limiting step in CO_2 fixation, immediately downstream of PRK [90]. In the case of MMO and methanol, SelFi identified four single-step selection pathways with one to three knockout targets. The pathway with two knockouts provided over twofold higher minimum selection flux (278.59 mmol/gDCW/hr) compared to the selection pathways with three knockouts (117.50 mmol/gDCW/hr and 127.50 mmol/gDCW/hr), and higher minimum selection flux compared to the selection pathway with a single knockout (196.18 mmol/gDCW/hr). In the case of aniline, SelFi identified one single-step selection pathway ending in 4-aminobenzoate, with one knockout target. The knockout coupled the selection pathway with cell survival, but the resulting minimum selection flux (0.001 mmol/gDCW/hr) was low. This result is explained by the low maximum demand for 4-aminobenzoate in producing biomass under all conditions. Currently, there are no KEGG reactions that allow for creating a more effective selection pathway for aniline.

SelFi utilizes *ProPath*, a probabilistic traversal algorithm that was designed to find synthesis pathways from the host to a desired useful compound. SelFi uses a derivative algorithm, *retroProPath*, to find pathways initiating from the desired product and terminating in the host. Like *ProPath,* SelFi utilizes reactions only in the KEGG database to construct pathways, limiting the search space to metabolites and reactions present in KEGG. Using multiple databases would expand SelFi, making it applicable to a broader range of metabolites and enzymatic products. This in turn would concurrently increase the repertoire and diversity of identifiable consumption pathways. Other search algorithms for synthesis or degradation pathways such as PathPred [47] can be integrated with

SelFi to create selection pathways for molecules not present in databases such as KEGG.

SelFi aims to improve the guaranteed minimum selection flux while meeting a lower bound constraint on cell growth. A non-zero minimum consumption flux guarantees that the cell will utilize this pathway - a goal that cannot necessarily be met by maximizing the selection flux. This focus on minimum flux optimization differentiates SelFi from prior knockout identification works that aim to maximize target production rates. For example, techniques such Optknock [50], MOMAKnock [53], OptGene [55] and OptORF [52], OptReg [51], OptForce [57], CosMos [58], CCOpt [56], and RobustKnock [54] aim to increase target production via gene up/down over expression or knockout. Many approaches are mathematically elegant utilizing bi-level programing (e.g., Optknock, MOMAKnock) or identifying required coordinated changes among reactions (e.g., OptForce, CosMos), we selected a simple greedy knockout heuristic to guarantee minimum yield. This strategy is optimal in selecting each successive knockout, and is shown effective in identifying effective knockout strategies.

To couple a consumption pathway to host survival, SelFi restricts alternate carbon sources and identifies possible knockout targets. While this coupling method guarantees a minimum non-zero flux through the consumption pathway, the maximum flux through the identified pathway is dependent on the demand of the terminal host metabolite for cell growth. A viable consumption pathway must end at a metabolite with high-demand for biomass production. Alternatively, the cell must be engineered to change such demand. We developed in this thesis a new methodology to evaluate such demand. We showed that the maximum flux potential of selection pathways correlates with cellular demand of the metabolites at which the pathways terminate. In particular, the limited demand of end metabolite 4-aminobenzoate for biomass production (0.05 mmol/gDCW/hr, Table 2) explains the low flux rate for the aniline consumption pathway. Cellular engineering utilizing knockouts did not result in increased consumption flux as the knockouts aimed to increase the guaranteed minimum flux. Using screens may be more desirable in such cases.

The host model utilized by SelFi impacts the quantity and quality of the identified selection pathways. Anecdotes within the community show that models released in the public domain often have undocumented inconsistencies, such as dead-end metabolites or reactions incapable of carrying fluxes. Model and constraint consistency checkers such as MC^3 [94] can detect some issues such as singly connected metabolites, as was the case for 2-3-dioxo-L-gulonate where zero consumption demand was reported. There are other issues, however, that cannot be detected automatically. In the *i*AF1260 model, L-xylulose is listed as a native metabolite in *E. coli*. Although this metabolite was present in the *i*AF1260 model, L-xylulose cannot metabolize in *E. coli* [95]. To take this issue into account, we excluded L-xylulose as a native metabolite, thus preventing the generation of selection pathways that end at this metabolite for the xylitol test case, and allowing for pathways SP_{X1} - SP_{X7} that utilize L-xylulose as an intermediate (Table 2).

In summary, SelFi addresses a major bottleneck in directed evolution of enzymes. SelFi is the first automated methodology that detects the formation of a desired enzymatic product. The results of applying SelFi for engineering Xylose Reductase and RuBisCO showed agreement with previously experimentally validated selections. SelFi promises to expedite the design of high-throughput selections in directed evolution of enzymes.

Chapter 4
Using Biological Filtering and Substrate Promiscuity to Advance Annotation in Untargeted Metabolomics

We present in this chapter a novel annotation workflow for untargeted metabolomics. Measured masses from the sample are *filtered* through a relevant biological context to identify a biologically relevant set of candidate compounds. The central premise is that identifying a biologically relevant set of candidate metabolites that correspond to the features detected in an untargeted experiment leads to savings in annotation runtime without comprising the quality of results. This set is based on the enzymatic reactions expected to occur in the system of interest, and can be identified using an Expanded Metabolic Model (EMM) that includes metabolites resulting from promiscuous action of the enzymes, in addition to those associated with the biological sample through catalogued canonical substrates and products of enzymes. An EMM-based candidate set not only guarantees the biological relevance of the search space, but also takes the search space for metabolite annotation beyond the metabolites already cataloged as part of the sample. This expanded biological search space enhances the chance of identifying new metabolites during annotation.

4.1 Methods

4.1.1 *EMM-based Annotation* (EMMA)

Model-based filtering annotation workflow (Figure 8A) consists of filtering the masses of the measured metabolites against those expected in the sample based on its metabolic model. Metabolites from the model with masses that match, within a small error, those in the measured data are considered the candidate set. The candidate set is then processed using annotation tools and ranked against observed spectra. While there is now a growing collection of annotated genome sequences and tools for the reconstruction of metabolic models [96, 97], identifying candidate metabolites using only the sample's metabolic model as a reference set can be limiting. Current genome-scale models largely represent well-conserved metabolic pathways [98]. Further, although traditionally assumed to be specific, many enzymes, if not all, have promiscuous activities by acting on substrates other than those for which they were evolved to transform [37, 99, 100]. Additionally, some enzymes exhibit catalytic promiscuity at different active sites [101]. As a result, a given enzyme could catalyze the formation of more than one metabolic product. For example, a recent study found that about one-third of enzymes in a genome-scale model of *Escherichia coli* metabolism are responsible for two-thirds of the known nonspontaneous metabolic reactions [101]. Even when using genome-scale models, there is often only a small set of measured metabolites that match to the masses in the model, and the size of the candidate set is relatively small. Measured masses that have no correspondence in the model cannot be annotated using this workflow.

In contrast to using the metabolic model as a filter, selecting the candidate set based on a large database of compounds (e.g., ChemSpider [102], PubChem [36]) can potentially enhance annotation (Figure 8B). Such an annotation workflow first identifies potential candidate metabolites by querying one or more specified compound databases for all molecules whose exact masses match experimentally observed masses of the sample. These mass-matched metabolites form the candidate set, and are then ranked based on how well the predicted fragments match the observed MS/MS spectra. As annotation is fraught with uncertainty, in this workflow some measurements are annotated with biologically irrelevant identities when using large databases that include biological and non-biological data. We define a biological relevant candidate as a metabolite that can be a potential product of an enzymatic reaction in the metabolic model. The end user sifts through ranked candidate metabolites to select biologically relevant candidates. The manual examination of these metabolites is time-consuming and relies on explicit domain knowledge. The selection can be aided by including only those metabolites that are expected to be present in the sample using the metabolic model. However, re-applying the model-based filtering here results in annotation outcomes similar to those in the model-based filtering annotation workflow (Figure 8A). Importantly, applying this workflow to large chemical structure databases is unfortunately computationally prohibitive as processing time in current annotation tools is a function of the number of candidate metabolites. For example, MetFrag combinatorially enumerates possible fragments for each candidate metabolite [67]. CFM-ID creates a trained probabilistic generative model of the fragmentation process for the input list of candidate metabolites [70]. CSI-FingerID computes fingerprints for each candidate molecule and compares each to a derived fingerprint associated with the query spectra [71]. Not all the computational cost however is necessary. It is highly unlikely that every compound in candidate sets derived from large databases is biologically relevant. Using biologically relevant databases (e.g. KEGG) to derive the candidate set is attractive as the size of candidate sets is reduced when compared to those derived from larger databases. However, as there is no database that includes all biologically relevant compounds, there are many biologically relevant compounds that are not catalogued in specialized databases.

Our novel annotation workflow (Figure 8C), EMMA (EMM-based Annotation), improves on these two annotation workflows by applying an EMM-based filter to identify the candidate set. To create this model, we adopt a previously described method, *PROXIMAL* [103], which utilizes lookup tables of enzyme-catalyzed chemical transformation patterns to generate plausible reaction products for substrates of interest. From the reactant-product pair(s) (RPAIR) of an enzymatic reaction, *PROXIMAL* identifies a molecular pattern that transforms the reactant into product. Each pattern is associated with a reaction center [104] and its second-level neighboring atoms. If a substrate of interest matches a pattern, then the corresponding operator is applied to generate a product, which we call a “derivative” metabolite. The main advantage of using *PROXIMAL* is its ability to generate a set of operators that reflect the chemical transformation capabilities of the enzymes specific to a biological system of interest as defined by the system’s genome (or genomes). The EMM is generated using *PROXIMAL* by applying the operators generated from the enzymatic reactions encoded in a biological system to the metabolites cataloged for the system. This generates a set of “derivative” metabolites. The calculated exact masses of derivative metabolites are used to filter the measured accurate masses. If a derivative has a mass that matches a measured mass, then the KCF or SMILES string of this derivative is searched against a chemical structure database to determine if it has been cataloged with a chemical name and identifier. The calculated masses of metabolites in the base model are also matched against the measured masses (as in Figure 8A). The union of matched derivatives and model metabolites constitute a set of compounds that could be present in the sample due to enzymatic activity, and are deemed biologically relevant candidate identities for the detected MS features. These candidate metabolites are then evaluated using *in silico* fragmentation analysis, where the measured MS/MS spectra of the mass-matched features are compared against the predicted spectra of the candidate compounds.

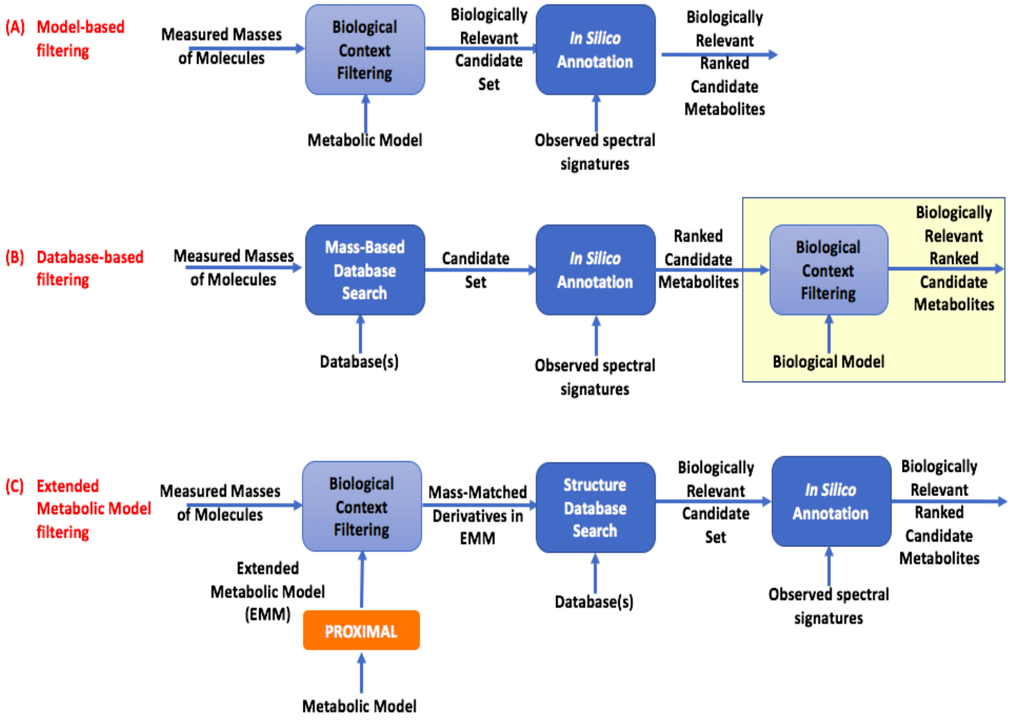

(A) Model-based filtering
Measured Masses of Molecules
Biological Context Filtering
Metabolic Model
Biologically Relevant Candidate Set
In Silico Annotation
Observed spectral signatures
Biologically Relevant Ranked Candidate Metabolites
(B) Database-based filtering
Measured Masses of Molecules
Mass-Based Database Search
Database(s)
Candidate Set
In Silico Annotation
Observed spectral signatures
Ranked Candidate Metabolites
Biological Context Filtering
Biological Model
Biologically Relevant Ranked Candidate Metabolites
(C) Extended Metabolic Model filtering
Measured Masses of Molecules
Biological Context Filtering
Extended Metabolic Model (EMM)
PROXIMAL
Metabolic Model
Mass-Matched Derivatives in EMM
Structure Database Search
Database(s)
Biologically Relevant Candidate Set
In Silico Annotation
Observed spectral signatures
Biologically Relevant Ranked Candidate Metabolites

Figure 8. Comparison between annotation workflows. The candidate set for annotation is derived by filtering the measured masses based on: (A) the metabolic model, (B) databases, and (C) extended metabolic model (EMM). The candidate sets in (A) and (C) are biologically relevant, while the ones in (B) may not all be biologically relevant.

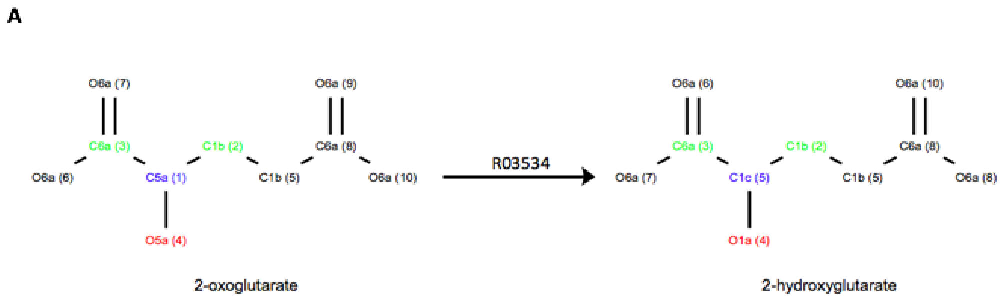
4.1.2 *Identifying biologically relevant molecules beyond those in the metabolic model*

The sample's metabolic model can be augmented into an expanded metabolic model based on enzyme promiscuity. To this end, we generalized the pattern matching method described in our earlier work, *PROXIMAL*, which was originally developed for identifying possible bio-transformation products of xenobiotic chemicals in the liver due to Cytochrome P450 (CYP) enzymes. The key idea in *PROXIMAL* [103] is to approximate enzyme activities through bio-transformation operators that act on molecular fragments. To expand the metabolic model, each bio-transformation operator is applied to each metabolite within the model.

The bio-transformation operators are constructed as follows. The transformation of each fragment is be specified using Reaction Center, Difference Region, and Matched Region (RDM) patterns [104]. The RDM patterns of metabolic enzymes are available from the KEGG reaction pair (RPAIR) database [104], and specify local regions of similarities/differences for reactant-product pairs based on chemical structure [48]. An RDM pattern consists of three parts: a Reaction Center (R) atom that exists in both the substrate and reactant molecule on the boundary between Matched and Non-Matched Regions, Difference Region (D) atoms that are adjacent to the R atom but also part of the Non-Matched Region, and Matched Region (M) atoms adjacent to the R atom in the Matched Region. A lookup table is constructed based on the RDM patterns of enzymes associated with reactions in the model. The "key" in the lookup table consists of the atom types of the R and M and adjacent neighbors in the reactant, while the "value" represents the atom types of the R and D in the product. For each potential R pattern matched in the query molecule, a set of transformations are looked up in the table and applied to the query molecule.

