Computational Methods for Pathway Synthesis and Strain Optimization

A dissertation

submitted by

Mona Yousofshahi, B. Sc., M.Sc.

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

February 2015

Advisors: Professor Kyongbum Lee,

and Professor Soha Hassoun

To my family

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Soha Hassoun, for her great advice and continuous encouragement throughout the course of my Ph.D. studies. She provided me with a rich and fertile environment to study and explore new ideas and supported me both scientifically and intellectually.

I would like to thank Prof. Kyongbum Lee for co-advising my work, during which I really enjoyed his calm and warm demeanor. His technical and editorial advice was essential to the completion of this dissertation.

Besides my advisors, I would like to thank the rest of my thesis committee members, Profs. Roni Khardon, Kathleen Fisher and Keith E.J. Tyo, for attending my Ph.D. defense, reading my dissertation and providing me with pertinent comments. I would also like to thank all my collaborators and colleagues at Tufts.

This journey would not have been possible without the support of my family. I thank my parents for encouraging me in all of my pursuits and inspiring me to follow my dreams. Finally, I thank my husband and best friend, Pooya, for his incredible support during my studies.

Computational Methods for Pathway Synthesis and Strain Optimization

Engineering and optimization of biological cells have been central to modern biotechnology, with applications ranging from drug discovery and development to production of commercially significant chemicals. Purely empirical approaches can benefit greatly when paired with computer-aided design methods that allow for design space exploration and optimization. Such methods may significantly contribute to reducing experimental efforts and expediting discoveries.

This thesis addresses two problems in metabolic engineering and synthetic biology. The first problem concerns the construction of synthetic pathways to produce a desired metabolite within a microbial cell. We present two approaches for solving this problem. The first approach, *ProbPath*, identifies non-native *de novo* synthesis pathways from reactions within a database by probabilistically sampling available reactions. *ProbPath* is shown effective in identifying synthesis pathways when compared to exhaustive exploration of the design space with limited path length in terms of generating similar yield profiles. Additionally, we were able with *ProbPath* to reproduce routes that were experimentally obtained for the production of several molecules. The second approach addresses the issue when a desired target metabolite is not present in known databases. To produce a synthesis pathway or such a metabolite, we develop a novel methodology based on identifying structural similarities between the target metabolite and existing metabolites within the database and developing transformation operators that predict the transformation outcome when applied to the target metabolite. To study this approach, we developed an algorithm, PROXIMAL, to construct transformation operators based on the set of xenobiotic transformations associated with human liver enzymes. We evaluated the prediction accuracy of PROXIMAL through case studies on two environmental chemicals. Comparisons with published reports confirm that our predictions have been experimentally validated in the literature.

The second problem addressed in this thesis concerns identifying optimal gene modifications when tuning a microbial cell to maximize the production of a desired compound. The novelty in our problem formulation lies in explicitly accounting for likely variations in flux capacities due to engineering modifications. The thesis presents a computational framework, CCOpt, which identifies an optimal set of gene modifications. CCOpt is based on chance-constrained programming, where constraints are probabilistically met at a user-specified confidence level. Evaluation of the approach demonstrates that CCOpt consistently finds a solution most-frequently found when using Monte Carlo sampling, but at a fraction of a computational cost. The CCOpt formulation is the first work to incorporate uncertainties when computing gene modifications.

Overall, the thesis contributes to and advances the state-of-the-art in design automation tools for metabolic engineering and synthetic biology.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Metabolic engineering of microbial hosts has shown promise in the production or overproduction of commercially useful biomolecule, including polyesters [1], building blocks for industrial polymers [2], biofuels [3], and therapeutic natural products derived from isoprenoids [4, 5, 6], polyketides [7, 8], and non-ribosomal peptides [6]. One successful example is microbially produced artemisnic acid as a viable source of antimalarial drugs [9]. Keasling's lab, which began to look at alternative sources for producing artemisinin in 2001, isolated genes that produce artemisinic acid from *Artemisia annua* (a common type of wormwood) and transferred them into yeast. To improve yield, productivity, and selectivity of artemisinic acid, metabolite flux was redirected by up- and down-regulating several endogenous yeast genes. Yeast was chosen over *Escherichia coli* because the latter was deemed incompatible with expression of plant-derived genes. The outcome, which took almost two decades to realize, allows for cheaper treatments (from $2.40 to $0.25 per dose) for malaria that annually threatens 300-500 million people and kills more than one million people [10]. Large-scale commercial production commenced in April 2013 through a licensed process to Sanofi & PATH that plans to manufacture and distribute artemisinin at-cost [11].

While this example highlights the potential of microbial biosynthesis as a production process, it also illustrates the challenges in synthesis pathway development process: A series of non-native genes must be identified and transferred into a genetically compatible host. The host must then be tuned to maximize the production of the non-native desired product. The example also highlights the effort, cost and time required for biological discoveries and for engineering synthetic biological systems.

While experimental efforts have yielded successes, it is becoming clear that tapping the full potential of biological engineering will require advances in computational methods to support and expedite the design cycle. Computing is poised to play a critical role for several reasons. First, Moore's law has predicted since the 1970s that transistor density doubles every two years, and has led to exponential increases in computational power [12, 13]. Computer memory density has exhibited similar exponential growth [14]. With increased computing capabilities, design variable search and optimization capabilities that were prohibitive a decade ago are now possible. Second, biological data is now catalogued electronically in databases and made available via web access. One example is the KEGG database [15, 16, 17], which catalogues compounds, reactions, biological pathways and modules. Another example is Ecocyc [18], a database specific to *E. coli* providing genomic information along with metabolic pathways and regulatory networks. The availability of such data has already had significant impact, enabling genome-reconstruction of metabolic networks (e.g. [19, 20]). Third, advances in computational techniques, including algorithms [21] and machine learning [22], can be adapted to enable the design and optimization of biological systems.

This thesis addresses several computational techniques that promise to expedite the design and tuning of synthetic systems. The thesis develops computational approaches for the identification and ranking of biosynthesis pathways, and for tuning the microbial hosts to achieve high yields. To solve these problems, we develop novel algorithms, and employ well-established chance-constrained optimization technique. We also utilize knowledge catalogued in databases to inform our techniques.

1.1 Computational Challenges in Metabolic Engineering

Several issues challenge the development of computational methods that aid and guide the design of biological systems. These challenges can be broadly classified into two categories: model development and design-space exploration challenges. We describe each challenge and highlight one or two recent works that have shown great promise in addressing these challenges.

1.1.1 Modeling Challenges

Stoichiometric models that capture the steady-state constraints of a metabolic network are traditionally constructed manually, based on earlier models and in combination with reaction availability from databases. The recent availability of genome, reaction, and organism specific databases has allowed for the automatic reconstruction of genome-scale models [23, 24]. The quality of the reconstruction, whether obtained manually or automatically, is as comprehensive as the availability of reconstruction and experimental data, and is a function of the reconstruction procedure. The resulting models may thus be incomplete or inconsistent for the purpose of steady-state analysis. A standalone tool that addresses model inconsistencies is *Model & Constraint Consistency Checker* (MC^3) [25], which has been developed to perform model and constraint consistency checking. MC^3 identifies model inconsistencies in the context of steady-state analysis.

Dynamic models represent changes in the metabolic concentration and reaction rates over time in a metabolic network. After determining kinetic rate expressions based on either mechanistic knowledge [26] or canonical rate expressions [27, 28, 29, 30, 31], kinetic parameters must be fitted based on experimental measurements. Parameter estimation methods minimize the error between model predictions and experimental data. Simulated annealing (SA) is a global optimization technique used in parameter estimation. This technique creates initial estimates of parameters and then iteratively refines the estimates [32]. SA is a stochastic method; it does not guarantee a global optimum. Recently, a new approach has been developed to avoid the traditional "best-fit" model. Ensemble kinetic modeling builds an collection of allowable dynamic models that reach the same steady state, and avoids the detailed characterization of kinetic parameters [33].

To enable cellular engineering to fully exploit computational advances, computational models must capture uncertainties inherent to biological systems due in part to stochastic fluctuations of molecular processes. The translation and transcription machinery expressing genes fundamentally undergo a stochastic process [34], resulting in cell-to-cell variations in mRNA and protein levels and phenotype variations across a wide range of organisms [35]. Discrete and stochastic models have been proposed based on a probabilistic interpretation of reaction events. For example, Gillespie's stochastic simulation algorithm (SSA) considers discrete amounts of molecules, and joint probability distributions are used to express the likelihood a particular cellular concentration at a given time [36]. While a number of modifications have been proposed to capture increasing levels of detail such as time delays in protein synthesis and burst increases in mRNA production [37], the impact of stochastic fluctuations in gene expression is not routinely utilized when analyzing complex cellular metabolism mainly due to lack of consistent data sets and methods to use the data.

1.1.2 Design-Space Exploration Challenges

Some computational problems in Metabolic Engineering are inherently difficult. One example problem is Elementary Flux Mode (EFM) analysis [38, 39], used to find all non-decomposable thermodynamically feasible pathways that can operate under steady-state conditions. EFM analysis has proven useful in analyzing robustness and regulation [38, 40, 41], microbial stress responses [42], product yield [43], and in assessing carbon conversion efficiency in plants [44]. While useful, enumerating all pathways is intractable, and EFM analysis is not possible for even moderate-size reconstructions. Alternate decomposition methods have emerged including sampling-based techniques [45] and finding pathways within sub-networks [46].

Computational intractability also manifests itself when optimizing design variables. With a large design space, exhaustive analysis is computationally prohibitive. Consider the problem of identifying optimal gene knockouts within a cellular network to maximize the yield of a desired metabolite. The feasibility of evaluating all possible knockouts individually, in pairs, and in groups of 3 or more requires evaluating the resulting cellular flux 2^n times, where n is the number of genes within the network. In OptKnock [47], this problem is formulated as a bi-level optimization problem, and solved using Mixed Integer Linear Programming. The runtime needed to solve this optimization problem grows exponentially with the network size [48]. While runtime is often cited as prohibitive, memory use proves to also be prohibitive. A problem arises when there are multiple options to choose from at each step of the computation, leading to exponential growth in the number of candidate solutions. Such exponential growth will eventually strain available memory resources and easily extend beyond reasonable processing times, severely limiting the scope of resulting solutions. An example problem is the pathway identification problem that will be addressed in this thesis.

While exploring the design space, it is also important to address uncertainties in engineering interventions. Consider the problem of modifying gene expression levels. Even the most skilled biological engineer is unable to implement an exact fold change. A nave way of accounting for such uncertainties is band-guarding, where worse case assumptions about the variations are assumed, and design decisions are made conservatively. Such approaches often lead to “over-design”, and may not lead to the best design options. Computational methods to effectively address biological and engineering uncertainties at the design stage will become increasingly important as biological engineering efforts progress from proof-of-principle to scaled-up manufacturing.

1.2 Thesis Overview

This thesis addresses two problems: identification of *de novo* synthesis pathways and optimal tuning of microbial hosts. The following sections describe these problems in detail.

1.2.1 Identification of *De Novo* Synthesis Pathways

Two different approaches are available to assemble biosynthesis pathways. The first involves combining partial pathways from various organisms and implanting them within a microbial host, as was demonstrated in producing artemisinic acid [9]. A second approach removes the restriction of utilizing known pathways, and constructs *de novo*synthesis pathways. Here, a series of reactions are assembled, one-by-one and from various organisms, to create a pathway capable of producing a target metabolite from an existing source within the host.

In some cases, a choice for the *de novo* pathway may be obvious. For example, there is only one known pathway for biosynthesis of 1,3-propanediol from glycerol [2]. This pathway consists of two reactions, each catalyzed by a singular enzyme. More generally, the number of alternative pathways for a given product may be too large for experimental exploration, especially if the goal is to exploit the diversity of metabolic enzymes across many different organisms. To date, more than 1000 prokaryotic genomes have been fully sequenced and annotated. Partial or draft genomes are available for more than 6000 species. The total number of reactions listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [15, 17, 16] currently exceeds 8000. In this light, computational approaches are warranted to analyze the growing number of possible metabolic and biosynthetic enzyme combinations as candidate pathways for heterologous synthesis of biomolecules.

Due to the combinatorial nature of the problem, an exhaustive search for candidate pathways is impractical. To routinely analyze ever-growing and continuously updated genome-scale databases, an effective search strategy needs to address several issues. Enzymes need to be selected from a large, multi-organism database such as KEGG [15, 17, 16], MetaCyc [49] or SEED [50] to form a logical reaction sequence, mapping the final product molecule to one or more reactant metabolites in the host organism. This selection process needs to take into account not only the main reactants, but also reducing equivalents and other cofactors. In the likely event that a large number of candidate pathways have been identified, the computational analysis needs to evaluate these pathways based on a performance metric such as maximal predicted yield. The evaluation needs to also assess whether the introduction of the synthesis pathways will negatively impact the host organism's capacity for balanced growth [51].

Over the last several years, a number of heuristic approaches have been developed to predict novel pathways for degradation of xenobiotics [52] or biosynthesis of native and non-native compounds [53, 52, 54, 55]. One such approach, PathMiner, seeks to build pathways that minimize the biochemical transformation cost [53]. This heuristic favors reactions involving the addition of smaller functional groups, which can select against canonical modifications such as phosphorylation. PathPred is another method to construct plausible reaction pathways based on chemical structure transformation patterns of small molecules [52]. PathPred specifically exploits the KEGG RPAIR database, which contains transformation patterns for substrate-product pairs (reactant pairs) of known enzymatic reactions.

The patterns are described by atom type changes at the reaction center atom and its neighboring atoms. A key advantage of PathPred is that it generates plausible pathways even when no matching compound is found for the queried molecule by utilizing pattern matches reflecting generalized reactions shared among structurally related compounds. The drawback is that the patterns need to be manually cured. OptStrain uses mixed integer programming to identify stoichiometrically balanced pathways by adding or deleting reactions to the host metabolic network [54]. A key advantage of this approach is to couple the selection of reactions with the ranking of the synthesis pathways in terms of theoretical yields. Success of the optimization however critically depends on thoroughly pre-processing the database, which remains a non-trivial task.

There currently is a lack of data and consensus on the best pathway scoring methods. The number of pathway steps does not necessarily correlate with yield or the implementation practicality [56]. Another metric for ranking the non-native pathway is metabolic burden which computes the reduction in the growth rate as a result of added reactions [57]. Another ranking strategy is the thermodynamic feasibility which tries to compute the change in the Gibbs free energy of the reaction along the pathways by using a group contribution method [58].

To overcome shortcomings of above prior approaches, this thesis develops a graph-based probabilistic algorithm, *ProbPath*, for identifying viable synthesis pathways compatible with balanced cell growth. Rather than exhaustive exploration, our approach applies probabilistic selection of reactions to construct the pathways. The algorithm considers not only the main reactants, but also the cofactors needed for the biosynthesis. *ProbPath* ranks identified pathways based on yield. We apply the algorithm to synthesize pathways producing commercially useful targets using E. coli as the host organism.

ProbPath proved effective in constructing synthesis pathways involving metabolites and reactions already catalogued in existing databases. However, some desired metabolites are not found in databases. For example, 3-Hydroxybutyrolactone (3-HBL), a key intermediate for various pharmaceuticals and a desired chemical for producing renewable energies [59], is not catalogued in the KEGG database. To address this problem, we take advantage of enzyme catalytic promiscuity defined as the potential to catalyze physiologically irrelevant reactions [60, 61, 62]. Transformations catalyzed by a reaction can be catalogued, and then applied judiciously to query molecules. We present a computational method to predict possible biotransformations of a target metabolite through a set of enzymes specified by the user using molecular substructural similarity between the target metabolite and all substrates of the enzymes. More specifically, we develop a method, PROXIMAL, that computes xenobiotic transformation pathways through human liver enzymes. The methodology used in PROXIMAL can be generalized to other sets of enzymes.

1.2.2 Optimal Tuning of Microbial Hosts

In recent years, increasingly sophisticated computational methods have been developed to identify optimal genetic modifications to achieve a desired metabolic engineering objective. The problem of identifying optimal genetic modifications can be expressed in terms of operating state variables such as reaction flux, and control (decision) variables such as the presence or absence of gene expression. The optimal design “tunes” these variables such that the solution meets the engineering objective while satisfying several constraints reflecting physico-chemical considerations, experimental observations and assumptions about the physiology of the cell or organism. Due to biological variability [63, 64], stochastic effects associated with gene expression, and imprecision in engineering implementation, it is questionable that enzyme levels can be precisely tuned to exactly match the target values calculated using computational design tools. More likely, the target enzyme levels, and thus the corresponding reaction flux capacities, can only be achieved with a finite degree of uncertainty. Addressing uncertainty at the design stage is a challenging issue that has become increasingly important not only for engineering biological systems, but also man-made systems such as electronic devices. Indeed, the past decade has witnessed a paradigm shift in design of electronics and computational design tools, where all modern electronic circuits are now designed to maximize tolerance to manufacturing and operational variations or to include tuning circuitry for post-manufacturing re-calibration. As metabolic engineering efforts progress from proof-of-principle to scaled-up manufacturing, computational methods to effectively address biological and engineering uncertainties at the design stage will become increasingly important in ensuring the identification of the most robustly optimal gene modifications.

The uncertainty in achieving targeted enzyme values suggests that the enzyme levels, and hence the corresponding flux carrying capacities (bounds), could be considered statistical distributions rather than fixed value parameters. In this statistical interpretation, a flux constraint in a conventional deterministic optimization problem represents the most conservative point in the flux capacity distribution, since a deterministic problem enforces all constraints with zero uncertainty.

Although the deterministic approach affords relatively straightforward problem formulation and is most commonly practiced [65, 47, 48], this approach might lead to choosing an intervention set that may be far from optimal in a statistical sense. Alternatively, a sampling-based optimization approach (e.g. Monte Carlo sampling [66]), with the obvious caveat of being computationally intensive, probabilistically explores a possible space of enzyme activities, *i.e.* flux capacity distributions, and solves for an optimal intervention set for each sam- pled instance of flux capacities. Repeated sampling produces multiple intervention sets and a corresponding distribution of objective function values. Another alternative for incorporating uncertainties in an optimization problem is chance-constrained programming (CCP), which selects an optimal solution with a user-defined degree of probabilistic confidence in meeting constraints. Chance-constrained programming was first introduced in [67] to solve the problem of temporal planning when uncertainty is present. Since then, CCP has been utilized in numerous applications, including circuit sizing [68], soil conservation [69], ground water management [70], energy management [71], and molecular property optimization [72].

Current strain optimization methods generally seek to identify combinations of gene-level modifications that will result in an improvement of the desired cellular objective. These modifications are commonly gene deletions, but may be also up- or down-regulations of gene expression. A notable example of a computational method to identify gene knockouts is OptKnock [47]. This method uses bi-level programming to identify gene deletions that satisfy the coupled objectives of metabolite overproduction and biomass formation. Another gene deletion strategy is Genetic Design through Local Search (GLDS) [48], which employs a heuristic and flux balance analysis (FBA) to iteratively find sets of zero flux reactions (corresponding to gene deletions) that would result in the maximization of the target reaction flux. Other, related methods for large-scale problems involve metaheuristic approaches to iteratively improve a candidate set of gene deletions by generating and selecting variants of the candidate set via assessment of the objective function. An example of this approach is OptGene, which uses an evolutionary algorithm to improve the set of gene deletions with respect to an objective function [73].

Optimization methods have also been described to identify targets for gene expression modification. OptReg [65] is a constraint-based method that uses bi-level programming to determine which sets of genes should be amplified or down-regulated to satisfy a coupled pair of engineering and cellular objectives. Another class of computational strain design methods utilizes elementary mode (EM) analysis. One recent example is Computational Approach for Strain Optimization aiming at high Productivity (CASOP), which ranks reactions based on their contributions to the yield of desired product [74]. Another example is Flux Design, which selects reactions for up-regulation or deletion based on their correlation with the objective flux computed from EMs that contribute to the target product [75, 76]. Despite increasing sophistication, these and other current computational strain design methods implicitly assume that reaction flux changes can be implemented precisely, and thus do not consider uncertainties as part of the problem formulation.

To address this problem, we propose a computational method, CCOpt, to optimize the selection of an intervention set that consists of gene up/down-regulation using uncertainty-aware chance-constrained optimization. CCOpt operates on both kinetic and stoichiometric models and depending on the model type, it approximates the distributions of flux capacities either using maximal reaction velocities or by analyzing the enzyme control flux (ECF) [77]. In contrast to deterministic approaches where constraints are met with 100% certainty, constraints in CCOpt are probabilistically met at a user-specified confidence level. We investigate the application of CCOpt to two case studies that utilize the Chinese Hamster Ovary (CHO) cell metabolism. Our results demonstrate that CCOpt is capable of identifying optimal intervention sets without the run-time cost of a sampling based (Monte Carlo) approach.

1.3 Thesis Contributions

This dissertation contributes computationally to the area of metabolic engineering with the goal of producing valuable molecules from microorganisms. The key contributions are:

- Developing a probabilistic search algorithm (*ProbPath*) to construct non-native, viable synthesis pathways.
- Comparing *ProbPath* against an exhaustive search of limited depth and demonstrating that *ProbPath* identifies synthesis pathways with yield profiles similar to those identified using exhaustive search at a fraction of the runtime.
- *ProbPath* reproduces experimentally obtained pathways reported in the literature, and predicts comparable maximum yields.
- Developing a generic methodology for predicting biotransformation pathways for target metabolites not catalogued in databases.

- Implementing PROXIMAL, a method for predicting xenobiotic transformations through human liver enzymes.
- Using PROXIMAL to identify derivatives of two environmental chemicals, bisphenol A (BPA) and 4-chlorobiphenyl (PCB3). Several identified molecules were previously reported in the literature as derivatives.
- Creating an optimization framework, CCOpt, that addresses engineering intervention uncertainties when identifying gene modifications to maximize cellular yield.
- Comparing CCOpt with other approaches that do not address uncertainties, and demonstrating that CCOpt identifies a more diverse set of interventions.
- Comparing CCOpt with Monte Carlo simulations, and showing that maximal fluxes predicted by CCOpt are always in the probable range identified by Monte Carlo simulations.

1.4 Thesis Organizations

This dissertation consists of five chapters. Chapter 2 describes a pathway construction algorithm, *ProbPath*, for identifying non-native viable synthesis pathways to overproduce a target metabolite. Chapter 3 presents a new algorithm for identifying possible biotransformation pathways when the target metabolite is not catalogued in the database. A method, PROXIMAL, is implemented to predict xenobiotic biotransformations through human enzymes. Chapter 4 describes an optimization method, CCOpt, that identifies reactions for flux capacity modifications to increase the production rate of a target metabolite. This method incorporates engineering uncertainties in the optimization problem using chance-constrained programming.

Chapter 5 summarizes the thesis and outlines directions for future research.

Chapter 2

Probabilistic Pathway Construction

In this chapter, we present a novel method for constructing synthesis pathways using a graph-based probabilistic-search approach. Our approach is based on searching the KEGG database for pathways and using flux balance analysis (FBA) [78] to rank the constructed pathways. The main challenge in this approach is to avoid exhaustive enumeration of all possible pathways as it yields an intractable number of pathways. Our algorithm, *ProbPath*, resembles backtracking algorithms that recursively explore the search space until a solution is identified [79]. Backtracking algorithms visit every possible combination in the solution space by employing a depth-first search scheme to build a solution incrementally. Starting from an initial point in the search space, a backtracking algorithm makes a choice at each step to expand the partial solution. If the partial solution cannot be expanded to a full solution (for example, if a predefined threshold is met), the algorithm backtracks to an earlier step and tries an alternative choice. The algorithm ends either when it reaches a solution or when the search space is explored and no feasible solution is found. Backtracking algorithms are useful in solving problems for which efficient methods are not available. One disadvantage of the standard backtracking algorithm is repeated failure due to the try-and-error nature of the algorithm [80]. Another drawback of the backtracking algorithm is performing redundant work by not remembering the conflicts that have already been observed during backtracking [80]. Unlike standard backtracking algorithms, *ProbPath* uses random sampling to explore the search space. At each step, *ProbPath* makes a random choice guided by one of its weighting schemes to extend the partial solution. *ProbPath* remembers failure routes and avoids them in subsequent steps.