To illustrate how *PROXIMAL* functions, an example is shown in Figure 9. In Figure 9A, a specific reversible reaction (KEGG reaction ID: R03534) transforms 2-oxoglutarate (KEGG compound ID: C00026) to 2-hydroxyglutarate (KEGG compound ID: C02630). The reactant and product molecules are encoded using KEGG atom types [48], while the atom numbers, extracted from KEGG KCF files, are specified in parenthesis following the type of atoms in the structure of each compound. Each reactant-product atom pair is then entered into a transformation table (Figure 9B). The transformation table identifies patterns of change in atom types along with a local context through the transformation of reactant to product. To identify transformation patterns, *PROXIMAL* aligns the atoms in reactant-product structures, and adds each atom in the reactant and its corresponding atom in the product as a new row to the transformation table. The ordering of the rows in the table is determined by the ordering of atoms in the reactant molecule structure (Figure 9B). Having the transformation table, any reactant atom, which is aligned to a product atom with a different type will be considered as a potential reaction center. In this example, rows 1 and 4 demonstrate two potential reaction centers in reactant compound: C5a and O5a. To add specificity to these transformations, the lookup table keys are augmented to include two-level nearest neighbors including the reaction center (Figure 9C). To visualize the concept of two-level nearest neighbors, we used a color code in Figure 9A illustrating this concept for one of the potential reaction centers, O5a**.** The potential reaction center O5a is shown in red. The first-level neighbor (adjacent neighbor) C5a is shown in blue, and the second-level neighbors (distant neighbors) C1b and C6a are shown in green. The same biotransformation can be derived by multiple reactions cataloged in KEGG. For this specific example, reactions with KEGG IDs R00267, R00342, R00709, R01000, R01388, R01392, R01394, R01513, R03104, R03688, and R07136 can lead to the same bio-transformation pattern. Similarly, the set of adjacent and distant neighbor atoms for the potential reaction center C5a can be extracted (Figure 9C). The set of distant neighbors always include the reaction center.

Given a query compound, *PROXIMAL* applies a select set of transformations from the lookup tables at one or more matching sites, or reaction centers, of the query compound, where several derivatives are possible (Figure 10). Considering each atom in the query molecule as a potential reaction center, *PROXIMAL* creates a neighbors table containing a list of adjacent and distant neighbors for each of the potential reaction centers. *PROXIMAL* then looks for matches between the generated list and keys in the lookup table. In case of a match, *PROXIMAL* applies the matched key’s value to the reaction center and its neighbors to generate a product. Query compound 4-hydroxyphenylpyruvate (KEGG compound ID: C01179) is demonstrated with atom types in Figure 10A. For each atom in the structure of the query compound, a list of adjacent and distant neighbors is generated and added to neighbors table (Figure 10B). Comparing the neighbors table against the keys in the lookup table (Figure 9C) shows row 4 of the neighbor table, with potential reaction center O5a, as a match. Application of the value found corresponding to the matched key to the reaction center and its neighbors leads to a biotransformation product 4-hydroxyphenyllactate with KEGG compound ID: C03672 (Figure 10C).



B

Reactant		Product	
Atom #	Atom type	Atom #	Atom type
1	C5a	1	C1c
2	C1b	2	C1b
3	C6a	3	C6a
4	O5a	4	O1a
5	C1b	5	C1b
6	O6a	7	O6a
7	O6a	6	O6a
8	C6a	8	C6a
9	O6a	10	O6a
10	O6a	9	O6a

C

Keys

Reaction center	Adjacent neighbor 1	Distant neighbors 1	Adjacent neighbor 2	Distant neighbors 2	Adjacent neighbor 3	Distant neighbors 3
C5a	O5a	C5a	C6a	C5a, O6a, O6a	C1b	C5a, C1b
O5a	C5a	O5a, C6a, C1b	--- *	---	---	---

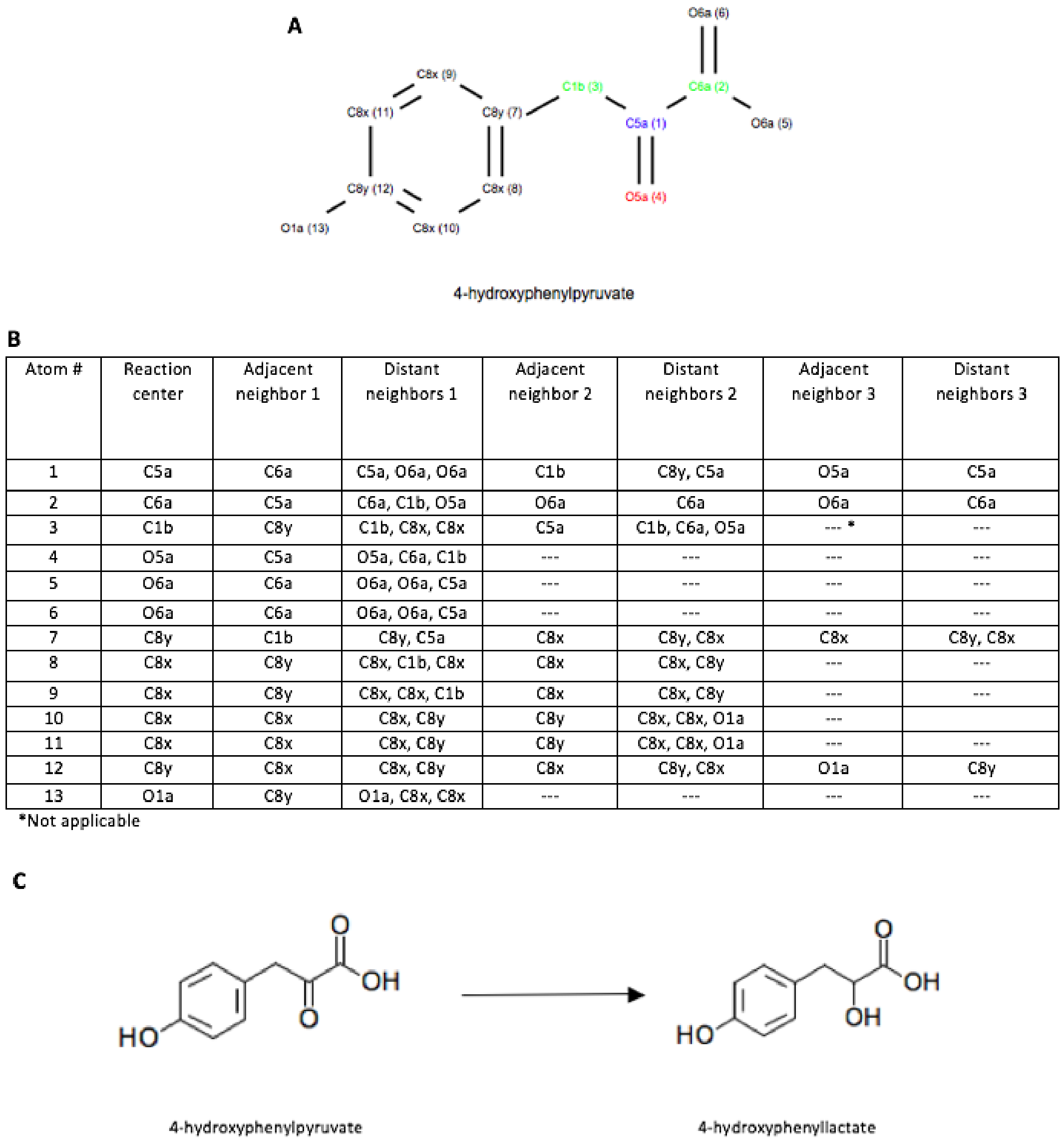
* Not applicable

Values

Reaction center	Adjacent neighbor	Added functional group
C1c	O1a, C6a, C1b	--- *
O1a	C1c	---

* Not applicable

Figure 9. Illustration of generating lookup tables by *PROXIMAL*. (A) Reactant and product of an enzymatic reaction R03534, for which *PROXIMAL* aims to derive possible corresponding bio-transformations (operators). (B) Transformation table containing matching atom pairs in reactant and product compounds. (C) Potential operators: key table specifies the transformed substructure in reactant. Value table specifies the modification in product corresponding to the content of key table.



Atom #	Reaction center	Adjacent neighbor 1	Distant neighbors 1	Adjacent neighbor 2	Distant neighbors 2	Adjacent neighbor 3	Distant neighbors 3
1	C5a	C6a	C5a, O6a, O6a	C1b	C8y, C5a	O5a	C5a
2	C6a	C5a	C6a, C1b, O5a	O6a	C6a	O6a	C6a
3	C1b	C8y	C1b, C8x, C8x	C5a	C1b, C6a, O5a	--- *	---
4	O5a	C5a	O5a, C6a, C1b	---	---	---	---
5	O6a	C6a	O6a, O6a, C5a	---	---	---	---
6	O6a	C6a	O6a, O6a, C5a	---	---	---	---
7	C8y	C1b	C8y, C5a	C8x	C8y, C8x	C8x	C8y, C8x
8	C8x	C8y	C8x, C1b, C8x	C8x	C8x, C8y	---	---
9	C8x	C8y	C8x, C8x, C1b	C8x	C8x, C8y	---	---
10	C8x	C8x	C8x, C8y	C8y	C8x, C8x, O1a	---	
11	C8x	C8x	C8x, C8y	C8y	C8x, C8x, O1a	---	---
12	C8y	C8x	C8x, C8y	C8x	C8y, C8x	O1a	C8y
13	O1a	C8y	O1a, C8x, C8x	---	---	---	---

*Not applicable

Figure 10. Illustration of application of lookup table to a query molecule by *PROXIMAL* to generate the potential bio-transformation products. (A) A query compound represented by KEGG atom types. (B) Table of neighbors generated considering each atom in the query compound as a potential reaction center. Row 4 in the generated table matches to one of the keys in the lookup table shown in Figure 9C. (C) The product 4-hydroxyphenyllactate is the result product of applying the matched key's value to the query compound.

To create an EMM given a reference catalogued metabolic model, one or more operators are derived from substrate-product pairs associated with each reaction. Operators are then applied to all metabolites within the model. The expanded model size depends on the number of operators and metabolites of the reference model.

4.1.3 *Details of the EMMA Annotation workflow*

Given a model (list of metabolites and reactions) as well as tandem MS data (mass measurements of parent molecules and associated spectral signatures) for a biological sample, the goal is to associate each mass measurement with a compound ID. The workflow of EMMA, Figure 8C, is outlined in Figure 11.

In step 1, *PROXIMAL* is used to create transformation lookup tables based on enzymatic reactions in the input model. In step 2a, the biotransformation information stored in the created lookup tables is applied to model metabolites to generate a set of potential derivatives in EMM. In step 2b, the monoisotopic masses of atoms are used to calculate the mass of each potential derivative. In step 2c, the calculated masses are compared within the specified error margin against measured masses to generate a list of mass-matched derivatives in EMM. In step 3, the mass-matched derivatives in EMM are structurally compared against compound databases to add structurally-matched metabolites to the list of biologically relevant candidate set. In step 4, biologically relevant candidate set metabolites are scored and ranked against the observed spectral signatures using *in silico* fragmentation leading to generate biologically relevant ranked candidate metabolites. We chose to use 10 PPM mass error margin in the implementation of the EMMA workflow. We used CFM-ID [70] as the fragmentation prediction tool for scoring the candidate metabolites.

EMMA workflow

```
Procedure EMMA (in metabolic model, in measured masses of molecules, in observed Spectral
signatures, in database(s), out biologically relevant ranked candidate metabolites)

Begin
        1. use model reactions in metabolic model to generate biotransformation lookup tables
        2. identify mass-matched derivatives in extended metabolic model (EMM)
        for each metabolite in metabolic model
                2a. apply biotransformation lookup tables on metabolite to generate potential
                derivatives
                for each derivative in potential derivatives
                        2b. calculate, M, the mass of derivative
                        for each mass measurement m in measured masses of molecules
                                2c. use an error margin to generate a mass interval
                                        if M falls into mass interval
                                                add derivative to mass-matched derivatives in EMM
                                        end if
                        end for
                end for
        end for

        3. compare mass-matched derivatives in EMM to database(s), add the ones that match
        structurally to a metabolite in a database into biologically relevant candidate set

        4. use an in silico fragmentation tool to score biologically relevant candidate set against
        observed spectral signatures and output biologically relevant ranked candidate metabolites
end
```

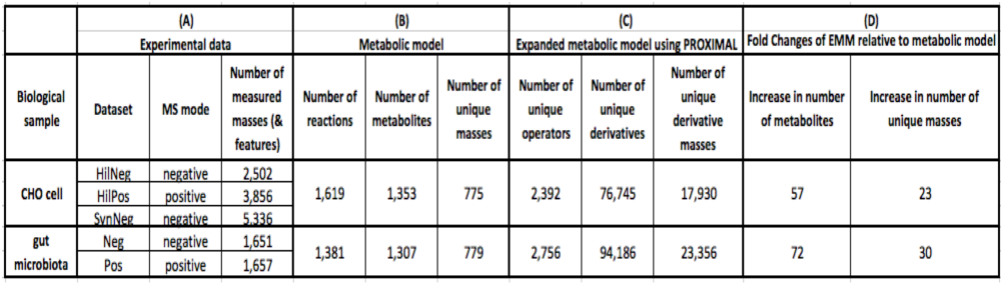
Figure 11. Pseudo code of the EMMA workflow

4.2 Results

4.2.1 *Datasets and models*

We compared the EMMA workflow with the other workflows shown in Figure 8 by analyzing untargeted LC-MS data collected on samples from two different biological systems (Table 12, column group A). One set of LC-MS experiments were performed on samples from Chinese hamster ovary (CHO) cell cultures grown in chemically defined media. The second set of experiments was performed on samples from anaerobic cultures of murine cecal isolates. To gain a better coverage in measured metabolites by MS, each set of LC-MS experiments comprised two or more MS methods, and the resulting datasets are treated independently. The cell culture and LC-MS experiments are described in [105]. The processed data were arranged into feature tables, where each feature was specified by a chromatographic retention time (RT), measured mass (m/z), and a set of associated product ion (fragment) masses and their relative intensities, i.e., MS/MS spectrum. The metabolic models for CHO cells and murine cecum microbiota were derived from genomes in the KEGG database. For the CHO cell, we obtained a listing of metabolites and reactions associated with the organism code *cge* in KEGG. The cecal culture is a consortium of many species. We used a community-level model that is assembled based on the taxonomic groups detected in the culture using a previously described procedure [106]. The numbers of reactions, metabolites and unique masses included in the two models and their corresponding EMMs are listed in Table 12 (column groups B and C). The EMM for a model includes additional metabolites that are not part of the reaction definitions for the cataloged enzymes in the model, but could result from a chemical transformation catalyzed by one or more enzymes in the system. This substantially increases the number of candidate metabolites by 57- and 72-fold for the CHO cell and microbiota models, respectively. Consequently, the number of unique masses also increase by 23- and 30-fold for the CHO cell and microbiota models, respectively (Table 12, column group D).

Table 12. Size of experimental data sets and models. (A) Three experimental datasets under different conditions were collected for the CHO cell, and two for the gut microbiota sample. (B) The size of the metabolic model in terms of number of, reactions, metabolites, and unique masses. (C) The size of the expanded metabolic model in terms of number of operators derived from *PROXIMAL*, unique derivatives generated by *PROXIMAL*, unique derivative masses due to *PROXIMAL*. For comparison purposes, the numbers of derivatives and derivative masses exclude those in the metabolic model. (D) Fold increase in number of metabolites and masses when comparing the size of these sets for EMM against the metabolic model.



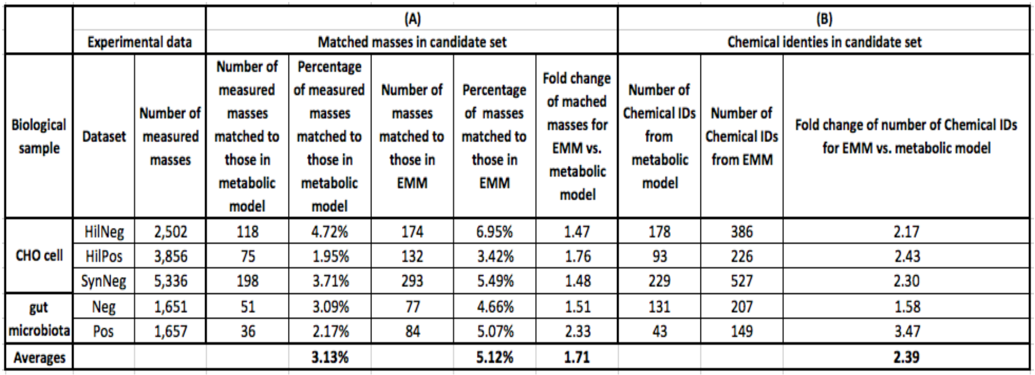
	(A) Experimental data			(B) Metabolic model			(C) Expanded metabolic model using PROXIMAL			(D) Fold Changes of EMM relative to metabolic model	
Biological sample	Dataset	MS mode	Number of measured masses (& features)	Number of reactions	Number of metabolites	Number of unique masses	Number of unique operators	Number of unique derivatives	Number of unique derivative masses	Increase in number of metabolites	Increase in number of unique masses
CHO cell	HilNeg	negative	2,502	1,619	1,353	775	2,392	76,745	17,930	57	23
	HilPos	positive	3,856								
	SynNeg	negative	5,336								
gut microbiota	Neg	negative	1,651	1,381	1,307	779	2,756	94,186	23,356	72	30
	Pos	positive	1,657								

4.2.2 *EMMA increased annotation opportunities when compared to metabolic model-based workflows*

Compared to a metabolic model for a biological sample, using EMM as the search space for metabolite annotation increases the size of the candidate set in terms of: (a) matching to a larger number of masses among the measured masses, and (b) suggesting a wider range of chemical identities. We therefore compare the size of the “biologically relevant candidate sets” in the model-based workflow and the EMMA workflow in Figure 8 using the increase in number of masses and in chemical identifies as metrics.

Using EMMs increases the size of the candidate set when compared to using the metabolic model. When using the metabolic model, a very small percentage of the measured metabolites are matched to the metabolic model. On average, 3.31% of measured masses can be annotated using the metabolic model only. This number increases for EMMA. On average, 5.12% of all measured masses can be annotated using the EMM, offering a 1.71-fold increase in the number of masses that can be annotated (Table 13, column group A). Unique masses in the candidate set when using the metabolic model correspond to compounds in the metabolic model. The number of such compounds varied from 43 to 229 across the data sets. When using the EMM, the number of chemical identities available for annotation ranged from 149 to 527 identities. There is therefore an average fold increase of 2.39 across all dataset in number of chemical identities that can be used for annotation when using EMMs (Table 13, column group B).

Table 13. Improvement due to EMMA workflow over using the metabolic model in terms of number of masses that match against the measured masses, and number of chemical identities in the candidate sets. (A) Number of masses in the candidate set for the metabolic model, the equivalent percentage in reference to the number of measured masses, number of masses in the candidate set for EMMA, and the equivalent percentage in reference to the number of measured masses, and the fold increase in number of masses. (B) Number of metabolites identified as the candidate set for metabolic model, number of metabolites identified as candidate set by EMMA, and the fold increase in number of metabolites in candidate sets.



			(A)					(B)		
	Experimental data		Matched masses in candidate set					Chemical identies in candidate set		
Biological sample	Dataset	Number of measured masses	Number of measured masses matched to those in metabolic model	Percentage of measured masses matched to those in metabolic model	Number of masses matched to those in EMM	Percentage of masses matched to those in EMM	Fold change of mached masses for EMM vs. metabolic model	Number of Chemical IDs from metabolic model	Number of Chemical IDs from EMM	Fold change of number of Chemical IDs for EMM vs. metabolic model
CHO cell	HilNeg	2,502	118	4.72%	174	6.95%	1.47	178	386	2.17
	HilPos	3,856	75	1.95%	132	3.42%	1.76	93	226	2.43
	SynNeg	5,336	198	3.71%	293	5.49%	1.48	229	527	2.30
gut microbiota	Neg	1,651	51	3.09%	77	4.66%	1.51	131	207	1.58
	Pos	1,657	36	2.17%	84	5.07%	2.33	43	149	3.47
Averages				**3.13%**		**5.12%**	**1.71**			**2.39**

4.2.3 *EMMA workflow reduces the annotation search space compared to large databases*

Current annotation methods using *in silico* fragmentation analysis require users to select a chemical database that can be searched for candidate compounds. Selecting a biological database such as KEGG database increases the likelihood that the candidate compounds are metabolites having an enzymatic origin. Further, selecting a smaller database also reduces computation time. For example, the number of compounds in PubChem, which includes both biological and non-biological compounds, is approximately 18 times larger than KEGG, and hence would require an 18-fold increase in computation time to analyze. The KEGG database, which includes metabolites from different kingdoms of life, is approximately 14 times larger than the number of compounds in the base models for the CHO cell and microbiota cultures (Figure 12).

The tradeoff for analyzing a smaller set of candidate compounds is that this can limit the potential for discovery. EMMA addresses this limitation by using chemical transformation operators derived from patterns identified for the enzymes specific to the system of interest. Applying *PROXIMAL* to the base models generates EMMs that are approximately 16 times larger in terms of number of compounds, but significantly smaller than databases KEGG and PubChem or their EMMs in which not all the compounds are biologically relevant (Figure 12). Expanding KEGG using *PROXIMAL* operators (11,091 operators derived from all reactions in KEGG applied to 21,270 compounds) expands the search space by at most 1.71 orders of magnitude compared to EMM (Figure 12). This expansion is akin to using a large derivative database such as MINEs [38] or MyCompoundID [39], which list 571,000 and 375,809 derivatives, respectively.

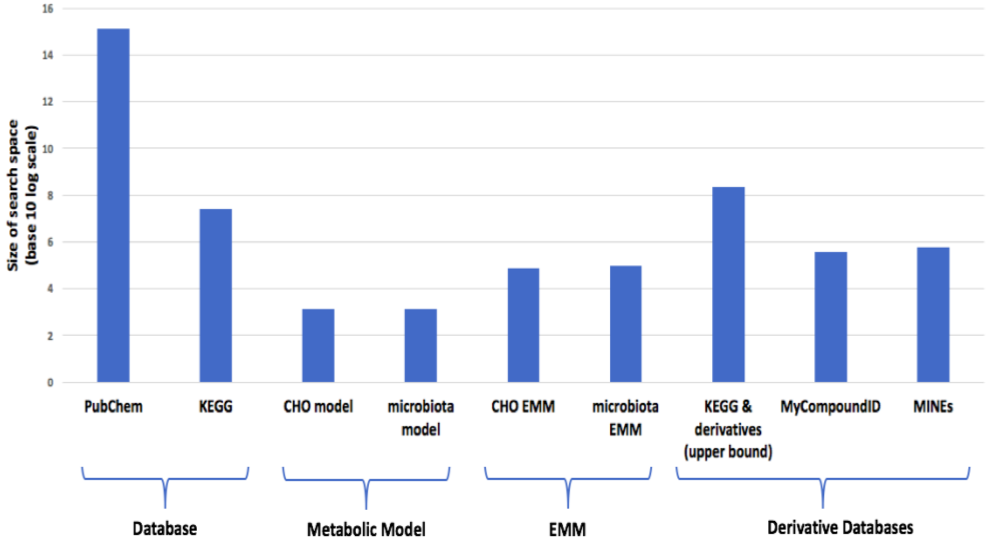


Figure 12. Comparison of the number of compounds in a large database (PubChem), a biologically relevant database (KEGG), metabolic model, expanded metabolic models, and expanded biological database.

By reducing the number of the candidate compounds, EMMA also reduces the number of candidate compounds with calculated exact masses that match the measured accurate masses of detected compounds. The distribution of exact masses when using KEGG and PubChem databases as the search space (Figure 13**,** histograms in red) resembles a long right-tailed distribution, a distribution that is in line with that for masses in these databases [107]. The masses from KEGG and PubChem filtered using EMMA (Figure 13**,** histograms in blue**)** shows a trend where there is a higher number of candidates for lighter masses. Overall, Figure 13 demonstrates the ability of EMMA to significantly reduce the number of candidate metabolites per measured mass to a biologically relevant set.

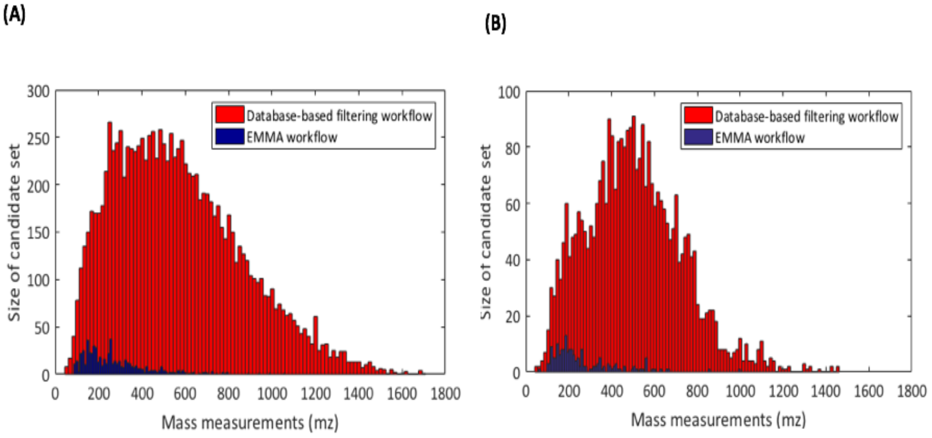
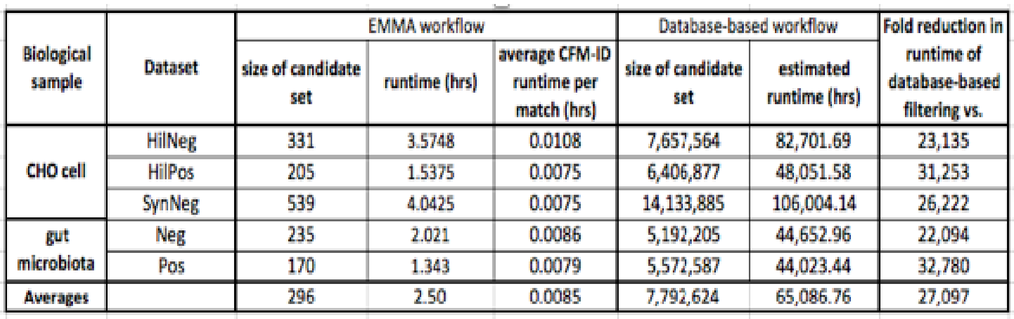


Figure 13. Distribution of masses in the candidate sets obtained through the database filtering workflow (red) and the EMMA workflow (blue) when searching KEGG and PubChem for the (A) CHO cell and (B) the gut microbiota.

To assess the computational savings of the EMMA workflow, we compared the time to perform an *in silico* annotation analysis in database-based workflow (Figure 8B) and EMMA (Figure 8C), for the LC-MS datasets. Using CFM-ID for *in silico* fragmentation and annotation of candidate sets identified using EMMA, we timed CFM-ID when scoring the candidates. For each dataset, the average runtime per match was recorded, and ranged from 0.1080 to 0.0075 hours (Table 14**)**. For all the runs, we used the same Windows machine with an Intel(R) Xeon(R) CPU E5-1620v2 processor, running at 3.70 GHZ, with 8 GIG RAM and 1 TB total memory.

To generate the candidate set as the input to *in silico* annotation analysis in database-based workflow (Figure 8B), we identified metabolites in the KEGG and PubChem databases that mass-matched to the masses in our experimental data for each dataset, within a 10 ppm error margin. It was computationally prohibitive to fragment all mass-matched metabolites from PubChem and KEGG (Table 14**)**. Instead, we estimated the runtime required by CFM-ID to *in silico* fragment each match by averaging the runtime of CFM-ID for EMMA through all datasets. Dividing the runtime by number of metabolites in the candidate set, on average, *in silico* fragmentation requires 0.0085 hours per match. Using this average, the estimated runtime for in silico fragmentation of database-based workflow is computed for each dataset. The average reduction in fragmentation time was 27,096x (Table 14).

Table 14. Computational speed up of EMMA workflow over database-based workflow for our datasets. For the EMMA workflow, the following data is provided: candidate set size generated by EMMA, relevant CFM-ID runtime for EMMA to perform annotation on the candidate set, and average runtime per match. For the database-based workflow, the size of the candidate and the estimated run time of CFM-ID is provided. The final column records the fold reduction in annotation runtime when comparing the database-based and EMMA workflows.

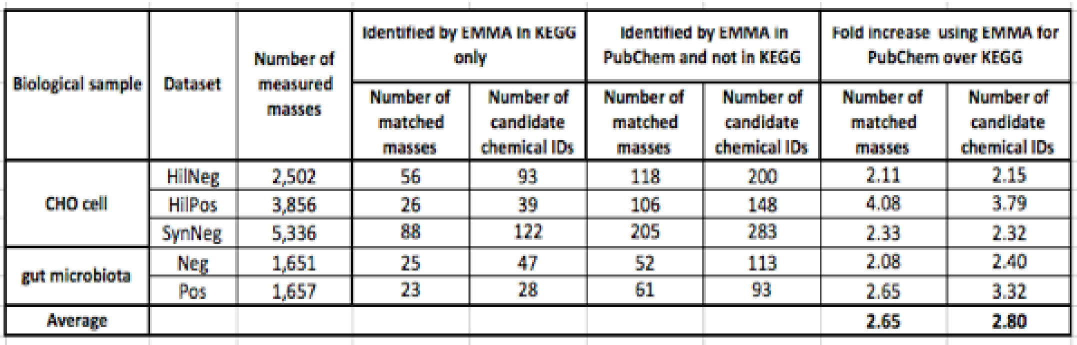


Biological sample	Dataset	EMMA workflow			Database-based workflow		Fold reduction in runtime of database-based filtering vs.
		size of candidate set	runtime (hrs)	average CFM-ID runtime per match (hrs)	size of candidate set	estimated runtime (hrs)	
CHO cell	HilNeg	331	3.5748	0.0108	7,657,564	82,701.69	23,135
	HilPos	205	1.5375	0.0075	6,406,877	48,051.58	31,253
	SynNeg	539	4.0425	0.0075	14,133,885	106,004.14	26,222
gut microbiota	Neg	235	2.021	0.0086	5,192,205	44,652.96	22,094
	Pos	170	1.343	0.0079	5,572,587	44,023.44	32,780
Averages		296	2.50	0.0085	7,792,624	65,086.76	27,097

While the speed up was assessed for CFM-ID, similar speedups are expected for spectral database lookups and other in silico fragmentation tools as the speed up directly correlates with the reduction in candidate metabolites.

Using a biological database such as KEGG for annotation guarantees biological relevance of candidate metabolites, and it may not be as computationally prohibitive (as was just shown) as it is to search a larger, general database that include non-biological compounds. A question that often arises is regarding the benefits of utilizing non-biological databases for annotation compared to when employing a biologically relevant database. Using EMMA, we are able to explore and quantify the benefits. Specifically, we utilized the EMMA workflow for our datasets once with the KEGG database, and once again with PubChem (Table 15). Using KEGG, we report the number of masses and metabolites in the candidate set. We report the numbers using PubChem excluding masses and metabolites already placed in the candidate set when using KEGG. For CHO cell and gut microbiota datasets, EMMA increases the number of annotated masses and biologically relevant candidate molecules by approximately 2.65- and 2.8-fold, respectively, when employing PubChem as the

Table 15. Capability of EMMA in expanding the search space for annotation. For each experimental dataset, the following is calculated. Number of metabolites in the candidate set identified by EMMA in KEGG, number of metabolites in the candidate set identified by EMMA in PubChem but not in KEGG, and fold change increase of number of metabolites in the candidate set resulted by employing PubChem as the search space vs. KEGG.



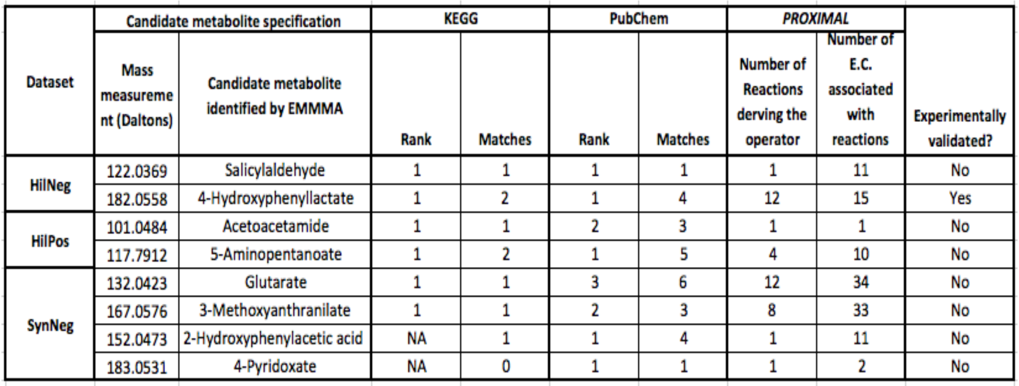
Biological sample	Dataset	Number of measured masses	Identified by EMMA in KEGG only		Identified by EMMA in PubChem and not in KEGG		Fold increase using EMMA for PubChem over KEGG	
			Number of matched masses	Number of candidate chemical IDs	Number of matched masses	Number of candidate chemical IDs	Number of matched masses	Number of candidate chemical IDs
CHO cell	HilNeg	2,502	56	93	118	200	2.11	2.15
	HilPos	3,856	26	39	106	148	4.08	3.79
	SynNeg	5,336	88	122	205	283	2.33	2.32
gut microbiota	Neg	1,651	25	47	52	113	2.08	2.40
	Pos	1,657	23	28	61	93	2.65	3.32
Average							**2.65**	**2.80**

search space beyond KEGG (Table 15).