2.1 Methods

2.1.1 Pathway Construction

We develop a graph-based, probabilistic search technique of the KEGG database, called *ProbPath*, to identify *non-native* synthesis pathways for a given product metabolite. We define a non-native synthesis pathway as a sequence of non-native reactions beginning with any native metabolite and ending with the specified product metabolite. The product metabolite may or may not be a native metabolite. Pathways are constructed as a graph, specifically a tree, by adding metabolite nodes and reaction edges selected from the KEGG database. The KEGG database was chosen for its breadth of coverage of metabolic pathways across many organisms.

Tree construction proceeds recursively, starting from the target metabolite, *i.e.* synthesis product, as the root of the tree Figure 2.1. A single reaction is selected

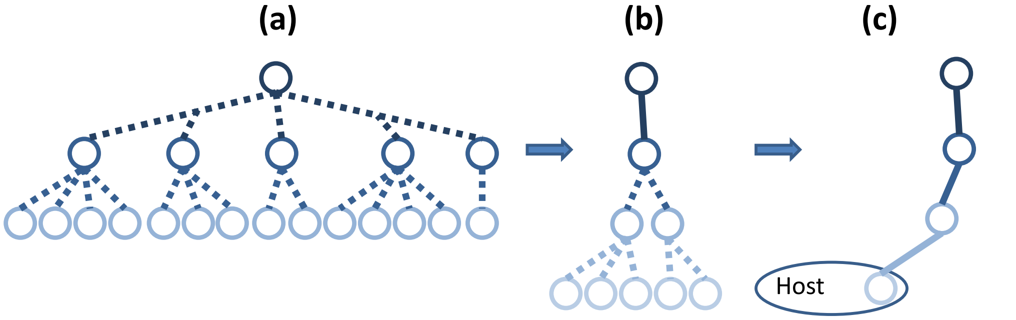
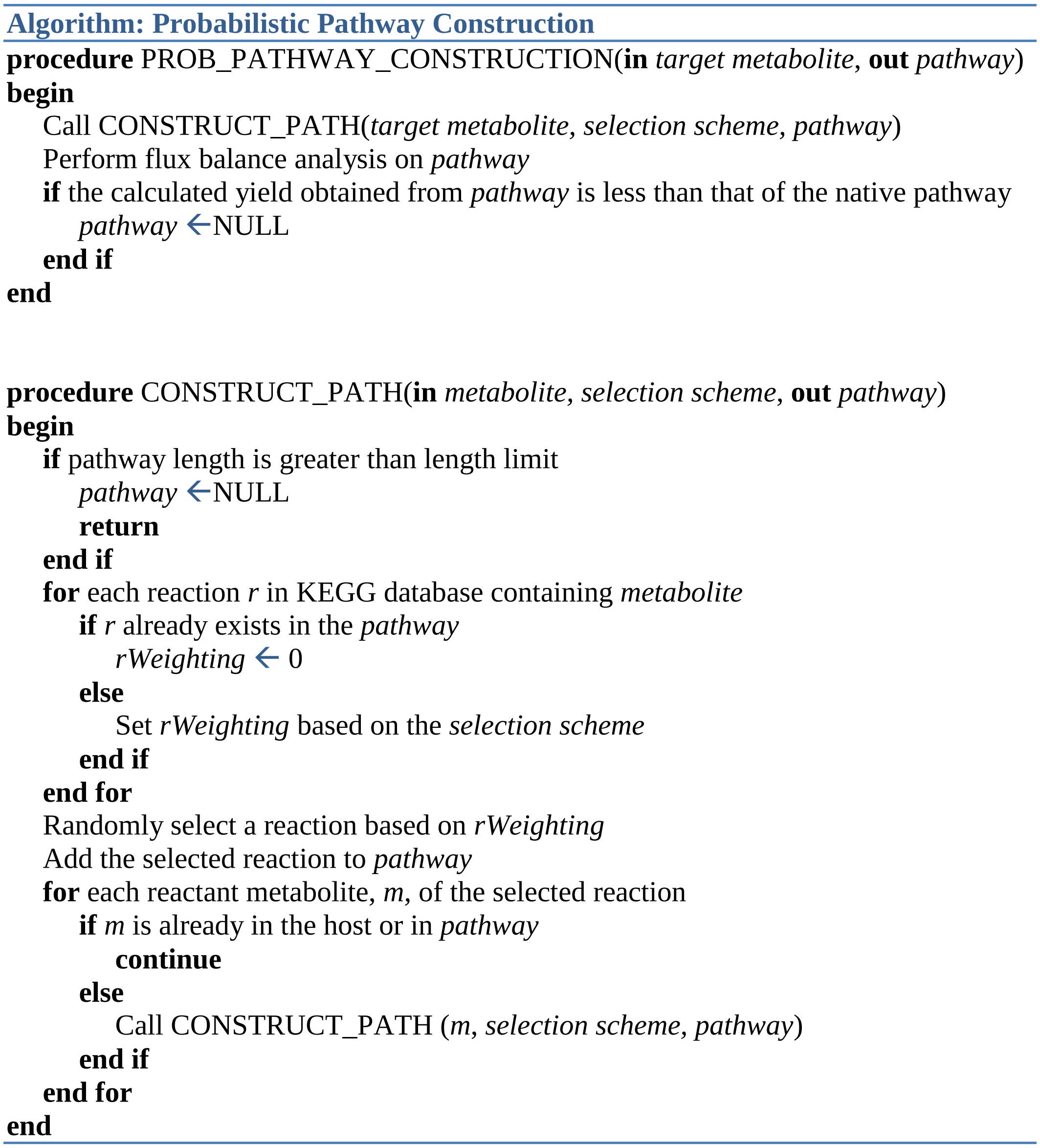


Figure 2.1: Schematic illustration of the probabilistic search. The dashed and solid lines show the possible routes and selected reactions, respectively. The tree expansion terminates when a metabolite found that is native to the host network: (a) all possible reaction choices to generate a target metabolite, two reactions away from the target; (b) only one reaction is explored in depth-first fashion and (c) recursive exploration terminates at a metabolite within the host network.

from a list of candidate reactions in the KEGG database that involve the target metabolite as a main product. Selection occurs probabilistically based on a weighting scheme determined by the connectivity of the candidate reactions' metabolites (see Section 2.1.2). The type of selection scheme is passed to the algorithm as a free parameter. The selected reaction is then added to the tree and represented by an edge. This edge expands the tree by attaching new nodes representing the reactant metabolites and cofactors of the selected reaction. The construction thus proceeds in a depth-first fashion. Each of these nodes is a new root for the recursion, unless the corresponding metabolite or cofactor is already present in the host organism or was previously added to the tree. Details of the algorithm are provided in Figure 2.2.

Because there is a practical limit to the number of heterologous genes that can be inserted into a typical host organism such as *E. coli* [7], we set a limit on the number of reactions that can be used to construct a pathway. The length limit is thus used to obtain candidate pathways of practical length, rather than to rank-



Algorithm: Probabilistic Pathway Construction

```
procedure PROB_PATHWAY_CONSTRUCTION(in target metabolite, out pathway)
begin
    Call CONSTRUCT_PATH(target metabolite, selection scheme, pathway)
    Perform flux balance analysis on pathway
    if the calculated yield obtained from pathway is less than that of the native pathway
        pathway ← NULL
    end if
end


procedure CONSTRUCT_PATH(in metabolite, selection scheme, out pathway)
begin
    if pathway length is greater than length limit
        pathway ← NULL
        return
    end if
    for each reaction r in KEGG database containing metabolite
        if r already exists in the pathway
            rWeighting ← 0
        else
            Set rWeighting based on the selection scheme
        end if
    end for
    Randomly select a reaction based on rWeighting
    Add the selected reaction to pathway
    for each reactant metabolite, m, of the selected reaction
        if m is already in the host or in pathway
            continue
        else
            Call CONSTRUCT_PATH (m, selection scheme, pathway)
        end if
    end for
end
```

Figure 2.2: Pseudo-code for the probabilistic pathway construction algorithm. Pathways are constructed recursively starting from the target metabolite which is assigned as the root node of the tree. The tree is expanded at each recursion by adding an edge which represents a randomly selected reaction among all candidate reactions linked to the nodal metabolite. The constructed pathway is evaluated by calculating the maximum yield of the target metabolite using flux balance analysis.

order or otherwise evaluate pathway quality. In the present study, the length limit was set to 23 reactions, which reflects state-of-the-art with respect to the number of simultaneous gene insertions [7]. When the addition of a reaction to the tree violates this limit, the search algorithm backtracks and 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 required reactant metabolites (including cofactors) that are native to the host organism. Therefore, the search is iterated many times to explore a diverse number of possible pathways.

As previously observed [81], a small subset of the reactions in the KEGG database are annotated as 'unclear' and/or lack corresponding enzyme commission number entries. Such reactions were excluded from the search.

To evaluate the effectiveness of *ProbPath*, we compare the probabilistic searches against an exhaustive search, which constructs all possible pathways in the form of a single tree. Tree construction proceeds recursively, similar to the probabilistic search, except that the algorithm adds all of the possible pathways. The output of the search is thus a set of pathways, rather than a single pathway, that satisfies the length limit and terminates at a metabolite in the host. For the tree shown in Fig. 1a, the exhaustive search recursively explores all possible additions to the tree.

Due to the prohibitive computational cost associated with exhaustive search, the pathway length limit was set to 10 reactions (as opposed to 23, which is the limit set for the probabilistic search).

2.1.2 Probabilistic Reaction Selection

We explore three different selection schemes based on metabolite connectivity of candidate reactions: high-degree connectivity, low-degree connectivity and uniform. Here, degree connectivity refers to the number of reactions in which a metabolite participates. The results of the three different selection schemes are compared based on the likelihood of identifying the pathway with the highest predicted yield.

2.1.2.1 High-connectivity Scheme

In this scheme, we use weighted probabilities to bias the selection in favor of reactions involving high-degree metabolites. It has been observed that scale-free networks include hub nodes of high degree through which lower degree nodes connect [82]. For example, if A, B and C are three metabolites with A having a high-degree connectivity and B and C having low degrees of connectivity, a path from B to C is likely to proceed through A rather than directly. To verify that the metabolites in the KEGG multi-organism database constitute the nodes of a scale-free network, we characterized the degree distribution by counting the number of times each metabolite participates in a distinct reaction. This analysis did not consider the directionality of the reactions, as most of the reactions are reversible. A log-scale histogram (Figure 2.3) showed that the degree distribution indeed followed a power law similar to other evolved, scale-free networks [83] with a scaling exponent

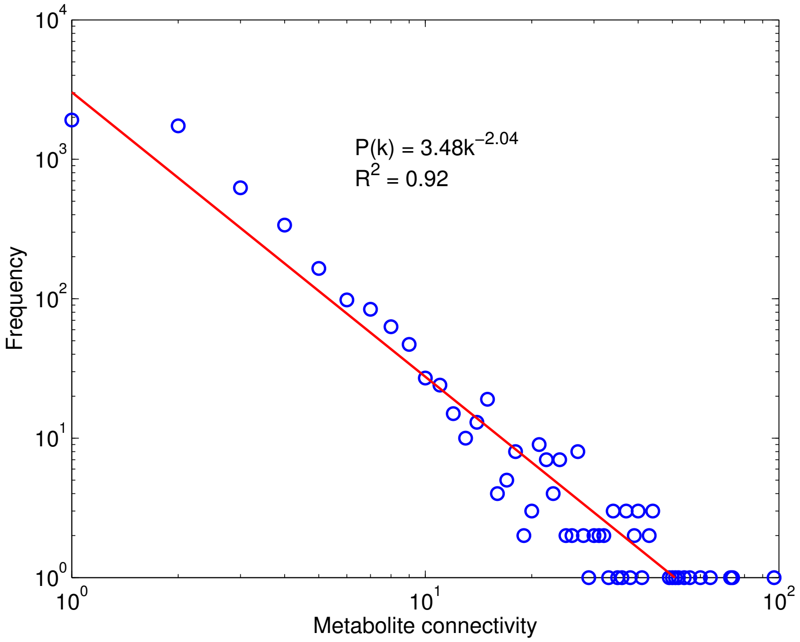


Figure 2.3: Degree connectivity distribution of metabolites in the KEGG database exhibiting a power-law distribution in which the probability of finding a metabolite with connectivity k is proportional to $k^{2.04}$. The blue circles depict the actual distribution and the red line represents the fitted power-law distribution.

value of 2.04. Motivated by this connectivity property of the KEGG database, we weighted the selection probabilities of candidate reactions to favor pathways whose intermediates are hub metabolites.