4.2.4 *Experimental validation of EMMA*

We next investigated whether any of the derivatives predicted by EMM and matched to a detected MS feature based on mass and MS/MS spectrum could be experimentally confirmed with a chemical standard. To this end, we selected eight predicted derivatives that had a match in the LC-MS feature tables for CHO cell samples (Table 16). The selection was based on two factors: the rank assigned by the *in silico* fragmentation tool and availability from a vendor. The selected derivatives are: Salicylaldehyde, one of the three isomers of hydroxybenzaldehyde; 4-Hydroxyphenyllactate, a tyrosine metabolite; Acetoacetamide, a monocarboxylic acid amide of acetoacetic acid; 5-Aminopentanoate, a lysine degradation product; Glutarate, produced in lysine and tryptophan metabolism; 3-Methoxyanthranilate, an ester of anthranilic acid; 2-Hydroxyphenylacetic acid, associated with styrene degradation pathway; and 4-Pyridoxate, a product of vitamin B_6. Almost all derivatives were identified as one of the top three candidates across two databases searches (KEGG and PubChem). In some cases, the derivative was the only match in a particular database (e.g, Salicylaldehyde in KEGG). In other cases, the derivative was one of several possible matches (e.g., 4-Hydroxyphenyllactate was one of 4 matches in PubChem). For the *PROXIMAL* operators, the number of their relevant reactions and enzymes varied. With compound 4-Hydroxyphenyllactate, the associated operator is derived from 12 enzymatic reactions, which are catalyzed by 15 enzymes in CHO cell. Comparing the RTs and MS/MS spectra of standards for these chemicals against the corresponding CHO cell culture sample features [105], we were able to confirm correct annotation of 4-Hydroxyphenyllactate (Figure 14), demonstrating that the EMM for the CHO cell can indeed predict the

Table 16. Experimental validation of EMMA. Eight metabolites identified by EMMA and highly ranked using annotation are selected for experimental validation. The ranking of each metabolite and the number of candidates that matched this metabolite when using KEGG and PubChem is reported.



Dataset	Candidate metabolite specification		KEGG		PubChem		*PROXIMAL*		Experimentally validated?
	Mass measurement (Daltons)	Candidate metabolite identified by EMMMA	Rank	Matches	Rank	Matches	Number of Reactions derving the operator	Number of E.C. associated with reactions	
HilNeg	122.0369	Salicylaldehyde	1	1	1	1	1	11	No
	182.0558	4-Hydroxyphenyllactate	1	2	1	4	12	15	Yes
HilPos	101.0484	Acetoacetamide	1	1	2	3	1	1	No
	117.7912	5-Aminopentanoate	1	2	1	5	4	10	No
SynNeg	132.0423	Glutarate	1	1	3	6	12	34	No
	167.0576	3-Methoxyanthranilate	1	1	2	3	8	33	No
	152.0473	2-Hydroxyphenylacetic acid	NA	1	1	4	1	11	No
	183.0531	4-Pyridoxate	NA	0	1	1	1	2	No

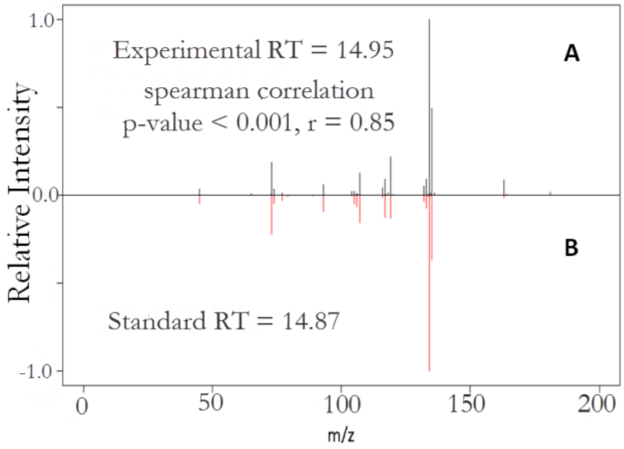


Figure 14. Mirror plot for 4-hydroxyphenyllactate. (A) experimental data. (B) data from high-purity chemical standard. This is considered a match by retention time (RT difference < 3 minutes) and by MS/MS (spearman rank correlation p-value < 0.05 and r-value > 0.

presence of a metabolite that is absent in KEGG’s catalog.

4.3 Discussion and Conclusions

Utilizing EMMA on metabolomics data from CHO cells, our results indicate that the use of biological context filtering during annotation can be powerful, yielding superior speedups and enhances annotation results. A handful of other studies have also suggested that biological knowledge could be exploited to enhance metabolite annotation. For example, a method is described for identifying substrate-product pairs based on the mass differences between pairs of detected MS features [72]. In this method, the mass difference between a pair of features is matched against mass differences between substrate-product pairs of common metabolic conversions (e.g., oxygenation, acetylation, etc.), with a match indicating a potential relationship between the pair of detected feature masses. These relationships can be explored to propagate metabolite annotation from an identified metabolite to its potential reactants and products. In contrast to this method, which limits discoveries to those possible using manually curated metabolic conversions, EMMs provide systematic discovery of derivative metabolites in a manner that is specific to the biological sample. Another method, iMet, suggests that neighboring metabolites within a metabolic network have similar MS/MS spectra and trains a classifier to predict if two metabolites are neighbors [73]. The classifier is trained using MS/MS spectra from spectral databases and mass differences between reactant pairs from KEGG that are not specific to the biological sample. In contrast, EMMA does not require any MS/MS training data and utilizes biological context that is specific to the sample.

From a workflow perspective, EMMA offers the advantage of separating the identification of the biological context from the annotation process, thus allowing for diverse and flexible annotation workflows that can easily incorporate multiple annotation tools and databases. For example, the EMMA workflow was used to assess how a large database such as PubChem can enhance annotation beyond what is possible using a smaller biological database such as KEGG. For our datasets, there was an average increase of 2.65 and 2.80x in the number of annotated masses and number of biologically relevant candidate molecules. This result provides the first evaluation of the benefit of using a non-biological database over a biological database for annotation, and emphasizes the need for biological filtering to make larger databases accessible for annotation.

EMMs are constructed using a previously developed pattern-matching method (*PROXIMAL*), which was developed to identify metabolic derivatives of ingested foreign chemicals due to Cytochrome P450 (CYP) enzymes, highly promiscuous enzymes utilized for detoxification. *PROXIMAL* derives operators from a specified set of KEGG reactions by analyzing transformations between substrate and product while focusing on a local molecular neighborhood centered on the reaction center [104]. It is beneficial to utilize *PROXIMAL* operators to create EMMs as *PROXIMAL* mimics actions expected through substrate promiscuity, where an enzyme recognizes multiple substrates and exhibits broad specificity

[108]. Prior approaches to computing substrate promiscuity relied on using a set of hand-curated rules. A list of 50 reaction rules, each associated with one or more reactions, was defined to explore novel synthesis pathways [109, 110]. The BNICE (Biochemical Network Integrated Computational Explorer) framework derives a set of hand-curated rules based on examining reactions at their third level of E.C. (Enzyme Commission) specificity [111]. The rules are applied repetitively to generate novel synthesis [111] or degradation pathways [112], but are not publically available. A list of biochemical conversions expected to occur frequently in metabolism was used to identify novel metabolic products not previously described in plants [72]. Further use of these types of rules allowed the compilation of predicted metabolic products into databases such as MINEs [38], using BNICE operators, or MyCompoundID [83], by the repeated (up to two times) application of addition or subtraction of expected functional groups. In contrast, using operators derived through *PROXIMAL* allows creation of an EMM specific to the biological sample under investigation.

To the best of our knowledge, we present the first experimental evidence for a computationally predicted metabolite derived through promiscuous action of an enzyme. Using a chemical standard, we confirm the presence of 4-hydroxyphenyllactate in a CHO cell culture, even though this metabolite is currently not listed as a CHO cell metabolite in KEGG CHO model. It is unlikely that the source of 4-hydroxyphenyllactate lies outside of CHO cell metabolism, as the cell culture medium was chemically defined and did not include this metabolite. This is a proof to the ability of EMM to expand the metabolic search space for annotation.

An EMMA limitation relates to the metabolic model used to generate the corresponding EMM. Genome-scale metabolic reconstructions can be inaccurate or incomplete, especially for non-model organisms. This problem is highlighted by the case study on the cecal culture, which comprises a complex microbial community of more than one hundred species. In contrast to the CHO cell, there is no well-curated reference genome annotation for this complex community.

The problem of incomplete models becomes more significant when it comes to the validation of promiscuous activity of an enzyme. Our biotransformation prediction by *PROXIMAL* suggests that the metabolite 4-Hydroxyphenyllactate may result from the promiscuous activity of one or more carboxylic acid dehydrogenases expressed in the CHO cell on the substrate molecule 4-Hydroxyphenylpyruvat. Although this can be a validation of the promiscuous activity of enzymes, it can also be due to the CHO model not being complete and missing the corresponding biotransformation.

Despite the limitations, EMMA demonstrates great utility in creating an expanded, biologically relevant annotation context and in utilizing this context to enhance annotation. EMMs provide annotation opportunities beyond those possible with metabolic models without the high cost of searching large structural databases that contain many non-biological compounds. Exploring a large database such as PubChem without the proposed filtering is simply prohibitive as it would require up to 27,097x machines in parallel must be deployed to achieve the same runtime obtained with the EMMA workflow. The reduced number of candidate metabolites followed by decreased runtime of *in silico* annotation of candidates make it practical to use large general databases, e.g. PubChem, as the search space for annotation. While we demonstrated EMMA workflow using specific tools and databases (e.g., CFM-ID for annotation; PubChem and KEGG databases) the overall workflow is generic and can be readily modified to use other *in silico* annotation tools and other databases.

In summary, applying EMMA to untargeted LC-MS data collected from cultures of Chinese hamster ovary (CHO) cells and murine cecal microbiota shows how EMMA enhances the chance of discovering previously uncharacterized metabolites, while reducing the computational burden associated with annotation. Compared to an *in silico* annotation workflow that analyzes every candidate compound in large chemical databases, EMMA reduces the number of calculations by 4 orders of magnitude. Further, EMMA increases the number of annotated masses and number of chemical identities by an average of 1.71 and 2.39-fold, respectively, when compared to using the sample's metabolic model. Further, the results show that EMMA increases the number of annotated masses and biologically relevant candidate molecules by the average of 2.65 and 2.80-fold, respectively, when compared to using candidate sets from KEGG. The experimental confirmation of the presence of 4-hydroxyphenyllactate, a CHO cell metabolite in the EMM that has not been previously identified as part of CHO cell metabolism, further demonstrates the effectiveness of EMMA. Collectively, our results show that it is necessary and practical to adapt innovative workflows as presented here to overcome annotation hurdles.

Chapter 5
Bayesian Probabilistic Modeling for Pathway Activity Analysis using Untargeted Metabolomics

We present in this paper a novel inference-based probabilistic approach, termed Probabilistic modeling for Untargeted Metabolomics Analysis (PUMA), for predicting the likelihood of activity of metabolic pathways and then deriving probabilistic assignment of measurements to candidate chemical identities. PUMA first constructs a graphical model [113] that captures the uncertainty of assigning observed measurements to pathways. PUMA then utilizes Gibbs sampling [114] to perform Bayesian inference [115] to approximate the posterior probabilities of pathway activities and metabolite annotations conditioned on the measurements.

5.1 Methods

To determine pathway activities, an untargeted metabolomics workflow (Figure 15A) begins with collecting measurements, followed by metabolite annotation using annotation tools (e.g. database look ups or annotation tools) and then applying pathway analysis tools (e.g. ORA or TA) to determine pathway activities. A pathway is assumed *active* when biological and environmental factors lead to the production of some or all of its metabolic products. In some cases, metabolite annotation is skipped and statistical pathway activity is computed directly from measurements [83]. In contrast, our inference-based approach utilizes a generative model (Figure 15B) that mimics biological processes inherent to the sample under study. In this work, the generative model assumes the following biological process. Within the sample, one or more pathways are active. An active pathway causes the presence of some its metabolites, which in turn results in observations of masses through untargeted metabolomics data collection.

A generative model is powerful because it captures complex relations among pathway activities, metabolites, and measurements in a single integrated model. Importantly, the generative model produces values that are observed (measured), as well as hidden variables of interest, which cannot be directly observed but rather inferred from those values that can be observed. In our case, the observations correspond to mass measurements collected through untargeted metabolomics. The hidden variables are pathway activities and the presence of a metabolite in a biological sample. The generative model is constructed using biological knowledge in the form of known relationships between pathways and metabolites, and metabolites and their masses. The generative model is parameterized with knowledge about the behavior of the biological process. Priors are provided for pathway activities.

Once the generative model is constructed, the next step in our approach is to perform inference on the model to compute posterior probabilities for variables of interest. As computing such probabilities is typically intractable, they are estimated using Gibbs sampling, a Monte Carlo Markov Chain (MCMC) sampling technique [115] Specifically, inference allows computing two types of probability distributions that are of interest: the probability of pathways being active and the probability of a metabolite being present in the sample, where both probabilities are conditioned on the observations.

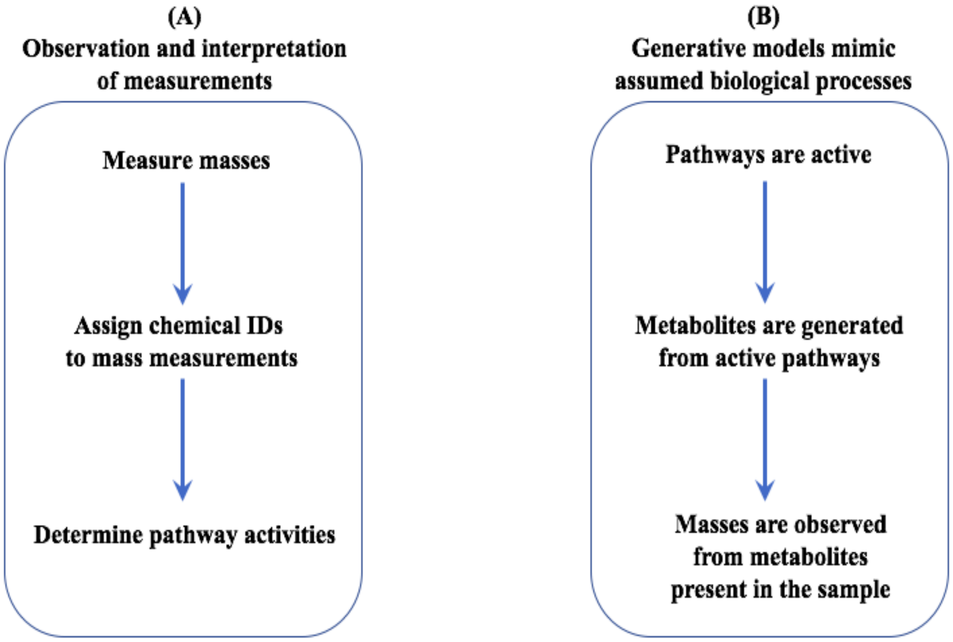

(A)
Observation and interpretation of measurements
Measure masses
Assign chemical IDs to mass measurements
Determine pathway activities
(B)
Generative models mimic assumed biological processes
Pathways are active
Metabolites are generated from active pathways
Masses are observed from metabolites present in the sample

Figure 15. Comparison of a workflow to collect and interpret observations (A), and a generative model that captures a biological process (B).

5.1.1 *Illustrative Example*

A small example is provided to illustrate some of the challenges in mapping measurements to metabolites and pathways, and to show inference's ability to address these issues. Figure 16 presents a snippet of a network that shows two pathways (ovals), Pathway 1 and Pathway 2. Metabolites with known chemical identities associated (circles) are either associated with one pathway (red circle) or more than one pathway (blue circles). Measurements (squares) correspond to masses that can be associated with one particular metabolite (red square) or multiple metabolites (blue squares). Not all metabolites within a sample are measured due to either instrument limitations or because they are simply not present in the sample due to biological or environmental factors. Some metabolites are thus not associated with any measurements (white circles), and some may be associated with one or more pathways.

There are two types of uncertainties in interpreting measurements from untargeted metabolomics. One type of uncertainty relates to assignment of metabolites to pathways (circles to ovals, Figure 16). For example, measurement w_3 is assigned to metabolite f_5. Because f_5 is a metabolite common to both Pathways 1 and 2, there is an uncertainty in assignment of the metabolite to the pathways: f_5 can be the product of activity in either Pathway 1 or Pathway 2. The other uncertainty relates to assignment of measurements to metabolites, when a measurement can map to multiple metabolites (squares to circles, Figure 16). Measurement w_4 can be attributed to one or two metabolites, f_6 and f_7, both sharing the same mass. The uncertainty in assigning w_4 to metabolites f_6 and f_7 manifests in further uncertainty. If w_4 is associated with f_6, then it contributes to the activity of Pathway 1 (and/or other pathways with which f_6 is associated), while, if w_4 is associated with f_7, then it contributes to the activity of Pathways 2 (and/or other pathways with which f_7 is associated). Not all measurements contribute to these uncertainties. For example, measurement w_5 is unique to metabolite f_{13}. In turn, f_{13} is unique to Pathway 2. Some measurements (such as w_5) clearly contribute more significantly than others (such as w_3 and w_4) in determining pathway activities.

Computing pathway activities using an enrichment ratio can be misleading, because it does not take into account the uncertainty in attributing measurements to metabolites and pathways. The enrichment ratio for Pathway 1 can be computed as the ratio of 4 putatively measured metabolites divided by 6 total metabolites in the pathway. While this enrichment ratio seems high, there is little confidence that Pathway 1 is active since all measured metabolites form this pathway could be due to active pathways other than Pathway 1. Pathway 2 has an enrichment ratio equal to 3 divided by 8. The significance or importance of this ratio is unclear. Inference will conclude that Pathway 2 is active with high probability, as it includes a measured metabolite that cannot be attributed to the activity of any other pathway. In contrast to enrichment methods, our inference-

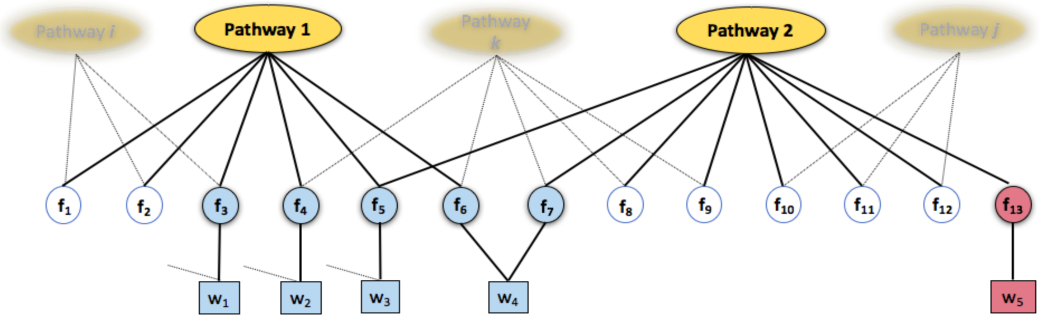


Figure 16. Illustrative example of uncertainty when mapping measurements to metabolites and pathways. Pathways (ovals) are associated with metabolites (circles), which in turn are associated with measurements (squares). White circles represent non-measured metabolites with membership in one or more pathways. Blue circles represent measured metabolites that have multiple-pathway memberships. The red circle represents a metabolite that has membership in only one pathway. Measurement w_5 uniquely maps to f_{13}, which uniquely maps to Pathway 1, while all other measurements map to multiple metabolites, as shown by solid or dotted lines.

based technique considers uncertainties in measurement-metabolite and metabolite-pathway relationships when computing the likelihood of pathway activities.

5.1.2 *Constructing Generative Models*

To create a generative model, we assume that a biological sample has a metabolic model with P pathways, F metabolites and K unique metabolite masses. A metabolite may have membership in one or more pathways. To map measurements to metabolites, masses of the model metabolites are compared to measurements with a predefined error margin (15ppm). A measured mass may be associated with one or more metabolites.

Let $a = (a_p : p = 1, \dots P)$ denote the status of P pathways in the biological sample, so a is a vector of binary random variables, where a value of 1 indicates that the corresponding pathway is active and 0 indicates inactivity. As a prior probability distribution on $a = (a_p : p = 1, \dots P)$, we assume that the a_i are distributed i.i.d. (independent identically distributed) Bernoulli(λ) with λ = 0.5. Pathway activity is a function of genetic and environmental factors specific to the biological sample under study. We assume a pathway is active with a Bernoulli probability.

In our generative model, a metabolite within a pathway can be generated due to pathway activity with some probability. Matrix μ is defined with P rows and F columns as a mapping of metabolites in the biological sample to the pathways, where $\mu_{p,f}$ is equal to the probability of generation of metabolite f due to pathway p. Let $o_{p,f}$ be a binary random variable indicating whether pathway p generates metabolite f in the sample. Each element in matrix o is defined as a Bernoulli random variable with a probability dependent on the corresponding elements in μ and a. In other words, we define the distribution of elements in o as $o_{p,f} \sim$ Bernoulli$(a_p \times \mu_{p,f})$.

A metabolite can be generated due to activity of one or more pathways in which the metabolite participates. Vector m is defined with F elements, where m_f represents the presence of metabolite f in the biological sample. The matrix o dictates the presence of metabolite m_f in the sample, where $m_f = 1 - \prod_p (1 - o_{p,f})$.

Metabolites that are present in the biological sample determine mass measurements that can be observed through untargeted metabolomics. We define τ to define the relationship between metabolites and mass measurements. τ is a matrix with F rows and K columns. Each element of matrix τ is a value between zero to one, representing the probability of mass of a metabolite to be measured. As some metabolites share the same chemical formula and the same mass, a mass can be observed if at least one of the metabolites with the same mass is present in the biological sample. Vector w with K elements is defined to represent observed masses, where w_k shows if mass k has been successfully measured. Each element of w is defined as a Bernoulli random variable. The probability of an observed mass is dependent on it being associated with one or more metabolites with that mass and the probability of generating any of these metabolites. w_k is equal to one if mass k is observed due to presence of at least one metabolite f with mass k in the biological sample. Thus, the distribution of elements in w is defined as, $w_k \sim$ Bernoulli (1 -$\prod_f (1 - \tau_{f,k} m_f)$.

A plate representation [116] of the model shows a list of dependencies in a graphical form as a directed acyclic network (Figure 17). The directed graph represents the joint probability distribution over random variables in the model.

The conditional probability of each random variable depends only on its parents in the graph. Each box shows the graphical representation of a conditional probability to calculate a specific variable based on the involved dependencies. To avoid representing all F metabolites, P pathways and K masses in the graph, we used the ‘plate’ notation by drawing one representative node per variable, and enclosing these variables in a plate (rectangular box). The number of instances of each enclosed variable is indicated by the fixed constant in the lower right corner of the box. The described model presents the joint probability distribution of random variables a, o, m and w defined as:

$$p(a, o, m, w) = p(a; \lambda)\ p(o|a; \mu)\ p(m|o)\ p(w|m; \tau).$$

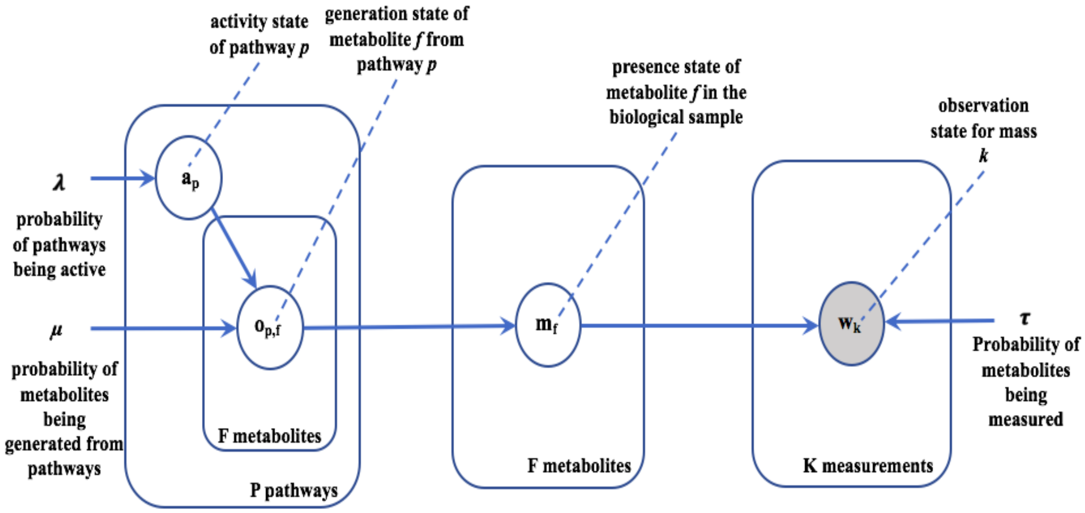

activity state of pathway p
generation state of metabolite f from pathway p
presence state of metabolite f in the biological sample
observation state for mass k
λ
probability of pathways being active
μ
probability of metabolites being generated from pathways
a_p
$o_{p,f}$
F metabolites
P pathways
m_f
F metabolites
w_k
K measurements
τ
Probability of metabolites being measured

Figure 17. Graphical representation of the generative model illustrates dependencies among random variables. Random variables of the model (a, o, m, w) are shown in white or shaded circles. The variable m represents a deterministic random variable. A shaded circle (w) represents an observed random variable. $\mu, \lambda, \tau$are parameters to the model.

5.1.3 *Inference*

Using the probabilistic model, we infer pathway activities and metabolite presence from mass measurements. Specifically, we calculate the following probabilities. For each pathway p in the biological sample we calculate $p(a_p|w)$, the posterior probability of pathway p being active. For each metabolite f, we then calculate $p(m_f|w)$, the posterior probability of f being present in the biological sample. We use the latter probabilities to rank a candidate set for each mass measurement, where a *candidate set* includes a list of metabolites with known chemical identities with the same mass (+/- error margin) as the observed one.

5.1.3.1. Inferring pathway activities

Gibbs sampling is employed to approximate the posterior probability $p(a|w)$. To avoid sampling all hidden variables, we marginalize out matrix o and vector m from the model. We first marginalize out o from the partial model $p(a, o, m)$ and divide by $p(a)$, giving $p(m|a)$, the probability of metabolites presence given the activity state of all the pathways. $p(m_f|a)$, the probability that at least one pathway in the biological sample generates metabolite m_f,is denoted as $\varphi_f(a)$, which is calculated as 1 - $\prod_p(1 - a_p\mu_{p,f})$. We now complete the model by multiplying by $p(w|m; \tau)$ and marginalize out m, giving $p(w|a)$. $\theta(a)$ is a vector with K elements, where each element is associated with the probability of mass w_k being observed given a. This probability is calculated as $\theta_k(a)$ = 1 – $\prod_f (1 - \varphi_f\tau_{f,k})$, denoting that weight w_k can be observed if at least one of the metabolites with this weight is present in the biological sample. With probabilities $p(a)$ and $p(w|a)$, we can draw samples from the posterior $p(a|w)$ through Gibbs sampling. After collecting Gibbs samples, we are able to estimate $p(a_p|w)$ for each pathway p.

5.1.3.2. Inferring metabolite annotations

With samples drawn from $p(a|w)$, we approximate $p(m_f|w)$, which is the posterior probability of metabolite f being present in the biological sample. The probability $p\ (m_f\ = 1|\ w)$ can be approximated by calculating $p\ (m_f\ = 1|\ a,\ w)$ as follows:

$$p\ (m_f = 1|\ w) = \sum_{a\in\{0,1\}^P} p\ (m_f = 1|\ a, w)\ p(a|w) \approx \sum_{a\in samples\ of\ p(a|w)} p\ (m_f = 1|\ a, w)$$

The probability $p\ (m_f|\ a, w)$ can be calculated by conditioning $p\ (m_f, w|\ a)$ on w as $p\ (m_f|\ a, w) =\ p(m_f, w|\ a)/p(w|a)$. Let k_f be the mass of m_f. Observation of any mass $w_{k'}$ such that $k' \neq k_f$ is independent of m_f given the pathway activities summarized in a. We conclude $p\ (m_f, \mathrm{w}|\ a)$ can be calculated from $p(m_f, w_{k_f}|a)$, which results in $p\ (m_f|\ a, w) =\ p(m_f, w_{k_f}|\ a)/p(w_{k_f}|a)$.

The calculation of probability $p\left(m_f = 1, w_{k_f}\middle|a\right)$ is now manageable. Let $\alpha_f = \{f':metabolite\ f'has\ weight\ k_f, f' \neq f\}$, then $p\left(m_f = 1, w_{k_f}\middle|a\right)$ is calculated as follows:

$$p\left(w_{k_f} = 1, m_f\middle|\ a\right) = \begin{cases} p(m_f = 1|a) & \text{for } m_f = 1 \\ p(m_f = 0|a)(1 - \prod_{f'\in\alpha_f}[1 - p(m_{f'} = 1|a)]) & \text{for } m_f = 0 \end{cases}$$

$$p\left(w_{k_f} = 0, m_f\middle|\ a\right) = \begin{cases} 0 & \text{for } m_f = 1 \\ p(m_f = 0|a)(\prod_{f'\in\alpha_f}[1 - p(m_{f'} = 1|a)]) & \text{for } m_f = 0 \end{cases}$$

The expression $p(w_{k_f} = 1, m_f|\ a)$ represents the probability of observing mass w_{k_f} due to two cases. In the first case, the observation of mass w_{k_f} is due to presence of m_f in the biological system. In the second case, the observation of mass w_{k_f} is due to presence of any metabolite other than m_f but with the same mass. Normalization of two probabilities resulting from these two cases generates the desired probability $p(w_{k_f} = 1, m_f \mid a)$. The probability $p(w_{k_f} = 0, m_f \mid a)$ indicates the absence of observation of mass w_{k_f}. As explained earlier, $p\ (m_f \mid a, w) = \ p(m_f, w_{k_f} \mid a)/p(w_{k_f} | a)$. This results in $p\left(m_f = 0 \mid a, w_{k_f} = 0\right) = 1$.

By marginalization of pathway activities (a) from posterior samples, we approximate the posterior probability $p(m_f|w)$ for each metabolite f. The derived probabilities are used as a scoring metric to rank a candidate set for each mass measurement.

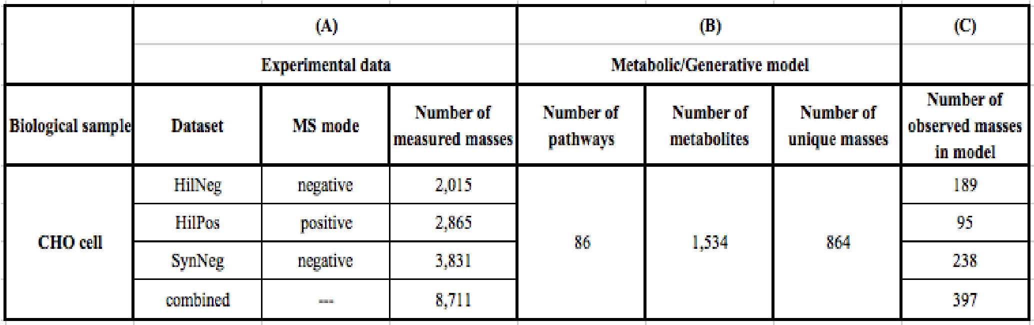
5.2 Results

We apply PUMA to untargeted data collected from cultures of Chinese Hamster Ovary (CHO) cells belonging to a low growth cell line [74]. The case study was selected based on the availability of untargeted metabolomics datasets and access to the relevant metabolic models.

Data was collected using liquid chromatography-mass spectrometry (LC-MS) based metabolomics (Table 17, column group A). To increase coverage, data is collected separately under three different combinations of liquid chromatography methods and ionization modes (positive or negative). When combined, the data provides a more comprehensive characterization of the sample in the form of 8,711 measurements. The metabolic model for the CHO cell was derived from KEGG [35], based on unique metabolites and pathways for the cricetulus griseus (Chinese hamster) under organism code cge. The number of pathways, metabolites and unique masses are listed in Table 17B.

Due to incompleteness of metabolic models and noisy data, only a small fraction of mass measurements, 397 masses, correspond to masses in the metabolic model (Table 17C). The number of observed masses in the model is used to initialize the observation vector w for each dataset.