The probability of selecting a reaction is proportional to its relative weight normalized by the sum of the weights of all candidate reactions, R_i. Mathematically put, $Prob(selecting R_i) = W_{Ri}/\Sigma W_{Rj}$. The weight of a reaction R_i, W_{Ri}, is computed as one less than the minimum degree of all metabolites connecting to the candidate reaction. The degree of a node (representing a metabolite) is defined as the number of edges (representing reactions) associated with each node. As an example, consider the hypothetical reactions shown in Figure 2.4. Metabolites B through E have the following degree values: deg(B)=5, deg(C)=4, deg(D)=3 and deg(E)=2. The weights for the reactions are W_{R1}=min(5, 4)-1=3, W_{R2}=3-1=2 and

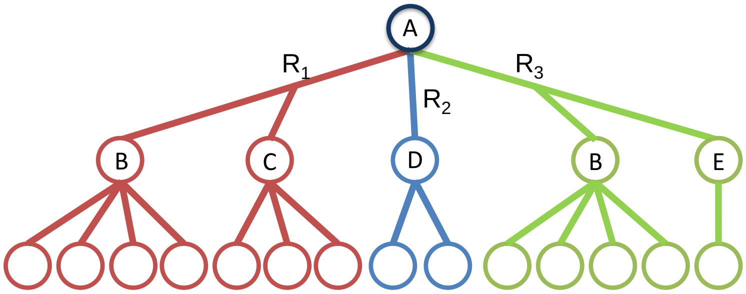


Figure 2.4: Metabolite connectivity-based weighting scheme. The nodes and edges represent metabolites and reactions, respectively. The three reactions R_1, R_2 and R_3 producing the metabolite A are defined as follows. $R_1 : A \Leftrightarrow B+C$, $R_2 : A \Leftrightarrow D$, and $R_3 : A \Leftrightarrow B + E$. Based on these definitions, the algorithm assigns weights of 3, 2 and 1 to reactions R_1, R_2 and R_3, in order.

W_{R3}=min(5, 2)-1=1. In this example, the probabilities of selecting reactions 1, 2 and 3 are 0.5, 0.33 and 0.17, respectively.

It should be noted that a clear distinction between cofactors and main reactants is not always possible without manually inspecting the reaction definition. One way to discriminate cofactors is based on the degree connectivity, which is generally higher than other metabolites. Therefore, we determine the weight of a reaction based on the metabolite with the lowest degree connectivity as calculated by the 'min' function in the formulas above. Also, all of the metabolites in a reaction are *and*-related, *i.e.* different pathway branches should be constructed for all of them.

2.1.2.2 Low-connectivity Scheme

In this scheme, we bias the selection in favor of reactions involving low-degree metabolites. The idea of identifying pathways using low-connectivity metabolites has been used previously to infer meaningful pathways in biochemical networks. One such pathway identification algorithm is Metabolic PathFinding, which first assigns metabolites a weight equal to their connectivity and then performs a search for the path with the smallest cumulative weight, thus reducing the likelihood of including currency metabolites such as ATP or H2O [84]. Another algorithm, MetaRoute, assigns a modified weight to each reaction based on metabolite degree and then performs a weighted path search to compute the first k-shortest paths between two given metabolites [81]. Metabolic PathFinding, MetaRoute, and path-pruning methods [85] are motivated by the idea that low-connectivity metabolites define the major pathways of carbon (or nitrogen) transfer in biochemical networks. In the present study, the low-connectivity selection probabilities for the reactions are calculated similar to the high-connectivity scheme, except that the inverse of one less than the smallest metabolite degree in a reaction is used as the weight for the reaction ($1/W_{Ri}$).

2.1.2.3 Uniform Scheme

Finally, we investigate a selection scheme where each reaction is assigned the same weight. This scheme does not favor either high- or low-degree metabolites as pathway intermediates, and thus should return the most diverse set of pathways.

2.1.3 Yield Calculation

We evaluate each pathway by calculating the maximum yield of the desired product, subject to constraints, using flux balance analysis (FBA) [86]. FBA is a computational approach to optimize the flux (flow) through the network subject to steady-state conditions and other thermodynamic and physiological constraints. FBA is commonly formulated as a linear optimization problem, where the objective function either maximizes or minimizes a desired reaction flux or a combination of desired fluxes. In the present study, the maximum yield of the synthesis pathway is used as an overall performance metric of the entire synthesis pathway. The yield also takes into account several (but not all) important biochemical and biophysical constraints of the host organism. The search space can be further pruned by evaluating thermodynamic favorability [87].

As the base model for FBA, we used a genome-scale model of *E. coli* metabolism (iAF1260) [19]. Reactions selected by the search algorithm to constitute a plausible pathway were then added to the base model to generate the modified strain model. Upper and lower flux bounds for the added reactions were set to 1000 and 1000 $mmol/gDW/h$, respectively. All other constraints were kept at the same default values of the base model as described in [19]. In brief, the glucose uptake upper and lower bounds were 1000 and 8 $mmol/gDW/h$, respectively, oxygen uptake bounds were 1000 and 18.5 $mmol/gDW/h$ and the ATP maintenance requirement was set to 8.39 $mmol/gDW/h$ [19]. The FBA objective was to maximize the flux forming the desired product subject to the constraint that the modified host strain produces at least 80% of the wild-type biomass yield. Pathways leading to zero product fluxes were considered non-viable. Viable pathways were rank ordered according to the maximum product yield. For comparisons with literature values, product yields were expressed as fluxes normalized by the corresponding glucose uptake flux or dry cell weight (DCW).

2.2 Results and Discussion

To analyze the effectiveness of *ProbPath*, we examined the synthesis of several native and non-native metabolites, which have previously been identified as commercially useful targets for over-production, using *E. coli* as the host organism. Moreover, experimentally determined yield or titer data (the concentration of a substance in solution) were available in the published literature, thus providing references for comparison. The test compounds belonged to four groups: precursors for natural products with therapeutic activity (isopentenyl diphosphate and taxa-4,11-diene); alcohols used as building blocks for polymer synthesis and other commercial applications (1,3-propanediol and 2,3-butanediol); a complex carbohydrate precursor for value added chemicals (*myo*-inositol); and lipid biofuels (fatty acid methyl and ethyl esters and triacylglycerol).

The analysis compared the performance of the probabilistic search algorithm for different weighting schemes (uniform, high connectivity, and low connectivity) based on the average and maximum yields of the synthesis pathways over all yields obtained from repeated runs. The probabilistic search with uniform weighting was also compared against an exhaustive search in terms of sampling efficiency as reflected in the yield diversity of the pathways. This comparison also involved an analysis of the computational cost, which showed that the run time of the exhaustive search increases exponentially with respect to the number of reactions in the *pathway*, whereas the runtime of the probabilistic search scales linearly with the number of reaction in the *database*. Finally, we compared the yield results calculated by the probabilistic and exhaustive searches against experimentally obtained values reported in the literature.

2.2.1 Yield Results for Different Weighting Schemes

Due to the probabilistic nature of *ProbPath*, meaningful interpretation of the search requires a large number of iterations. To estimate the number of iterations needed to identify viable, high-yield pathways, we executed the probabilistic search for varying numbers of iterations ranging from 100 to 1500 and recorded the maximum product yield obtained for each iteration number. To also determine an average yield, we repeated this process 500 times for each iteration number. The results of this analysis for fatty acid methyl esters are shown in Figure 2.5. The *overall* maximum product yield remained constant for all iteration numbers. The *average* maximum product yield increased steadily with the iteration number, and gradually leveled off, reaching a plateau around iteration number 1500. Similar trends were observed for all other test cases (data not shown). These trends suggest that the likelihood of finding a pathway with the overall maximum yield increases with the iteration number, but only up to a point. Once a sufficiently large iteration number has been reached, the likelihood of finding a pathway with the overall maximum yield remains essentially unchanged.

In addition to the overall maximum yield, we also recorded the standard deviation of the maximum yields for the 500 repeats at the various iteration numbers. To more clearly visualize the trend, we plotted the standard deviations for iteration numbers up to 4000 (Figure 2.6). The decreasing standard deviations suggest that increasing the iteration number improves the predictability of the search outcomes resulting from repeated runs. Given that we calculate and record the maximum yield for each run in a batch of repeats, the convergence in yield trends toward the overall maximum.

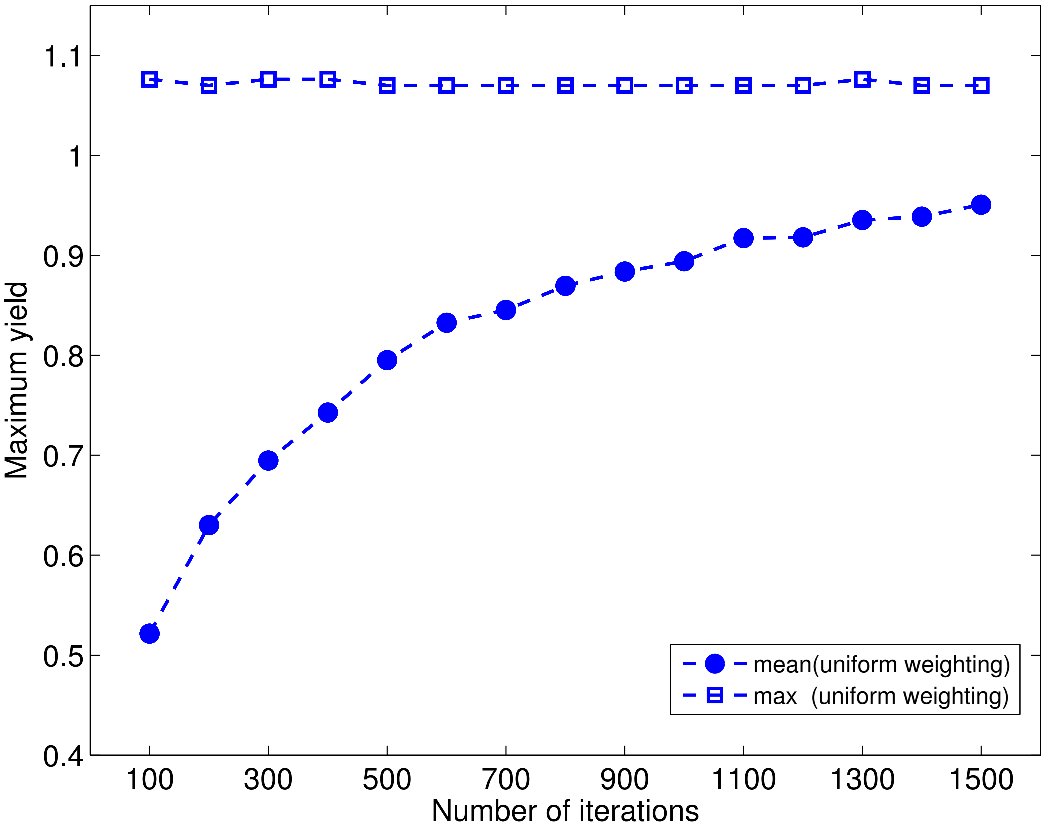


Figure 2.5: Dependence of the overall and average maximum yields on the number of iterations for the fatty acid methyl ester test case. At each iteration number, the probabilistic search was repeated 500 times. The overall maximum refers to the largest of the 500 maximum yields for the iteration number calculated using flux balance analysis (FBA). The average maximum yield refers to the arithmetic mean of the 500 FBA yields.

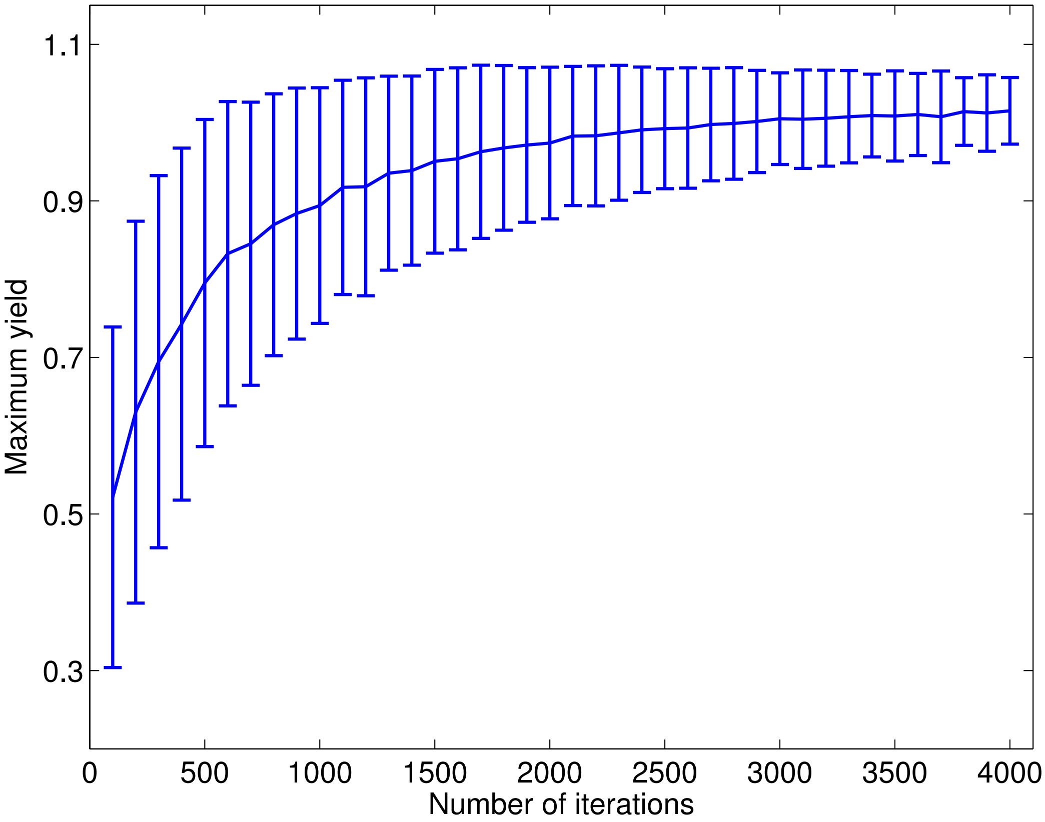


Figure 2.6: Average yield (solid line) and standard deviation (error bars) vs. number of iterations for the fatty acid methyl ester test case.

To examine the impact of the weighting scheme on search performance, we repeated the probabilistic search with uniform, low-connectivity and high-connectivity selection of reactions. Comparisons of overall and average maximum yields for fatty acid ethyl esters (FAEEs) are shown in Figure 2.7. The uniform weighting scheme consistently outperformed the connectivity-based weighting schemes, as it needed fewer iterations to identify viable, high-yield pathways. This improvement in the search performance suggested that the high-yield pathways involve intermediates with both high and low metabolite connectivity. This indeed was the case for fatty acid ethyl esters. The pathway with the highest maximum product yield (3.58 $mmol/gDW/h$) was identified only when the reactions were selected based on a uniformly random probability. The connectivity values of metabolites in the reaction sequence for this pathway were 3, 8, 6, 44, 7, 24, 11, 22, 7 and 8, in order. The maximum product yield calculated using the high-connectivity weighting scheme was 3.37 $mmol/gDW/h$. The metabolite connectivity values for this reaction sequence were 3, 8, 6, 44, 7, 24, 11, 22 and 12, in order, differing only slightly from the higher yield pathway at the end of sequence. As a consequence of its bias for reactions involving high-degree metabolites (in this case, bias for the reaction involving a product with degree 12 over degree 7) the high-connectivity weighting scheme favors the lower yield sequence, whereas the uniform weighting scheme is equally likely to select either sequence.

Similar results were obtained for triacylglycerol, with greater average maximum yields calculated by the uniform weighting scheme compared to the connectivity weighting schemes (Appendix A Figure 13). In the case of fatty acid methyl esters, the results of the low-connectivity weighting method were similar to those

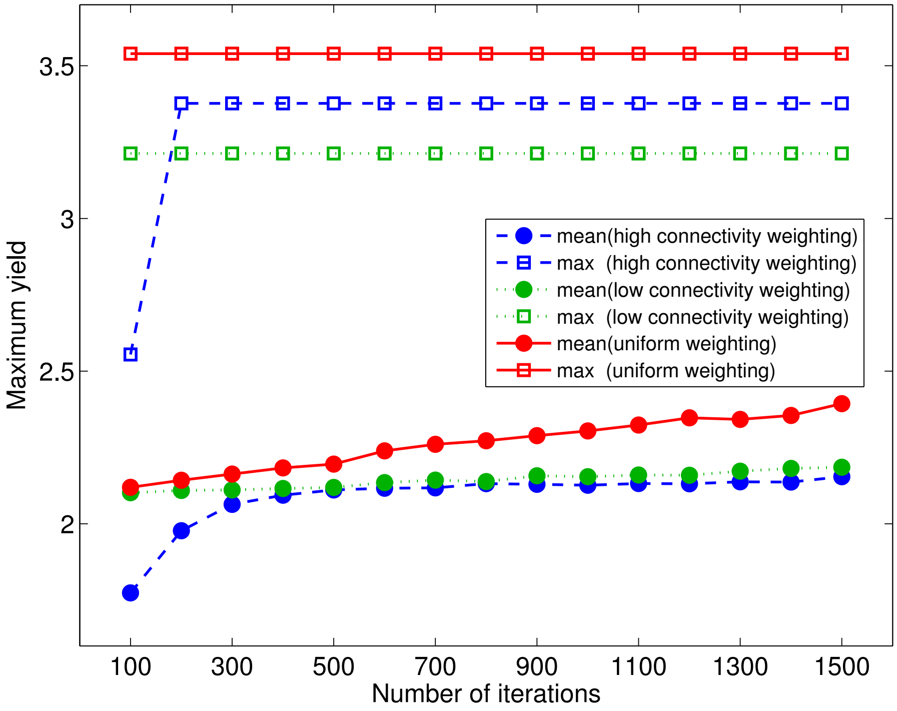


Figure 2.7: Yield comparisons for fatty acid ethyl esters obtained using connectivity-based and uniform weighting schemes.

of the uniform weighting method (Appendix A Figure 12), suggesting that there are some highest-yielding pathways (1.065 $mmol/gDW/h$) that proceed through metabolites with low-connectivity values. Since the uniform probabilistic method is able to find not only the pathways found by the low-connectivity method but also other pathways with the same yield, which otherwise cannot be identified using low-connectivity weighting due to some high-connectivity metabolites in them, uniform probabilistic method outperforms the low-connectivity probabilistic search in this case. One possible reason the uniform weighting scheme outperforms the connectivity-based schemes could be due to interactions with the host metabolic network. Integrating a high-connectivity pathway with the host could subject many native pathways to competition with the non-native pathway. Likewise, a pathway

Metabolite name	Weighting scheme	Normalized average maximum yield
Isopentenyl diphosphate	Uniform	1
	High-connectivity	1
	Low-connectivity	1
Myo-Inositol	Uniform	1
	High-connectivity	1
	Low-connectivity	1
Taxadiene	Uniform	1
	High-connectivity	1
	Low-connectivity	1
1,3-Propanediol	Uniform	1
	High-connectivity	1
	Low-connectivity	1
2,3-Butanediol	Uniform	1
	High-connectivity	1
	Low-connectivity	1
Fatty acid ethyl esters	Uniform	0.64
	High-connectivity	0.60
	Low-connectivity	0.58
Fatty acid methyl esters	Uniform	0.77
	High-connectivity	0.69
	Low-connectivity	0.76
Triacylglycerol	Uniform	0.40
	High-connectivity	0.38
	Low-connectivity	0.38

Table 2.1: Effect of the weighting scheme on search performance. The normalized average maximum yield was calculated by dividing the average maximum yield with the overall maximum yield. Average and overall maximum yields were determined for runs of 1,000 iterations repeated 50 times as described in the text.

with low-connectivity metabolites could add to the scarcity of metabolites with one or few native routes of production.

The three weighting schemes returned similar performances for all other test cases (Table 2.1 and Appendix A Figures 7-11), where the synthesis pathways did not exhibit significant variations in the obtained yield.

2.2.2 Sampling Efficiency

For every test case, the probabilistic search with the uniform weighting scheme identified viable synthesis pathways supporting a non-zero yield of the target product and at least 80% of the maximum wild-type biomass flux. A summary of the search results is shown in Table 2.2. In general, the exhaustive search returned a greater number of pathways than the probabilistic search, despite the lower length limit (10 vs. 23 reactions). Equally small numbers of pathways were identified for taxadiene and 1,3-propanediol (2 and 1, respectively), presumably reflecting the involvement of singular reactions. In the cases of isopentenyl diphosphate (IPP) and *myo*-inositol, the number of pathways identified by the probabilistic search was greater than the exhaustive search. In the cases of the lipids, the exhaustive search generated a larger number of pathways than the probabilistic search, with the fold differences in the number of pathways ranging from 30 to 58.

Despite the differences in the total number of pathways, the maximal yields calculated by the probabilistic (uniform weighting) and exhaustive searches were identical in all test cases, except for fatty acid methyl esters, where the difference was less than 1%.

For the sake of completeness, we also compared the results obtained from the uniform probabilistic and exhaustive search with length limit of 10 reactions (Appendix A Table 1). For some of the test cases, the number of identified pathways using the probabilistic method decreased. However, the maximum calculated fluxes were the same, because the search found pathways with lengths less than 10 reactions and the same highest yield (Table 2.2).

The similarity of the maximal yields suggested that the two search methods

Metabolite name	Native yield	Method	# pathways	Max. yield/rate	Pathway lengths (pathways with yields larger than 95% of the max. yield)
Isopentenyl diphodphate[a]	301.13 $mg/gDW/h$	Uniform *ProbPath*	11	314.24 $mg/gDW/h$	4, 5, 8, 10, 19
		Exhaustive search	9	314.24 $mg/gDW/h$	4, 5, 8, 10
		Literature	NA	27.4 g/L amorphadiene [88] (1.95 $mg/gDW/h$)	NA
Myo-inositol[b]	0 mol/mol glucose	Uniform *ProbPath*	71	0.2 g/g glucose	2, 7, 8, 9, 10, 11, 13
		Exhaustive search	42	0.2 g/g glucose	2, 7, 8, 9, 10
		Literature	NA	0.23 g/L [89] (0.08 g/g glucose)	NA
Taxadiene[c]	Non-native	Uniform *ProbPath*	2	0.06 g/g glucose	2, 5
		Exhaustive search	2	0.06 g/g glucose	2, 5
		Literature	NA	1.3 mg/L [90] (0.06 mg/g glucose)	NA
1,3-Propanediol	Non-native	Uniform *ProbPath*	1	2.19 $mmol/gDW/h$	2
		Exhaustive search	1	2.19 $mmol/gDW/h$	2
		Literature	NA	2.3 $mmol/gDW/h$ [47]	NA
2,3-Butanediol	Non-native	Uniform *ProbPath*	9	0.11 g/g glucose	2, 3, 4
		Exhaustive search	9	0.11 g/g glucose	2, 3, 4
		Literature	NA	0.31 g/g glucose [91]	NA
Fatty acid ethyl esters[d]	Non-native	Uniform *ProbPath*	19	3.58 $mmol/gDW/h$	8, 9, 10, 11
		Exhaustive search	1092	3.58 $mmol/gDW/h$	8, 9, 10
		Literature	NA	647 mg/L [3] (0.34 g/g glucose)	NA
Fatty acid methyl esters[e]	Non-native	Uniform *ProbPath*	45	1.07 $mmol/gDW/h$	7, 8, 9
		Exhaustive search	1353	1.08 $mmol/gDW/h$	7, 8, 9
		Literature	NA	0.3 g/gDW [92]	NA
Triacylglycerol	Non-native	Uniform *ProbPath*	51	1.65 $mmol/gDW/h$	8, 9
		Exhaustive search	2900	1.65 $mmol/gDW/h$	8, 9
		Literature	NA	0.06 mol/mol glucose [93]	NA

For comparisons with FBA calculations, the reported titer values were converted as follows:

[a] IPP ($27.4\ g/L\ /\ 88\ gDW/L$) / $160hr = 1.95\ mg/gDW/hr$

[b] *Myo*-inositol $10\ g/L - 7\ g/L = 3\ g/L$ (glucose consumed)
$0.23\ g/L\ /\ 3\ g/L = 0.08\ g/g$ glucose

[c] Taxadiene 1.3 mg/L / 20 g/L = 0.06 mg/g glucose (We assumed all glucose in LB medium is consumed.)