Table 17. CHO cell case-study data: (A) untargeted metabolomics datasets, (B) metabolic model, (C) number of observations that match to metabolites in CHO



	(A)			(B)			(C)
	Experimental data			Metabolic/Generative model			
Biological sample	Dataset	MS mode	Number of measured masses	Number of pathways	Number of metabolites	Number of unique masses	Number of observed masses in model
CHO cell	HilNeg	negative	2,015	86	1,534	864	189
	HilPos	positive	2,865				95
	SynNeg	negative	3,831				238
	combined	---	8,711				397

5.2.1 *PUMA implementation and parameter initialization*

We implemented PUMA using PyMC3 [117], a probabilistic programming framework that allows for automatic Bayesian inference on user-defined models. To draw samples from a posterior distribution, PyMC3 utilizes Gibbs sampling, a Markov Chain Monte Carlo (MCMC) sampling technique [118] [119]. The generative model for each case was derived from the metabolic model for the sample under study. Each such metabolic model specifies pathways, metabolites, and membership of metabolites in pathways. The mass of each metabolite is available through KEGG or other databases. The model parameters were initialized as follows. Each of the P elements in the vector λ are set to 0.5, implying that all pathways have the same prior of being active. Matrix μ is initialized based on the mapping of metabolites to pathways in the metabolic model of the sample under study. Each entry in μ is assumed to be 0.9 or 0 according to whether a metabolite is associated with a pathway. To generate a vector of observations w, we compare the unique masses of metabolites in the metabolic model to mass measurements collected through untargeted metabolomics with 15ppm margin of error. The default setting for w_k is 1 if the mass of metabolite k falls within 15ppm from any of the measurements. Each entry in τ is set to zero or one according to whether a metabolite is associated with a mass. T, the number of samples to draw from the model, is a variable that can be set in PyMC3 with default value equal to 500. The sampler was run multiple times with values of T equal to 500, 1000 and 1500. For all the reported runs, increasing number of drawn samples did not affect the computed probabilities for pathways activities.

5.2.2 *Pathway activity probabilities are inferred from the probabilistic model*

We applied inference on the generative model for the CHO cell. A list of pathways in CHO cell that are identified as active, with $p(a_p|w)$ equal to or greater than 0.5, in at least one of the datasets is provided (Table 18). The KEGG IDs and names are listed. The number of metabolites within a pathway is designated as the pathway size. The number of mass measurements that could be mapped to each pathway is reported, columns 4-7, for the various datasets. The last four columns indicate predicted pathway in datasets HilNeg, HilPos, SynNeg and combined.

As mass observations differ from one set of measurements to another, the predicted activity differs among the datasets. We expect that the combined dataset, with the highest number of mass measurements, is the most reliable predictor of pathway activity. To investigate, we analyze the reported activity levels in Table 18. There are several cases to consider. In some cases, e.g. cge00785 and cge00970, pathways that are predicted active by each individual dataset and the combined dataset. In other cases, e.g. cge00072, and cge00053, pathways are predicted active in the combined dataset, but not predicted active for all other individual datasets. In such cases, individual dataset measurements when considered independently of others did not provide inference sufficient evidence to conclude that the pathway is active. As an example, for pathway cge00053, with size nine, the number of mass measurements in SynNeg and combined datasets that can be mapped to the pathway is seven. PUMA predicts this pathway active in both datasets. However, the same pathway is not predicted active in HilNeg and HilPos, where the number of mass measurements that can be mapped to the pathway is reduced to four and zero, respectively.

In other cases, some pathways (e.g. cge00730, cge00040) are predicted active by at least one of the individual datasets while predicted not active by the combined dataset. Additional evidence in the form of a larger number of mass measurements in the combined dataset affects the predicted activity for pathways with common metabolites. For example, pathway cge00730, with size seven, is predicted active by HilNeg (probability of activity is 0.72) but not predicted active by the combined dataset (probability of activity is 0.43). In both datasets, three mass measurements can be mapped to the pathway, while two of these mass measurements can also be mapped to cge00970. With an increase in the number of mass measurements that can be mapped to cge00970 from 12 in HilNeg to 19 using the combined dataset, cge00970 has a higher probability of being active (probability of activity is 1.0) compared to cge00730 (probability of activity is 0.43). For the rest of the CHO cell analysis, we utilize the combined dataset, as it is the most predictive dataset with highest number of measurements.

We investigated the biological relevance of some of the pathways predicted active by the combined dataset. Cge00780 (Biotin metabolism) is a pathway involved with synthesis of Biotin (vitamin B7) supports adrenal function and aids in maintaining the nervous system. Cge00785 (lipoic acid metabolism) is involved with synthesis of lipoic acid, an essential cofactor for the activity of dehydrogenase enzymes that assist in energy production. Cge00020 (TCA cycle) is essential for cellular metabolism, playing an important role in the energy production. Cge00970 (Aminoacyl-tRNA biosynthesis) is involved in protein synthesis and the translation from genes to proteins, which is a crucial process for a cell. Cge00053 (Ascorbate and aldarate) is involved with synthesis of Ascorbate (vitamin C), which is essential for health. It is expected that such pathways are active in the sample.

Table 18. List of CHO cell pathways predicted active by PUMA in at least one dataset. For each pathway, the table lists the pathway ID in KEGG, pathway name, pathway size given in number of metabolites per pathway, number of mass measurements that can be mapped to metabolites in the pathway, and the prediction of being active (Y, if active; N, otherwise) for each individual dataset and the combined dataset. Pathway IDs with an * are identified as *not* statistically enriched by Fisher's Exact test.

Pathway ID	**Pathway name**	**pathway size**	**#masses mapped to the pathway (HilNeg)**	**#masses mapped to the pathway (HilPos)**	**#masses mapped to the pathway (SynNeg)**	**#masses mapped to the pathway (combined)**	**Identified active Due to HilNeg dataset?**	**Identified active due to HilPos dataset?**	**Identified active due to SynNeg dataset?**	**Identified active due to the combined dataset?**
cge00780	Biotin metabolism	12	1	1	3	3	N	N	Y	Y
cge00785	lipoic acid metabolism	11	2	3	2	4	Y	Y	Y	Y
cge00020	Citrate cycle (TCA cycle)	19	5	2	9	10	N	N	Y	Y
cge00970	Aminoacyl-tRNA biosynthesis	69	12	13	18	19	Y	Y	Y	Y
cge00561*	Glycerolipid metabolism	18	6	0	3	7	Y	N	N	Y
cge00565	Ether lipid metabolism	31	3	2	1	5	N	N	N	Y
cge00072*	Synthesis and degradation of ketone bodies	6	1	2	3	5	N	N	N	Y

cge00591	Linoleic acid metabolism	5	0	2	0	2	N	Y	N	Y
cge00053	Ascorbate and aldarate metabolism	9	4	0	7	7	N	N	Y	Y
cge00290	Valine, leucine and isoleucine biosynthesis	8	2	3	7	8	N	N	Y	Y
cge00524	Neomycin, kanamycin and gentamicin biosynthesis	2	1	0	1	1	Y	N	Y	Y
cge00730	Thiamine metabolism	7	3	2	2	3	Y	N	N	N
cge00040	Pentose and glucuronate interconversions	18	10	0	6	10	Y	N	N	N
cge00472*	D-Arginine and D-ornithine metabolism	4	2	1	2	2	Y	N	Y	N

5.2.3 *Comparison of predicted pathway activities to enrichment ratios*

We investigate how inference compares with pathway enrichment ratios [79]. We define the enrichment ratio for a particular pathway as the ratio of measured masses that map to metabolites within the pathway to its size. Pathways that are labeled as *statistically enriched* based on statistical significance of the ratios using Fisher's Exact Test (FET). The null hypothesis is that there is no difference between the enrichment ratio of pathway *p* and ratios of other pathways in the sample. A *p-value* equal to or less than 0.05 is considered significant. Of the 14 pathways designated as active using PUMA, all but three pathways (cge00561, cge00472, and cge00072) are statistically enriched.

Enrichment ratios of CHO cell pathways are contrasted against pathway activities that are predicted by PUMA (Figure 18). While there is some consensus between the two techniques (upper right and lower left parts of Figure 18, there are important differences. In some cases, PUMA designates pathways as active despite low enrichment ratios. For example, pathways cge00780 and cge00970, with pathway sizes 12 and 69, respectively, have a predicted activity of 0.99 and 1.0, respectively, when using the combined dataset. The enrichment ratios for these two pathways are 0.25 and 0.27, respectively. The low enrichment ratio may indicate inactivity, and enrichment ratios for both pathways are statistically enriched. Inference however predicts them both as active. In another set of cases, PUMA predicts low pathway activity, while enrichment assumes a high enrichment ratio. For example, statistically enriched pathway cge00400 has an enrichment ratio of 0.66, but assigned active by PUMA with probability 0.11. cge00400 includes six metabolites, of which four were observed using measurements. Two of the four observed mass measurements from cge00400 can also be mapped to cge00970. Cge00970 is unique in generating a unique measurement that cannot be generated by any other pathway in model (similar to the case of w_5 in our illustrative example). As the result, cge00970 is predicted active with high probability, which in turn reduces the probability of cge00400 being active. The remaining two observed mass measurements in cge00400 are not unique to this pathway as they can be generated by at least one other pathway

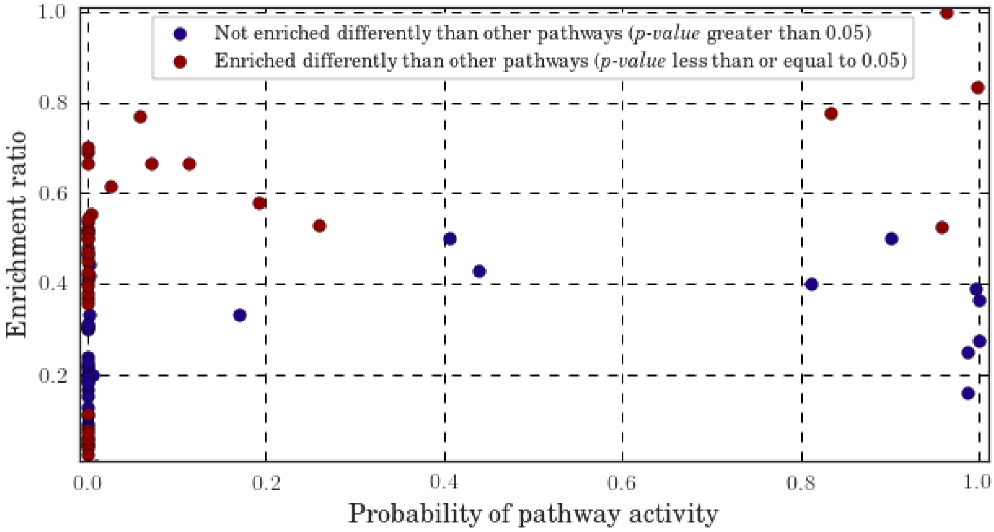


Figure 18. Contrasting probabilities of pathway activities as computed by PUMA vs. enrichment ratios in CHO cell. Each data point is marked as either statistically enriched (red) or non-statistically enriched (blue) based on Fisher's Exact Test with a significance level of 0.05 for the *p-values*.

in the biological sample. Despite its high enrichment ratio, PUMA does not assign a high probability of activity of cge00400.

5.2.4 *PUMA annotations show agreements with other tools and annotate new metabolites*

Among the 1,534 model metabolites (Table 17), there were 352 metabolites that map to 397 mass measurements in the combined data set. A particular mass measurement was associated with a model metabolite if its mass matched the measured mass within 15ppm error margin. Therefore, each measurement may have zero, one or more putative annotations. The probabilities of each metabolite being present in the sample as inferred by PUMA are used to score and rank metabolites. Here, only the top ranked metabolite(s) for each mass is considered as the PUMA *candidate set.* The lowest probability of assignment was 0.17, which occurs in 3 annotation cases. All other probabilities were equal to or greater than 0.29. All but 22 of the 397 annotations had a likelihood of greater than 0.5.

We assess the accuracy of PUMA annotations by comparing the level of agreement of PUMA annotations with annotations using two other techniques, spectral database searches and BioCAN (Figure 19). Spectral signatures collected through untargeted metabolomics were looked up in s METLIN and HMDB, and were previously reported [74]. The highest scoring metabolites for each measurement in METLIN and in HMDB formed the spectral database candidate set. Out of 397 mass measurements, 85 were identified as either in HMDB or METLIN. For each measurement, the PUMA candidate set was compared against the candidate set identified by HMDB and METLIN. The comparison leads to four different scenarios. One scenario is “agreement”, where the highest-ranked candidate metabolite identified by PUMA exactly matches the candidate set from HMDB and METLIN. Such agreement occurs in 64 cases. Another scenario is “semi-agreement”, in which the candidate set from HMDB and METLIN is a subset of the top candidate set obtained from PUMA annotation. There are 15 cases of semi-agreement. Another scenario is “disagreement”, where the candidate set from METLIN and HMDB does not overlap with the PUMA candidate set. We investigate the six disagreements. In three cases, the candidate metabolite from METLIN and HMDB is the second likely putative annotation identified by PUMA. These putative annotations, which were not included in the PUMA candidate set, had a high activity score and close to that of the metabolite(s) in the candidate set. In the remaining three cases, however, the candidate metabolite from METLIN and HMDB is assigned a low score by inference-based annotation workflow, a score far from the one assigned to the metabolite in the PUMA candidate set. These three cases are considered as genuine disagreement in annotation. Importantly, in the final scenario with 312 cases, there were not matching annotations in METLIN and HMDB. These cases are new annotations provided by PUMA and labeled as “Only PUMA”.

We compare PUMA annotations to annotation results by BioCAN [74]. BioCAN aggregates the results from spectral database searches and *in silico* fragmentation tools, and estimates the confidence in an annotation for a mass measurement not only based on a consensus but also by the confidence of presence of metabolites that are connected to the mass measurement through the substrate-product relationships. BioCAN annotates 346 out of 397 mass measurements that are annotated by PUMA. We analyze the various scenarios as we did when comparing with annotations using METLIN and HMDB. There are 273 cases of agreement, 54 cases of semi-agreement, 19 cases of disagreement, and 51 new annotations by PUMA. The disagreements fell into two categories. In 11 out of 19 cases, there was disagreement on the top candidate, where PUMA ranked BioCAN's candidate as second best. There were genuine disagreements in 8 cases were the annotation by BioCAN was assigned a low score by PUMA.

In summary, comparing PUMA annotations against those obtained through spectral database and BioCAN shows significant levels of agreement. METLIN, HMDB and BioCAN incorporate spectra signatures during annotation while PUMA relies solely on pathway organization and mass measurements. Importantly, PUMA increased annotation by 367% over spectral databases and by

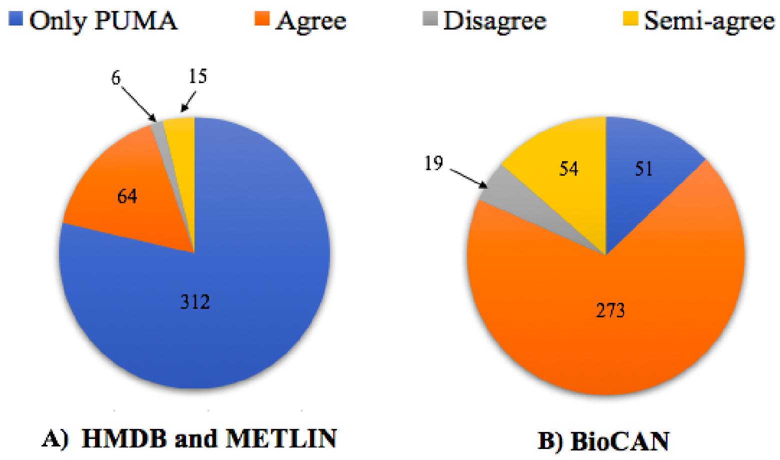


Figure 19. Comparison of the number of metabolite annotations attained with PUMA against those identified by: (A) searching spectral databases, HMDB and METLIN, and (B) BioCAN. The blue slice in each pie represents the number of mass measurements that could only be annotated by PUMA. Other slices indicate agreement (orange), semi-agreement (yellow), and disagreement (gray) between the annotations found by PUMA and the other tools.

14% over BioCAN. PUMA identifies candidate metabolites that are not identified by other annotation tools. The quality of these annotations can be further assessed by employing an *in silico* fragmentation tool (e.g. CFM-ID [33]) to compare and score computationally generated spectral signatures against experimentally collected spectral signatures.

5.2.5 *Evaluation of PUMA in overcoming uncertainty in annotation*

To reduce the uncertainty inherent in mapping measurements to metabolites with a same mass, we incorporated annotation data obtained through spectral database search based on spectral signatures in METLIN and HMDB into the generative model. Specifically, for each mass *k* annotated using METLIN or HMDB as metabolite *f*, matrix τ is modified. All column entries other than $\tau_{f,k}$ are set to zero, indicating that mass *k* uniquely maps to metabolite f. Using the updated τ for the probabilistic model, PUMA calculated posteriors for pathway activities. There is a slight change in predicted posteriors (an average of 0.002) compared to those obtained using the original τ matrix. The change however does not alter the posterior probabilities sufficiently to modify the list of active pathways (Table 18).