[d] FAEE 647 mg/L / 2 g/L = 0.34 g/g glucose

[e] FAME The yield was reported for thraustochytrid Aurantiochytrium sp. strain T66 (as opposed to *E. coli*).

Table 2.2: Summary of search results obtained using uniform probabilistic and exhaustive methods. Results are shown for 1,000 iterations of the probabilistic search and one single run of the exhaustive search for each test case. The total number of pathways includes only those with specific productivities greater than the wild-type organism represented by the base model. The maximal yield refers to the pathway with the highest ratio of product flux to biomass flux. Product fluxes were calculated using FBA with the constraint that the biomass flux exceeds 80% of the wild-type.

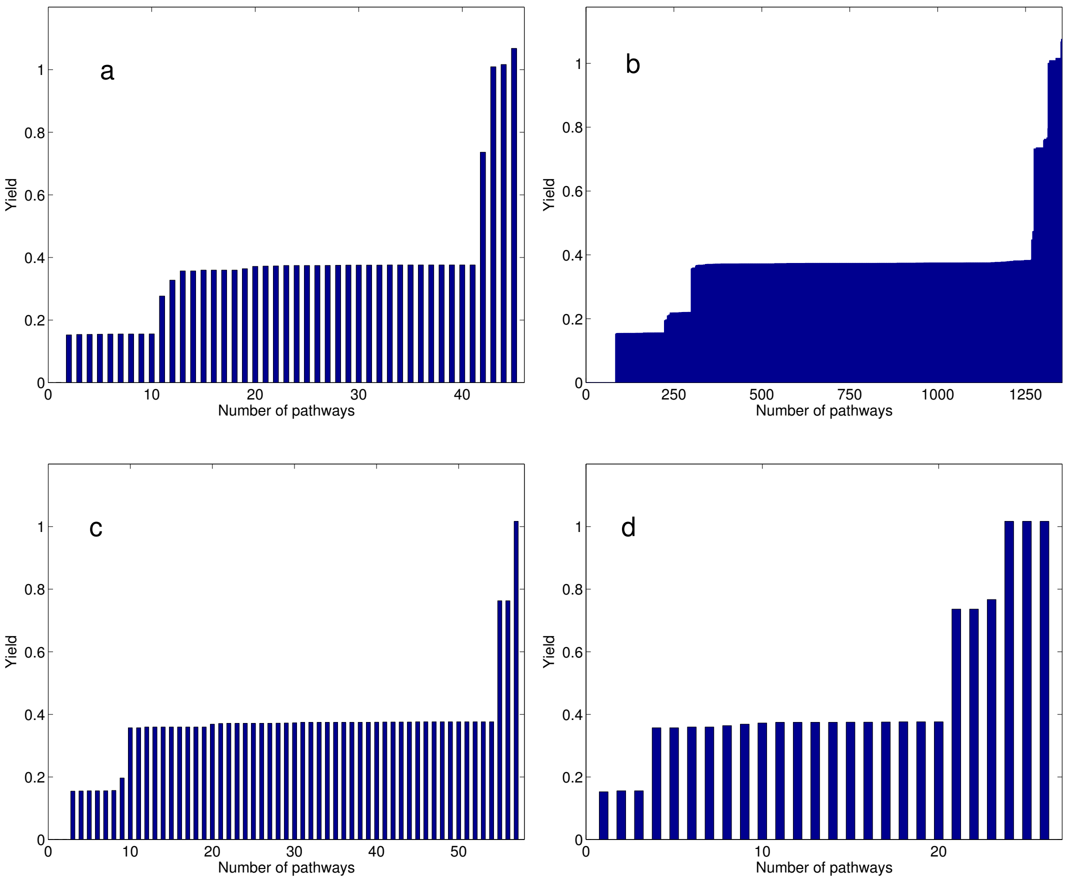


Figure 2.8: Yield distributions for fatty acid methyl esters obtained using uniform probabilistic search (a), exhaustive search (b), high-connectivity probabilistic search (c) and low-connectivity probabilistic search (d). The histograms for the probabilistic searches represent the cumulative results of 1000 iterations. The histogram for the exhaustive search reflects a single run.

identified at least partially overlapping sets of pathways. To evaluate the extent of the overlap, we compared the yield distributions resulting from the two methods for each test case (Appendix A Figures 1-6). In the case of fatty acid methyl esters (Figures 2.8 a, b), the yield distributions were essentially identical, even though the total number of pathways returned by the probabilistic search was less than 4% of the total returned by the exhaustive search. Similar trends were observed for all other test cases, suggesting that the probabilistic search representatively sampled the space of possible pathways in the KEGG database.

We also compared the yield distributions for fatty acid methyl esters obtained from different weighting schemes (Figures 2.8 a, c, d). In all three cases, similar patterns were generated, suggesting that there is no correlation between connectivity and yield distribution. In conclusion, using uniform weighting proves superior in all our test cases in finding highest-yielding pathways compared to other weighting schemes.

2.2.3 Runtime Comparisons

To examine the computational efficiency of the probabilistic search, we compared its runtime against the exhaustive search. The runtime for the exhaustive search grew exponentially with the number of reactions used in the pathway (Figure 2.9), rendering the algorithm intractable for longer pathways. For example, a single run of the exhaustive search with a pathway length limit of 23 was projected to require a runtime exceeding 400 years on a workstation with four Quad-Core 2.3 GHz processors (AMD Optron 8356) and 64 GB of physical memory. The runtime for the probabilistic search shows a linear dependence on the number of reactions in the KEGG database. With the identical pathway length limit of 23, 1000 iterations of the probabilistic search required a runtime of 6 minutes on the same workstation.

2.2.4 Experimental Support

We next examined the quality of the results from the probabilistic search (uniform weighting) by comparing the FBA calculations against published data on experimentally obtained yields. In general, direct comparisons of the predicted yields against published values were not possible. The immediate output of FBA is spe-

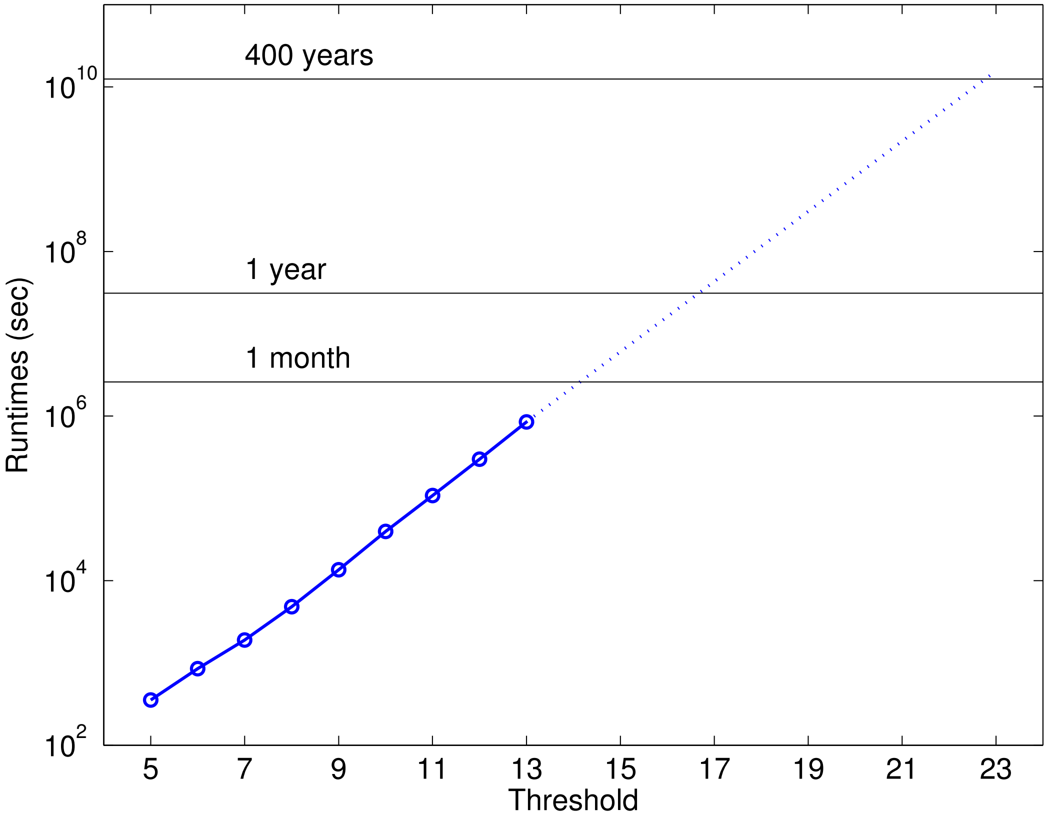


Figure 2.9: Runtimes of the exhaustive search as a function of pathway length limit. Data shown are for the fatty acid methyl esters test case. Symbols indicate recorded values from simulations performed on a four Quad-Core 2.3 GHz AMD Optron 8356 with 64 GB of physical memory. The dashed line extrapolates the runtimes for length limits up to 23 reactions, which is the maximum number of steps allowed for the probabilistic search. Extrapolation is performed based on linear regression of log-transformed runtime data against length limit.

cific productivity (rate of target production normalized to dry cell weight) during balanced growth, which can then be normalized to glucose uptake or another flux. Typically reported values are volumetric productivity (product titer and cell density) measurements obtained from shake flask or fed-batch experiments [88] [3]. In some cases, quantitative details regarding the carbon source and other culture parameters needed to derive the specific productivity or yield were not reported. In such cases, representative values were used based on a survey of the literature. For example, we used the default value of 0.4 absorbance unit per $gDCW/L$ to convert the reported OD 600 readings into cell concentrations. The details of converting the reported titer values varied from case to case, and are therefore described separately for each case in the caption of Table 2.2.

In the case of IPP, the search algorithm identified 11 distinct pathways, all of which led to higher maximal yields compared to the native pathway. Comparisons with published data suggested that the predicted maximal yield is two orders of magnitude greater than previously observed yields. Similarly large differences were found between the FBA calculation and reported values for the other natural product, taxadiene. In the case of the alcohols and *myo*-inositol, the FBA results were of comparable magnitude as previously achieved production rates [47] [89] [91]. In the case of the lipids, comparisons with published values were further confounded by the generic representation of hydrocarbon side chains and unspecified stoichiometries of the relevant reactions in the KEGG database. Computing mass yields was especially problematic for synthesis pathways involving chain elongation reactions catalyzed by multi-functional enzyme complexes such as fatty acid synthase.