This finding is significant as it shows that inference accounts for uncertainty in mapping masses to multiple metabolites without employing additional annotation information from spectral databases. PUMA can be used to accelerate the process of pathway activity analysis by direct use of mass measurements and bypassing metabolite annotation using spectral databases.

We repeated the analysis, but incorporated the annotation data available from BioCAN instead of that obtained through spectral lookups. The change in τ caused a slight change in predicted posteriors (an average of 0. 006 per pathway) compared to those obtained using the original τ matrix. The one significant change was for pathway cge00360, where pathway activity changed from 0.102 to 0.508. The cge00360 pathway, the phenylalanine pathway, is responsible for producing Tyrosine. This finding is also significant as it shows that substantial additional annotations, as provided in the form of added annotations by BioCAN over using spectral databases, are required to further inform inference in regards to pathway activities.

5.2.6 *Runtime and complexity of the probabilistic model*

We estimated the time and space complexity of the proposed model in terms number of pathways (P), number of metabolites (F), number of unique masses of metabolites (K), number of drawn samples (T). The time and space complexity in sampling the model is O(T x P x F). We timed the runtime of PUMA for constructing the model and performing the sampling. The runtime for the combined CHO cell dataset was 795 seconds. The time and space complexity in annotation is O(T x P x F). The runtime for the combined CHO cell dataset was 17,253 seconds. The runs were performed on a Dell PowerEdge R815 server with four AMD Opteron 6380 processors running at 2.5GHz.

5.3 Discussion and Conclusions

We presented in this chapter PUMA, a probabilistic approach to interpret mass measurements collected through untargeted metabolomics. Because it is based on inference, PUMA allows drawing stronger conclusions about activities of the biological sample under study by folding in what is already known about the sample. In doing so, levels of uncertainty in mapping measurements to metabolites and pathways are significantly reduced, and a clearer view of the likelihood of pathway activity levels and metabolite annotation emerges. PUMA consists of the two steps. Using the metabolic model for a biological sample under study and measured masses, a generative model is constructed to capture known complex relations among pathway activities, metabolites, and measurements. Next, Gibbs sampling is applied to approximate posteriors for pathway activities and metabolites being present in the sample. PUMA was used to analyze untargeted metabolomics data collected from two biological samples.

PUMA provides significant contributions in advancing both pathway analysis and metabolite annotation. Pathways identified by PUMA as highly active are ones with essential biological roles in the samples under study. Further, PUMA offers a perspective on pathway activity that is distinctly different from that offered by statistical enrichment approaches. PUMA identifies pathways that have a high likelihood of being active but have statistically low enrichment ratios, and pathways with low activity probabilities yet with statistically high enrichment ratios. Because inference reduces the uncertainty in mapping measurements to chemical identities, PUMA was able to infer pathway activities without the additional burden of metabolite annotation. For the CHO cell tests case, PUMA was able to infer pathway activity levels similar to those identified with additional annotation information from other tools. In terms of advancing annotation, PUMA results had high agreement to annotations using spectral database lookups and BioCAN. This high level of agreement occurs despite the fact that PUMA does not utilize additional information in form of spectra signatures, as employed by other techniques. Importantly, PUMA suggested annotations for measurements that were not previously annotated by other techniques. In the case of the CHO cell test case, PUMA increased the percentage of mass annotation by 367% over spectral lookups and by 14% over BioCAN.

The performance of the Gibbs sampler is a function of the number of random variables that are sampled. Several optimizations were necessary to reduce the runtime. We selected to predict the likelihood of pathway activities first, as opposed to directly predicting the likelihood of assigning weights to masses, as proposed in ProbMetab [120]. This sped up sampling, as the number of metabolites in a metabolic model is significantly higher than the number of metabolic pathways. For example, in the CHO cell, there are 1535 model metabolites but only 86 pathways. Further, we marginalized out random variables when appropriate. In addition, it was necessary to vectorize computations instead of using for loops to speed up sampling.

A range of studies has employed probabilistic modeling in the field of metabolomics. ProbMetab [121] is a Bayesian annotation tool that uses a probabilistic method [120] to assign empirical formulas to a list of measured mass spectrometric peaks, given a list of potential formulas and possible biochemical transformations. ProbMetab assigns higher probability to formulas that could be created from metabolites in the sample based on the input set of chemical transformations. In another annotation tool, a Bayesian model is proposed to improve metabolite annotation by ranking candidate metabolites matched to a spectral signature using a competing score in addition to a similarity score [122].

The competing score accounts for the likelihood of a candidate metabolite to be matched to other compounds in the spectral database based on its spectral signature similarity to others in the database. The competing spectra however are from a larger spectral database that may not be relevant to the sample. This approach has similarity to PUMA when considering annotations from METLIN and HMDB. Bayesian modeling can be used for purposes other than metabolite annotation in the field of metabolomics. Cancer biomarker discovery studies that use mass spectrometric analysis of human biospecimens can greatly benefit from purification of the data prior to statistical and pathway analyses. To this end, a Bayesian model is proposed, to computationally analyze metabolomics data to identify cancer cells in a biological sample [123]. The proposed approach models the metabolomics data from a heterogeneous biological sample as a weighted mixture of cancerous and non-cancerous features coming from various biomolecule origins and assigns each feature a probability of being cancerous. To the best of our knowledge, PUMA is the first use of Bayesian modeling for the purpose of pathway activity analysis by utilizing metabolomics data.

PUMA may be improved by augmenting the generative model with additional information describing the biological process. One possibility is to modify the value for μ, the probability of generation of metabolite f due to pathway p, to reflect either the centrality of the metabolite within the pathway or to account for the number of ways reactions within the pathway can act on a metabolite. It is also possible to augment the metabolic products within pathways using the concept of enzyme promiscuity, where an enzyme acts on substrates in addition to its natural substrate [101], thus increasing the number of measurements that can be annotated [105]. Hierarchical pathway organization or modularity can also be incorporated into the model. For example, in the metabolic model curated for CHO, some pathways are subsets of others. PUMA predicted pathways cge00072 (Synthesis and degradation of ketone bodies) and cge00290 (Valine, leucine and isoleucine biosynthesis) as active. Both pathways are part of parent pathways cge00650 (Butanoate metabolism) and cge01230 (Biosynthesis of amino acids), respectively. However, PUMA predicted neither parent pathways as active. In addition, instead of using a fixed noise model, a probabilistic noise model, as proposed in ProbMetab [120], can be adapted to reflect uncertainties in measurements. Further, mass differences between measurements as evidence of biochemical transformations [124] can be incorporated in the model as evidence of enzymatic reactions taking place within pathways. Using mass differences between measurements has proven effective when using inference to assign higher probability to metabolites that can be created from others in the sample [120], in constructing networks of putative transformation routes [72], and in identifying related pairs of compounds that have similar spectral signatures [73]. We expect these improvements to further enhance PUMA's ability in interpreting metabolomics data.

In summary, PUMA was shown effective in interpreting data collected through untargeted metabolomics. Using untargeted metabolomics datasets for CHO cells, PUMA identified with high probability a list of pathways considered essential in cellular metabolism. PUMA overcame the uncertainty caused by possibility of matching a mass measurement to multiple candidate metabolites, as incorporating the annotation knowledge from BioCAN and spectral search databases METLIN and HMDB to the model did not significantly change the predicted list of active pathways. Thus, PUMA accelerates pathway activity analysis by bypassing metabolite annotation as a prior step without compromising the quality of results. Comparing the top metabolite candidate sets generated by PUMA annotation per mass measurement to the candidate metabolites identified by BioCAN and spectral database search in HMDB and METLIN shows a high level of agreement between the annotations. The high level of agreement occurs despite the fact that PUMA does not utilize additional information in form of spectral signatures, as in BioCAN, HMDB and METLIN. Importantly, PUMA improved the annotation results from BioCAN, HMDB and METLIN by identifying candidate sets for measurements that were not previously annotated by other annotation techniques.

Chapter 6
Conclusions and Future Directions

The computational methods presented in this thesis include SelFi, a selection finder for directed evolution of enzymes; EMMA, a metabolite annotation workflow in untargeted metabolomics; and PUMA, a probabilistic predictor of pathway activities and metabolite annotation using metabolomics data. These methods advance the state-of-the-art in computational techniques targeting synthetic biology and metabolic engineering as well as metabolomics data analysis.

6.1 Research Summary

This thesis presented several innovative contributions. SelFi is the first tool to provide an automated way of designing a selection mechanism that isolates a desired enzymatic phenotype. SelFi's contribution is in integrating the synthesis of the selection pathway with knockout identification to couple the selection pathway with cell survival. The results of applying SelFi to identify selection pathways for several target enzymatic products demonstrate that it is possible to synthesize high-quality selection pathways that match in quality to those already confirmed experimentally in the literature. The results also present additional selection strategies that can expand current experimental practices.

The EMMA workflow contributes two key advances. First, filtering the list of possible candidate chemicals through an Expanded Metabolic Model (EMM) specific for the system of interest can eliminate unnecessary and time consuming computations on chemicals that are likely irrelevant to the measured data. When compared to utilizing an *in silico* annotation workflow that utilizes a large database, the results on our datasets demonstrated a reduction in the number of calculations by 4 orders of magnitude. Second, filtering candidate chemicals using an EMM allows for the identification of novel metabolites that are missing from a genome-scale model reconstruction. When compared to using the biological sample's metabolic model, the results on our datasets show that EMMA expands the search space during metabolite annotation. For our datasets, there was a 2.39-fold increase in the number of chemical identities that can be used for annotation, and a 1.71-fold increase in the number of masses than can be annotated. This second advance addresses the need to enable discovery, which is inherently limited in the simpler approach of using a model comprising only metabolites associated with the sample to filter the candidate chemicals, or when using a small biological database. The experimental verification of the presence in the EMM of a CHO cell metabolite (4-hydroxyphenyllactate) that was not previously identified as part of CHO cell metabolism confirmed the utility of the EMMs, and the need to enable discovery beyond a simple metabolic model.

PUMA is the first tool to demonstrate the utility of using probabilistic inference to estimate the likelihood of metabolic pathway activities and metabolite annotation. Our results show the capabilities of PUMA in evaluating pathway activities despite uncertainties in metabolite annotation, and how predictions regarding pathway activities can be utilized to reduce the uncertainties in assigning chemical identities to measurements. The results are substantial as they show a significant increase in the number of annotated mass measurements compared to the number of annotations possible by other state-of-the-art tools. For annotations made by PUMA and other tools that utilize additional information in the form of spectral signatures, there was a high level of agreement in ranking the candidate chemical identities.

6.2 Future Research Directions

SelFi utilized *ProPath* to construct selection pathways using existing reactions within the KEGG database. The target enzymatic molecule is assumed to be available in KEGG. The construction of selection pathways can be extended to engineer pathways that consume product molecules that are not associated with known reactions. Potential transformations of the product molecule can be explored using a tool that predicts outcomes of substrate promiscuity such as *PROXIMAL*. The consecutive application of such a tool can generate novel selection pathways.

The discovery of a novel metabolite in the CHO samples that is not cataloged as part of the CHO metabolic model in the KEGG database shows the utility of constructing EMMs for metabolite annotations. Integrating available metabolite concentration and gene expression data can refine EMM models. For example, low gene expression indicates that the corresponding enzyme is unlikely to act promiscuously and if it did, the metabolite concentration of the resulting product is likely small in comparison of metabolite concentrations of other molecules in the cell. Further, providing a tool that allows the user to automatically create EMM models can streamline the automated construction of EMMs for annotation and other applications.

The experimental validation of EMMA resulted in confirming the identity of only one out of eight predicted metabolites. This could be due to inaccuracies in the rankings by the fragmentation tool. For example, CSI:FingerID reports an accuracy of only 39.5% when annotating a data set from MassBank by searching PubChem. The low confirmation rate can also be due to the assumption that all enzymes are promiscuous. As an enhancement, it is possible to improve *PROXIMAL* to rank the predicted derivatives based on enzyme designations as generalists or specialists [123], on participation in primary or secondary metabolism [125], and other kinetic data available through the BRENDA database [126].

Appendix A

Supplementary Material for Chapter 3

Table 19. Full names of chemical abbreviations in Chapter 3

Chapter 3 abbreviations	Full names of chapter 3 abbreviations
2pg	D-Glycerate-2-phosphate
3pg	3-Phospho-D-glycerate
3php	3-Phosphohydroxypyruvate
AcCoA	Acetyl-CoA
ADP	Adenosine Diphosphate
AKG	2-Oxoglutarate
ara5p	D-Arabinose-5-phosphate
ATP	Adenosine Triphosphate
Co2	Carbon Dioxide
CoA	Coenzyme-A
db4p	3-4-dihydroxy-2-butanone-4-phosphate
dmlz	6-7-Dimethyl-8--1-D-ribityl-lumazine

FUM	Fumarate
g3p	Glyceraldehyde-3-phosphate
H	Hydrogen
ICT	Citrate
kdo2lipid4	KDO-2--lipid-IV-A--with-laurate
kdo8p	3-Deoxy-D-manno-octulosonate-8-phosphate
MAL	Malate
NAD	Nicotinamide-adenine-dinucleotide
NADH	Nicotinamide-adenine-dinucleotide---reduced
OAA	Oxaloacetate
Pi	Phosphate
ribflv	Riboflavin
ru5p-D	D-Ribulose-5-phosphate
SUCC	Succinate
xu5p-D	D-Xylulose-5-phosphate
xyl-D	D-Xylose
xylu-D	D-Xylulose

Bibliography

1. U Nair, N., C. A Denard, and H. Zhao, *Engineering of enzymes for selective catalysis.* Current Organic Chemistry, 2010. **14**(17): p. 1870-1882.
2. Bastian, S., et al., *Engineering of pyranose 2-oxidase from Peniophora gigantea towards improved thermostability and catalytic efficiency.* Applied microbiology and biotechnology, 2005. **67**(5): p. 654-663.
3. Hao, J. and A. Berry, *A thermostable variant of fructose bisphosphate aldolase constructed by directed evolution also shows increased stability in organic solvents.* Protein Engineering Design and Selection, 2004. **17**(9): p. 689-697.
4. Miyazaki, K., et al., *Thermal stabilization of Bacillus subtilis family-11 xylanase by directed evolution.* Journal of Biological Chemistry, 2006. **281**(15): p. 10236-10242.
5. Seng Wong, T., F.H. Arnold, and U. Schwaneberg, *Laboratory evolution of cytochrome P450 BM-3 monooxygenase for organic cosolvents.* Biotechnology and bioengineering, 2004. **85**(3): p. 351-358.
6. You, L. and F. Arnold, *Directed evolution of subtilisin E in Bacillus subtilis to enhance total activity in aqueous dimethylformamide.* Protein Engineering, Design and Selection, 1996. **9**(1): p. 77-83.
7. Bryan, P.N., *Protein engineering of subtilisin.* Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 2000. **1543**(2): p. 203-222.
8. Vasserot, A.P., et al., *Optimization of protein therapeutics by directed evolution.* Drug discovery today, 2003. **8**(3): p. 118-126.
9. Kurtzman, A.L., et al., *Advances in directed protein evolution by recursive genetic recombination: applications to therapeutic proteins.* Current Opinion in Biotechnology, 2001. **12**(4): p. 361-370.