In addition to yield, we also examined the compositions of the pathways resulting from the probabilistic search. For every test case, the computational search is able to reconstruct viable pathways involving reactions whose insertion or overexpression has been shown to improve the yield. In the case of IPP, the search results included the mevalonate pathway, which has been shown to be a preferred synthesis route in *E. coli* [5]. Two pathways were identified for taxadiene, one of which started with IPP, as previously shown in [90]. The other started with farnesylfarnesylgeraniol and had the same yield. In the case of *myo*-inositol, several pathways started from acetyl-CoA as reported in [94]. The search also identified several other pathways that started from d-glucose-6-phosphate, acetaldehyde, pyruvate, propanoyl-CoA or malonyl-CoA. The highest yield belonged to the pathway which started with acetyl-CoA. Other identified pathways had slightly less yields. Only one pathway was identified for 1,3-propanediol, which started from glycerol, consistent with the analysis in a review [2]. The synthesis pathways for (R,R)-2,3-butanediol could begin with a variety of metabolites native to *E. coli*, including (S)-2-acetolactate, 3-methyl-2-oxobutanoic acid, (R)-2,3-dihydroxy-3-methylbutanoate (all three had the highest yield), thiamin diphosphate and pyruvate. Of these, the pathway starting from pyruvate has already been demonstrated experimentally in *E. coli* [91]. Pathways producing fatty acid ethyl esters (FAEEs) started with different metabolites, including 1,2-diacyl-sn-glycerol with the highest yield of 3.58 $mmol/gDW/h$, (3R)-3-hydroxyacyl-[acyl-carrier protein], acetyl-CoA, phosphatidylethanolamine, 1-acyl-sn-glycerol 3-phosphate, 2-acyl-sn-glycero-3-phosphoethanolamine and phosphatidate, all with lesser yields. Depending on the length of the desired hydrocarbon side chain, fatty acid methyl esters (FAMEs) could be produced from various glycerolipids and phospholipids such as (3R)-3-

hydroxyacyl-[acyl-carrier protein] with the highest yield of 1.07 $mmol/gDW/h$, CDP-diacylglycerol, phosphatidylethanolamine, choline, 1,2-diacyl-sn-glycerol, phosphatidate and 1-acyl-sn-glycerol 3-phosphate. In the case of triacylglycerol, most pathways involved acyl-CoA [95] starting with (3R)-3-hydroxyacyl-[acyl-carrier protein] which generated the highest yield of 1.65 $mmol/gDW/h$, phosphatidate, phosphatidylethanolamine and 1-acyl-sn-glycerol 3-phosphate. Other pathways, which did not involve acyl-CoA, began with CDP-diacylglycerol, choline and 1,2-diacyl-sn-glycerol.

2.3 Conclusion

We designed a probabilistic graph-based search algorithm to identify novel, non-native synthesis pathways for metabolite over-production using heterologous hosts. Importantly, the algorithm considers not only the main reactants, but also the cofactors needed for the biosynthesis. The probabilistic search for pathways is based on uniform weighting and the degree of metabolite connectivity determined from the KEGG database. Results demonstrate that uniform weighting outperforms the connectivity weighting in terms of average maximum yield with the same number of iterations. *ProbPath* is much faster (minutes, independent of the pathway length) than the exhaustive search algorithm ($\sim$ years for the longest pathways) and takes into account all possible pathways in a probabilistic way.

Using this method, we were able to reproduce experimentally obtained pathways reported in the literature as in the cases of IPP, *myo*-inositol, taxadiene, 1,3-propanediol, (R,R)-2,3-butanediol, fatty acid ethyl and methyl esters and triacylglycerol. The corresponding maximum yields are also comparable with those reported in the literature. We also compared the yield results of *ProbPath* with those of an exhaustive search method which looks for all possible pathways leading to the target metabolite production in a higher rate. Our calculations show that for reasonable number of iterations ($\sim$ 1000 with a runtime on the order of minutes) the results of both methods are comparable.

Chapter 3

PROXIMAL: A Method for Prediction of Xenobiotic Metabolism

One limitation of *ProbPath*, our probabilistic construction pathway algorithm, is that it cannot identify synthesis pathways for a target metabolite not catalogued within a database. A more general approach would accommodate any target metabolite, and allow for exploring reaction steps that are not necessarily catalogued in the databases. To investigate this approach, we focused on the specific problem of predicting the biotransformation of xenobiotics by human enzymes. We develop a new method, termed PROXIMAL, that analyzes Phase I and Phase II xenobiotic transformations that are cataloged in public databases (*i.e.* DrugBank [96, 97, 98] and KEGG [15, 99]), and builds look-up tables linking specific molecular substructures with matching biotransformation operations that modify these substructures. To achieve specificity, the look-up tables take into account molecular substructures that consist of a reaction center and its two-level nearest neighbors. Given a query compound, PROXIMAL applies a select set of transformations from the look-up tables at one or more matching sites, or reaction centers, of the query compound. PROXIMAL then ranks the transformation results based on the activity and abundance of the enzymes involved in the transformations. To evaluate the predictive power of PROXIMAL, we investigate two case studies involving bisphenol A (BPA) and 4-chlorobiphenyl (PCB3), two environmental chemicals with suspected endocrine disrupting activity. While the presentation in this chapter is specific to predicting xenobiotic biotransformation, the approach can be generalized to other classes of enzymes.

3.1 Background: Metabolism of Ingested Drugs and Foreign Chemicals in Human Livers

When ingested, drugs and other foreign chemicals can be metabolized by xenobiotic transformation enzymes, which are expressed throughout the body, in particular the liver and intestine. In mammals, including humans, clearance of xenobiotic chemicals from the body involves two to three phases, with the first two phases carrying out key structural modifications. Typically, Phase I activates the chemical by introducing a reactive and polar functional group, whereas Phase II conjugates the activated chemical with a charged compound, increasing the molecular weight, reducing reactivity, and improving transport properties. An additional Phase III step can follow the conjugation step to eliminate the conjugated chemical from the cell into the extracellular medium. The enzymes mediating these reactions have broad specificity, and thus are capable of generating a variety of metabolic products. Cytochrome P450 (CYP) enzymes play an especially important role in Phase I modification, which often involves oxidizing the substrate by introducing a hydroxyl group or oxygen atom. Depending on the substrate, a CYP reaction can produce a highly reactive derivative that can bind and modify other molecules in the cell, including macromolecules, and thus pose a cytotoxicity risk [100].

In some cases, the products of xenobiotic transformation can be biologically active, and interact with endogenous enzymes or regulator molecules to interfere with critical physiological processes. An example where this might be a concern is in the case of emerging contaminants. For example, diethylhexyl phthalate (DEHP) is a widely used plasticizer that can be hydrolyzed in the body to form monoethylhexyl phthalate (MEHP). *In vitro* experiments using molecular and cellular assays have shown that MEHP can selectively activate the nuclear receptor peroxisome proliferator-activated receptor-γ (PPAR-γ) to promote adipogenesis [101], and thus could contribute to the development of obesity. Hydroxylated derivatives of polychlorinated biphenyls (PCBs) can inhibit the sulfation of thyroid [102] and steroid hormones [103], and thereby disrupt an important mechanism for regulating the levels of these hormones. In the case of PCBs, hydroxylation can also enhance the toxicity of these chemicals [104], possibly through a mechanism involving oxidative DNA damage [105]. A similar increase in toxicity has also been reported to result from hydroxylation of polybrominated diphenyl ethers (PBDEs) [106].

One approach for identifying endogenously formed xenobiotic transformation products is to experimentally profile bodily fluids such as blood or urine for compounds that are structurally related to the xenobiotic chemical. For example, Dhakal et al. utilized mass spectrometry (MS) to identify several Phase I and Phase II derivatives of a PCB in urine and fecal samples from mice [107]. This metabolite profiling study utilized selective ion monitoring, a MS method for targeted analysis. While this targeted approach affords quantitative analysis, it may not be comprehensive. By definition, targeted analysis requires a priori knowledge of the chemicals of interest, and consequently limited in its potential for discovery. Ideally, investigation of xenobiotic transformation is sufficiently comprehensive to characterize the breadth of metabolic products that can be derived from a chemical of interest, while also focused enough to robustly identify and quantitate the products. To this end, complementing experimental analysis with computational prediction of transformation products could be a powerful strategy to enhance the discovery potential of quantitative analytical experiments.

Several computational approaches have been developed to predict xenobiotic transformations that result in structural modifications to the chemical. Examples of well-known approaches are UM-PPS [108, 109, 110], Meta [111, 112, 113] and Meteor [114, 115, 116]. The University of Minnesota Pathway Prediction System (UM-PPS) [108, 109, 110] is a rule-based method specifically developed to predict microbial catabolism of organic compounds. Based on curated information on microbial reactions cataloged in the University of Minnesota Biocatalysis/Biodegredation Database (UM-BBD) [117] and documented in the published literature, UM-PPS generates a set of rules that specify how predefined functional groups may be modified through a metabolic reaction. These rules are ranked based on their thermodynamic feasibility and applied to the matching functional groups in the query compound to predict the compound's metabolic products. Meta, another rule-based method, predicts xenobiotic transformations in mammals by generating rules compiled from reviews and textbooks [111, 113]. Given a query compound, Meta searches for a molecular fragment in the compound that is recognized by a specific enzyme and transforms it into a product fragment. Meta uses a genetic algorithm to optimize the rules based on experimental observations to improve the predictions [112]. For a given metabolite, Meteor predicts the possible transformation steps using a reasoning engine that has a knowledge base composed of generic rules [115, 116]. Several hundred (841) rules are derived from 217 known biotransformation reactions. The rules are assigned a rank according to the lipophilicity and molecular weight of the query metabolite [114]. Kirchmair et al. provide a comprehensive review of computational approaches for predicting outcomes of xenobiotic transformations [118]. A drawback of rule-based approaches is that they rely heavily on generic transformation rules, leading to a large number of predictions that may be difficult to evaluate and interpret.

3.2 Methods

The PROXIMAL method has three steps. The first step catalogs known xenobiotic transformation reactions (enzymes, substrates and products) recorded in databases. In this study, we focused on CYP enzymes (Phase I enzymes) and transferases (Phase II enzymes), which account for the bulk of xenobiotic chemical modifications and conjugations in mammals. The cataloged information is used to build look-up tables that associate a particular molecular substructure with a specific pattern of modification or conjugation. The second step uses the look-up tables to apply a select set of transformations to a matching substructure within the chemical

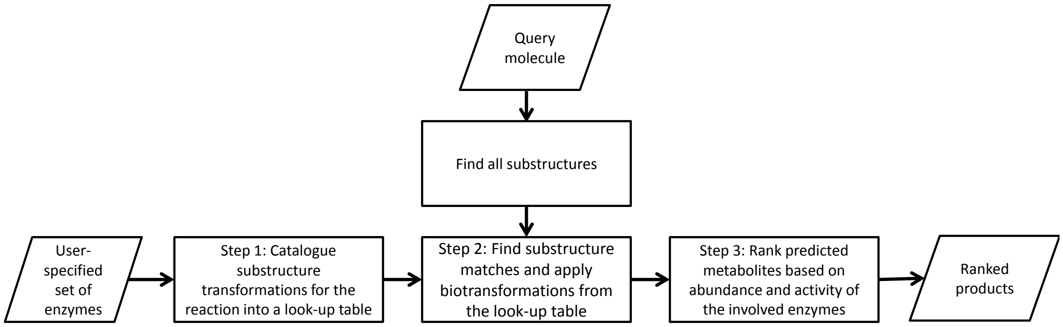
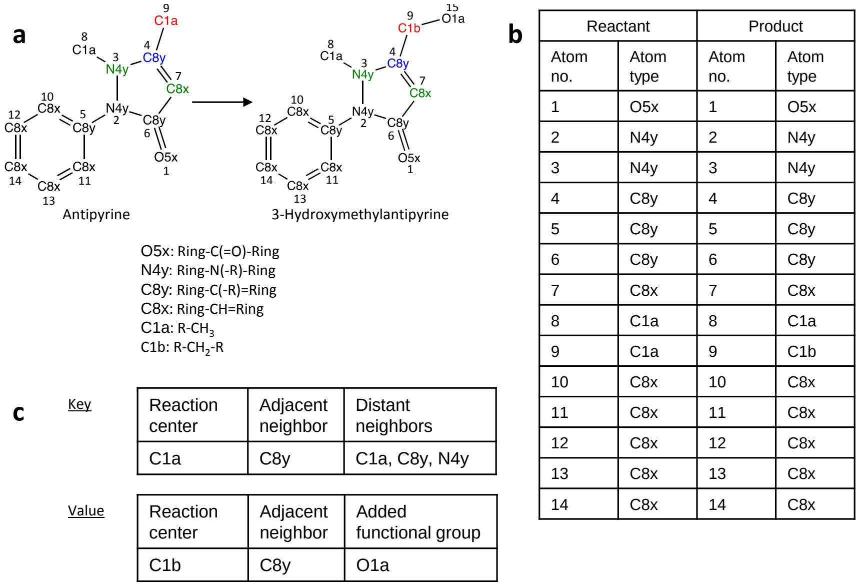

Query molecule
Find all substructures
User-specified set of enzymes
Step 1: Catalogue substructure transformations for the reaction into a look-up table
Step 2: Find substructure matches and apply biotransformations from the look-up table
Step 3: Rank predicted metabolites based on abundance and activity of the involved enzymes
Ranked products

Figure 3.1: A flowchart illustrating PROXIMAL's inputs, outputs and steps.

of interest. Depending on the chemical, this step may generate a large number of possible transformation products due to the number of sites available for modification. The third step ranks the predicted transformation products using available data on the activity and abundance of the enzymes associated with the transformations. Figure 3.1 shows PROXIMAL's steps, inputs and outputs. The inputs to PROXIMAL are a user-specified set of enzymes and a query molecule. The outputs are predicted products for the query molecule and their relative ranks.