10. Vellard, M., *The enzyme as drug: application of enzymes as pharmaceuticals.* Current Opinion in Biotechnology, 2003. **14**(4): p. 444-450.
11. Delagrave, S. and D.J. Murphy, *In vitro evolution of proteins for drug development.* Assay and drug development technologies, 2003. **1**(1, Supplement 2): p. 187-198.
12. Marshall, S.H., *DNA shuffling: induced molecular breeding to produce new generation long-lasting vaccines.* Biotechnology advances, 2002. **20**(3-4): p. 229-238.
13. Nair N.U., Z.H., *Improving protein function by directed evolution*, in *The metabolic pathway engineering handbook: fundamentals*. 2009, CRC press.
14. Leemhuis, H., R.M. Kelly, and L. Dijkhuizen, *Directed evolution of enzymes: library screening strategies.* IUBMB life, 2009. **61**(3): p. 222-228.
15. Xiao, H., Z. Bao, and H. Zhao, *High throughput screening and selection methods for directed enzyme evolution.* Industrial & engineering chemistry research, 2014. **54**(16): p. 4011-4020.
16. Dietrich, J.A., A.E. McKee, and J.D. Keasling, *High-throughput metabolic engineering: advances in small-molecule screening and selection.* Annual review of biochemistry, 2010. **79**: p. 563-590.
17. Cobb, R.E., R. Chao, and H. Zhao, *Directed evolution: past, present, and future.* AIChE Journal, 2013. **59**(5): p. 1432-1440.
18. Schrimpe-Rutledge, A.C., et al., *Untargeted metabolomics strategies—challenges and emerging directions.* Journal of The American Society for Mass Spectrometry, 2016. **27**(12): p. 1897-1905.
19. Wang-Sattler, R., et al., *Novel biomarkers for pre-diabetes identified by metabolomics.* Molecular systems biology, 2012. **8**(1): p. 615.
20. Suhre, K., et al., *Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting.* PloS one, 2010. **5**(11): p. e13953.
21. Spratlin, J.L., N.J. Serkova, and S.G. Eckhardt, *Clinical applications of metabolomics in oncology: a review.* Clinical cancer research, 2009. **15**(2): p. 431-440.
22. Bogdanov, M., et al., *Metabolomic profiling to develop blood biomarkers for Parkinson's disease.* Brain, 2008. **131**(2): p. 389-396.
23. Jansson, J., et al., *Metabolomics reveals metabolic biomarkers of Crohn's disease.* PloS one, 2009. **4**(7): p. e6386.
24. Wikoff, W.R., et al., *Pharmacometabolomics reveals racial differences in response to atenolol treatment.* PLoS One, 2013. **8**(3): p. e57639.
25. Corona, G., et al., *Pharmaco-metabolomics: An emerging "omics" tool for the personalization of anticancer treatments and identification of new valuable therapeutic targets.* Journal of cellular physiology, 2012. **227**(7): p. 2827-2831.
26. Bundy, J.G., M.P. Davey, and M.R. Viant, *Environmental metabolomics: a critical review and future perspectives.* Metabolomics, 2009. **5**(1): p. 3.

27. Smith, C.A., et al., *METLIN A Metabolite Mass Spectral Database.* Proceedings of the 9Th International Congress of Therapeutic Drug Monitoring & Clinical Toxicology, 2005. **27**: p. 747-751.

28. Wishart, D.S., et al., *HMDB: The human metabolome database.* Nucleic Acids Research, 2007. **35**.

29. Horai, H., et al., *MassBank: a public repository for sharing mass spectral data for life sciences.* J Mass Spectrom, 2010. **45**(7): p. 703-14.

30. Lam, H., et al., *Development and validation of a spectral library searching method for peptide identification from MS/MS.* Proteomics, 2007. **7**(5): p. 655-667.

31. Wolf, S., et al., *In silico fragmentation for computer assisted identification of metabolite mass spectra.* BMC Bioinformatics, 2010. **11**: p. 148.

32. Heinonen, M., et al., *FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data.* Rapid Commun Mass Spectrom, 2008. **22**(19): p. 3043-52.

33. Allen, F., et al., *CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra.* Nucleic Acids Research, 2014. **42**(Web Server issue): p. W94-W99.

34. Dührkop, K., et al., *Searching molecular structure databases with tandem mass spectra using CSI: FingerID.* Proceedings of the National Academy of Sciences, 2015. **112**(41): p. 12580-12585.

35. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

36. Kim, S., et al., *PubChem substance and compound databases.* Nucleic Acids Research, 2016. **44**: p. D1202-D1213.

37. Nobeli, I., A.D. Favia, and J.M. Thornton, *Protein promiscuity and its implications for biotechnology.* Nat Biotechnol, 2009. **27**(2): p. 157-67.

38. Jeffryes, J.G., et al., *MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics.* Journal of cheminformatics, 2015. **7**(1): p. 44.

39. Huan, T., et al., *MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites.* Anal Chem, 2015. **87**(20): p. 10619-26.

40. Moura, M., L. Broadbelt, and K. Tyo, *Computational tools for guided discovery and engineering of metabolic pathways*, in *Systems Metabolic Engineering*. 2013, Springer. p. 123-147.

41. McShan, D. and I. Shah, *Heurstic search for metabolic engineering: de novo synthesis of vanillin.* Computers & chemical engineering, 2005. **29**(3): p. 499-507.

42. McShan, D.C., S. Rao, and I. Shah, *PathMiner: predicting metabolic pathways by heuristic search.* Bioinformatics, 2003. **19**(13): p. 1692-1698.

43. Pharkya, P., A.P. Burgard, and C.D. Maranas, *OptStrain: a computational framework for redesign of microbial production systems.* Genome research, 2004. **14**(11): p. 2367-2376.

44. Yousofshahi, M., K. Lee, and S. Hassoun, *Probabilistic pathway construction.* Metabolic engineering, 2011. **13**(4): p. 435-444.

45. González-Lergier, J., L.J. Broadbelt, and V. Hatzimanikatis, *Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways.* Journal of the American Chemical Society, 2005. **127**(27): p. 9930-9938.
46. Henry, C.S., et al., *Genome-scale thermodynamic analysis of Escherichia coli metabolism.* Biophysical journal, 2006. **90**(4): p. 1453-1461.
47. Moriya, Y., et al., *PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res 38.* 2010.
48. Hattori, M., et al., *Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways.* J Am Chem Soc, 2003. **125**(39): p. 11853-65.
49. Hou, B.K., L.B. Ellis, and L.P. Wackett, *Encoding microbial metabolic logic: predicting biodegradation.* Journal of Industrial Microbiology and Biotechnology, 2004. **31**(6): p. 261-272.
50. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnology and bioengineering, 2003. **84**(6): p. 647-657.
51. Pharkya, P. and C.D. Maranas, *An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems.* Metabolic engineering, 2006. **8**(1): p. 1-13.
52. Kim, J. and J.L. Reed, *OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.* BMC systems biology, 2010. **4**(1): p. 53.
53. Ren, S., B. Zeng, and X. Qian. *Adaptive bi-level programming for optimal gene knockouts for targeted overproduction under phenotypic constraints*. in *BMC bioinformatics*. 2013. BioMed Central.
54. Tepper, N. and T. Shlomi, *Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways.* Bioinformatics, 2009. **26**(4): p. 536-543.
55. Patil, K.R., et al., *Evolutionary programming as a platform for in silico metabolic engineering.* BMC bioinformatics, 2005. **6**(1): p. 308.
56. Yousofshahi, M., et al., *Probabilistic strain optimization under constraint uncertainty.* BMC systems biology, 2013. **7**(1): p. 29.
57. Ranganathan, S., P.F. Suthers, and C.D. Maranas, *OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions.* PLoS computational biology, 2010. **6**(4): p. e1000744.
58. Cotten, C. and J.L. Reed, *Constraint-based strain design using continuous modifications (CosMos) of flux bounds finds new strategies for metabolic engineering.* Biotechnology journal, 2013. **8**(5): p. 595-604.
59. Horai, H., et al., *MassBank: A public repository for sharing mass spectral data for life sciences.* Journal of Mass Spectrometry, 2010. **45**: p. 703-714.

60. *NIST Mass Spectrometry Data Center. Mass Spectrum Interpreter, ver. 2*. 2011; Available from: http://www.chemdata.nist.gov/mass-spc/interpreter/.
61. *METLIN*. Available from: https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage.
62. *MassBank of North America*. Available from: http://mona.fiehnlab.ucdavis.edu/downloads.
63. *HMDB Statistics*. Available from: http://www.hmdb.ca/statistics.
64. *Mass Frontier software, version 7.0*. 2014; Available from: http://highchem.com/index.php/support/.
65. *ACD/MS Fragmenter, version 12*. 2012; Available from: http://www.acdlabs.com/products/adh/ms/ms_frag/.
66. Zhou, J., et al., *HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries.* Bioinformatics, 2014. **30**(4): p. 581-3.
67. Wolf, S., et al., *In silico fragmentation for computer assisted identification of metabolite mass spectra.* BMC Bioinformatics, 2010. **11**: p. 148.
68. Heinonen, M., et al., *FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data.* Rapid Communications in Mass Spectrometry, 2008. **22**(19): p. 3043-3052.
69. Wegner, A., et al., *Fragment formula calculator (FFC): determination of chemical formulas for fragment ions in mass spectrometric data.* Anal Chem, 2014. **86**(4): p. 2221-8.
70. Allen, F., et al., *CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra.* Nucleic Acids Research, 2014. **42**.
71. Dührkop, K., et al., *Searching molecular structure databases with tandem mass spectra using CSI:FingerID.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**: p. 12580-5.
72. Morreel, K., et al., *Systematic structural characterization of metabolites in Arabidopsis via candidate substrate-product pair networks.* The Plant Cell, 2014. **26**(3): p. 929-945.
73. Aguilar-Mogas, A., et al., *iMet: A computational tool for structural annotation of unknown metabolites from tandem mass spectra.* arXiv preprint arXiv:1607.04122, 2016.
74. Alden, N., et al., *Biologically Consistent Annotation of Metabolomics Data.* Analytical chemistry, 2017. **89**(24): p. 13097-13104.
75. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32**(Database issue): p. D277-80.
76. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2008. **36**(Database issue): p. D623-31.
77. Jewison, T., et al., *SMPDB 2.0: big improvements to the Small Molecule Pathway Database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D478-84.

78. Goeman, J.J., et al., *A global test for groups of genes: testing association with a clinical outcome.* Bioinformatics, 2004. **20**(1): p. 93-9.

79. Xia, J. and D.S. Wishart, *MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W71-7.

80. Xia, J., et al., *MetaboAnalyst 3.0--making metabolomics more meaningful.* Nucleic Acids Res, 2015. **43**(W1): p. W251-7.

81. Kamburov, A., et al., *Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA.* Bioinformatics, 2011. **27**(20): p. 2917-8.

82. Marco-Ramell, A., et al., *Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data.* BMC bioinformatics, 2018. **19**(1): p. 1.

83. Li, S., et al., *Predicting network activity from high throughput metabolomics.* PLoS Comput Biol, 2013. **9**(7): p. e1003123.

84. Werpy, T., et al., *Top value added chemicals from biomass. Volume 1-Results of screening for potential candidates from sugars and synthesis gas*. 2004, Department of Energy Washington DC.

85. Varma, A. and B.O. Palsson, *Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.* Applied and environmental microbiology, 1994. **60**(10): p. 3724-3731.

86. Galanie, S., et al., *Complete biosynthesis of opioids in yeast.* Science, 2015. **349**(6252): p. 1095-1100.

87. Alper, H., et al., *Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli.* Metabolic engineering, 2005. **7**(3): p. 155-164.

88. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Molecular systems biology, 2007. **3**(1): p. 121.

89. Nair, N.U. and H. Zhao, *Evolution in Reverse: Engineering a D-Xylose-Specific Xylose Reductase.* ChemBioChem, 2008. **9**(8): p. 1213-1215.

90. Cai, Z., et al., *Development of an activity-directed selection system enabled significant improvement of the carboxylation efficiency of Rubisco.* Protein & cell, 2014. **5**(7): p. 552-562.

91. Basch, H., et al., *Mechanism of the methane→ methanol conversion reaction catalyzed by methane monooxygenase: a density functional study.* Journal of the American Chemical Society, 1999. **121**(31): p. 7249-7256.

92. Chen, M.M., P.S. Coelho, and F.H. Arnold, *Utilizing Terminal Oxidants to Achieve P450-Catalyzed Oxidation of Methane.* Advanced Synthesis & Catalysis, 2012. **354**(6): p. 964-968.

93. Chadwick, S.S., *Ullmann's encyclopedia of industrial chemistry.* Reference Services Review, 1988. **16**(4): p. 31-34.

94. Yousofshahi, M., et al., *MC^3: a steady-state model and constraint consistency checker for biochemical networks.* BMC systems biology, 2013. **7**(1): p. 129.

95. Sanchez, J.C., et al., *Activation of a cryptic gene encoding a kinase for L-xylulose opens a new pathway for the utilization of L-lyxose by Escherichia coli.* Journal of Biological Chemistry, 1994. **269**(47): p. 29665-29669.

96. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nature Reviews Microbiology, 2009. **7**(2): p. 129-143.

97. Schellenberger, J., et al., *BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions.* BMC bioinformatics, 2010. **11**(1): p. 213.

98. Monk, J., J. Nogales, and B.O. Palsson, *Optimizing genome-scale network reconstructions.* Nature biotechnology, 2014. **32**(5): p. 447.

99. D'Ari, R. and J. Casadesus, *Underground metabolism.* Bioessays, 1998. **20**(2): p. 181-6.

100. Tawfik, O.K. and S. Dan, *Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective.* Annual Review of Biochemistry, 2010.

101. Khersonsky, O., et al., *Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions.* Biochemistry, 2011. **50**(13): p. 2683-2690.

102. Pence, H.E. and A. Williams, *ChemSpider: An online chemical information resource*, in *Journal of Chemical Education*. 2010. p. 1123-1124.

103. Yousofshahi, M., et al., *PROXIMAL: a method for Prediction of Xenobiotic Metabolism.* BMC systems biology, 2015. **9**: p. 94.

104. Oh, M., et al., *Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways.* J Chem Inf Model, 2007. **47**(4): p. 1702-12.

105. Hassanpour, N., et al., *Using Biological Filtering and Enzyme Promiscuity Prediction to Advance Annotation in Untargeted Metabolomics.* (in preparation), 2018.

106. Sridharan, G.V., et al., *Prediction and quantification of bioactive microbiota metabolites in the mouse gut.* Nat Commun, 2014. **5**: p. 5492.

107. May, J.C. and J.A. McLean, *Advanced Multidimensional Separations in Mass Spectrometry: Navigating the Big Data Deluge.* Annual Review of Analytical Chemistry, 2016. **9**: p. 387-409.

108. Hult, K. and P. Berglund, *Enzyme promiscuity: mechanism and applications.* Trends Biotechnol, 2007. **25**(5): p. 231-8.

109. Cho, A., et al., *Prediction of novel synthetic pathways for the production of desired chemicals.* BMC Syst Biol, 2010. **4**: p. 35.

110. Campodonico, M.A., et al., *Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path.* Metab Eng, 2014. **25**: p. 140-58.

111. Li, C., et al., *Computational discovery of biochemical routes to specialty chemicals.* Chemical Engineering Science, 2004. **59**(22-23): p. 5051-5060.

112. Finley, S.D., L.J. Broadbelt, and V. Hatzimanikatis, *Computational framework for predictive biodegradation.* Biotechnol Bioeng, 2009. **104**(6): p. 1086-97.

113. Jordan, M.I., *Learning in graphical models*. Vol. 89. 1998: Springer Science & Business Media.

114. Gelman, A., et al., *Chapter 11: Basics of Markov chain simulation*, in *Bayesian data analysis*. 2014, CRC press Boca Raton, FL.

115. Gelman, A., et al., *Basics of Markov chain simulation*, in *Bayesian data analysis*. 2014, CRC press Boca Raton, FL.

116. Koller, D. and N. Friedman, *Probabilistic graphical models: principles and techniques*. 2009: MIT press.

117. Salvatier, J., T.V. Wiecki, and C. Fonnesbeck, *Probabilistic programming in Python using PyMC3.* PeerJ Computer Science, 2016. **2**: p. e55.

118. Yildirim, I., *Bayesian inference: Gibbs sampling.* Technical Note, University of Rochester, 2012.

119. Casella, G. and E.I. George, *Explaining the Gibbs sampler.* The American Statistician, 1992. **46**(3): p. 167-174.

120. Rogers, S., et al., *Probabilistic assignment of formulas to mass peaks in metabolomics experiments.* Bioinformatics, 2008. **25**(4): p. 512-518.

121. Silva, R.R., et al., *ProbMetab: an R package for Bayesian probabilistic annotation of LC–MS-based metabolomics.* Bioinformatics, 2014. **30**(9): p. 1336-1337.

122. Jeong, J., et al., *An empirical Bayes model using a competition score for metabolite identification in gas chromatography mass spectrometry.* BMC bioinformatics, 2011. **12**(1): p. 392.

123. Wang, M., et al., *Topic model-based mass spectrometric data analysis in cancer biomarker discovery studies.* BMC genomics, 2016. **17**(4): p. 545.

124. Breitling, R., A.R. Pitt, and M.P. Barrett, *Precision mapping of the metabolome.* Trends in biotechnology, 2006. **24**(12): p. 543-548.

125. Bar-Even, A. and D.S. Tawfik, *Engineering specialized metabolic pathways—is there a room for enzyme improvements?* Current opinion in biotechnology, 2013. **24**(2): p. 310-319.

126. Schomburg, I., et al., *BRENDA, the enzyme database: updates and major new developments.* Nucleic acids research, 2004. **32**(suppl_1): p. D431-D433.