3.2.1 Step 1: Generating Look-up Tables

To generate the look-up tables, we cataloged CYP oxidoreductases as representative enzymes responsible for Phase I modifications. We also cataloged several transferases that play a major role in Phase II conjugations of drug chemicals. These Phase II enzymes are: UDP-glucuronosyltransferase (UGT; EC 2.4.1.17), sulfotransferase (SULT; EC 2.8.2.1), N-acetyltransferase (NAT; EC 2.3.1.5), glutathione S-transferase (GST; EC 2.5.1.18), thiopurine S-methyl transferase (TPMT; EC 2.1.1.67) and catechol O-methyl transferase (COMT; EC 2.1.1.6) [119]. We used two publicly accessible databases, DrugBank [96, 97, 98] and KEGG [120, 99], to obtain the reaction data for these enzymes. At the time of completion of this work, DrugBank reported reaction data on 409 CYP enzymes and 70 transferases associated with Phase II drug metabolism, whereas KEGG listed reaction data on 154 and 61 Phase I and II enzymes, respectively. These two databases were mined to identify a master list of reactant-product pairs for reactions catalyzed by Phase I oxidoreductases or Phase II transferases. PROXIMAL accepts the list as an input and downloads the .mol files from the databases. Next, PROXIMAL aligns the structures of the reactant and product in each reaction using SIMCOMP [121], finds structural similarities and differences. SIMCOMP is based on a graph comparison algorithm to search for the maximal common subgraph from two chemical graphs. This problem can be reduced to finding maximal clique in the corresponding association graph [121]. The key ideas in SIMCOMP algorithm are labeling nodes by their chemical environments and weighing the edges in the association graph to allow for partial matching. SIMCOMP improves the clique finding algorithm by stopping the calculation of clique finding after a reasonable number of recursion steps and extending only large simply connected common subgraphs. SIMCOMP also generates a list of matched atoms for the reactant-product pair. Each atom in the list is represented by an atom number and a KEGG atom type [122], where the atom type describes the chemical element of the atom and its adjacent atoms (*i.e.* neighbors) in the molecule. Rows that contain different KEGG atom types for the reactant and product represent potential reaction centers. From the list of potential reaction centers, PROXIMAL removes those that are not directly connected to a functional group that is absent or new in the product. PROXIMAL then generates a list of reaction centers and their adjacent and distant neighbors to construct the look-up tables. An adjacent neighbor of an atom x is directly connected to atom x by a covalent bond. A distant neighbor of atom x is an adjacent neighbor of any one of atom x's adjacent neighbors. The look-up table stores information as a 'key' and 'value' pair. The 'key' consists of the information for the reaction center, its adjacent neighbors, and distant neighbors. The 'value' part of the look-up table holds the information for the changes to the atoms comprising the keys, including any added functional groups. We illustrate the construction of a look-up table using a Phase I modification of the drug antipyrine as an example (Figure 3.2). In this example, the cataloged CYP reaction hydroxylates the drug to form 3-hydroxymethylantipyrine (Figure 3.2a). Based on the .mol files for the drug and its transformation product, PROXIMAL represents the atoms in KEGG atom types, compares the reactant and product structures using SIMCOMP, and finds the reaction center, its adjacent neighbors, and distant neighbors. Reactant atoms that correspond to matching product atoms are arranged into a table such that each row contains a matching pair of reactant and product atom types (Figure 3.2b). The ordering of the rows is determined by the atom order (numbering) in the reactant .mol file. Reactant atoms corresponding to rows that contain different reactant and product atom types are marked as reaction centers. In the present example, the 9th row has different atom types (C1a changes to C1b). Therefore, C1a is the atom type of the reaction center. This reaction center has only one adjacent neighbor, atom number 4, which has atom type C8y. The distant neighbors are atom numbers 3, 7, and 9. The set of distant neighbors always includes the reaction center. Atoms 3 and 7 are type N4y and C8x, respectively. The reaction center (shown in red in Figure 3.2a), its adjacent atom (shown in blue), and distant neighbors (shown in



Reactant		Product	
Atom no.	Atom type	Atom no.	Atom type
1	O5x	1	O5x
2	N4y	2	N4y
3	N4y	3	N4y
4	C8y	4	C8y
5	C8y	5	C8y
6	C8y	6	C8y
7	C8x	7	C8x
8	C1a	8	C1a
9	C1a	9	C1b
10	C8x	10	C8x
11	C8x	11	C8x
12	C8x	12	C8x
13	C8x	13	C8x
14	C8x	14	C8x

Key

Reaction center	Adjacent neighbor	Distant neighbors
C1a	C8y	C1a, C8y, N4y

Value

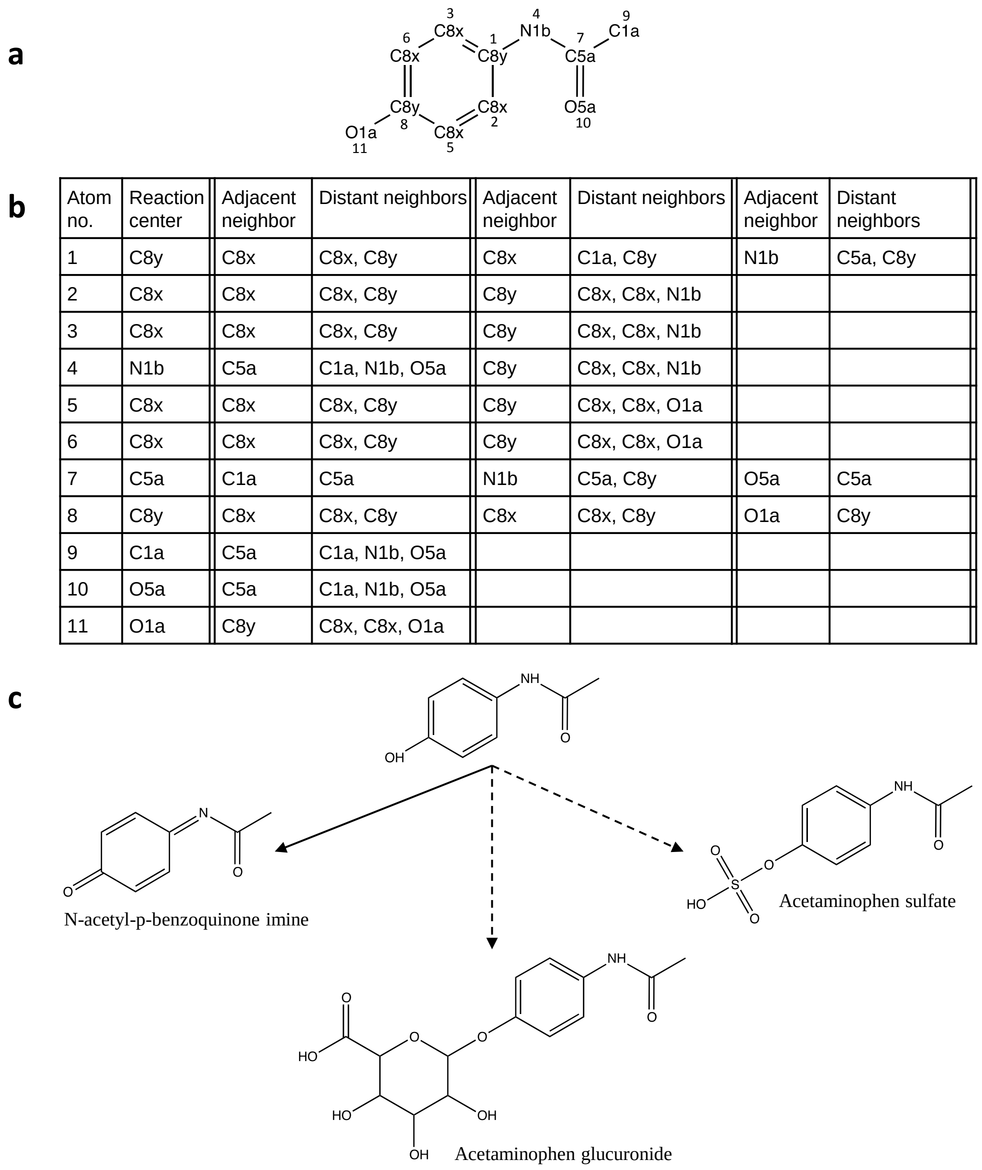
Reaction center	Adjacent neighbor	Added functional group
C1b	C8y	O1a

Figure 3.2: Schematic illustration of look-up table construction (PROXIMAL step 1). (a) A CYP reaction transforms the drug antipyrine to 3-hydroxymethylantipyrine, both represented in KEGG atom type format. Red, blue and green atoms are the reaction centers, adjacent neighbors and distant neighbors, respectively. The .mol files were downloaded from DrugBank. (b) List of matched atoms for the reactant and product. (c) The transformation look-up table has two parts, consisting of a key that specifies the modified reactant substructure and a corresponding value that describes the modifications resulting in the product.

green) together comprise the substructure where the CYP-mediated chemical modification occurs. This substructure forms the ‘key’ of a look-up table (Figure 3.2c). The corresponding ‘value’ is O1a, which is the atom type added to the reactant through the CYP reaction. In the case where there are additional reaction centers, the corresponding rows in the alignment table (Figure 3.2b) are processed similarly to generate additional keys and values. The same process is also used to construct look-up tables for substructures around reaction centers identified in the substrates of the Phase II enzymes.

3.2.2 Step 2: Generating Potential Products

Given a chemical of interest, PROXIMAL applies the transformation patterns represented in the look-up tables to generate possible products of Phase I and/or Phase II reactions. The chemical of interest is specified using an input .mol file. PROXIMAL processes this input file to represent the chemical in KEGG atom type format (Kanehisa et al., 2004) using the KEGG API (Application Programming Interface). PROXIMAL treats all atoms in the chemical as possible reaction centers, and builds corresponding lists of adjacent and distant neighbors. A query is performed to identify any substructures that match to a key in the look-up tables. If there is a match, then the key's value is applied to generate a biotransformation product. Depending on the number of matches, PROXIMAL may generate multiple products for a given chemical. The use of the look-up table in predicting xenobiotic transformation products is illustrated through an example involving acetaminophen (Figure 3.3). After converting the drug's .mol file to the KEGG atom type format (Figure 3.3a), PROXIMAL generates a list of potential reaction center atoms and their neighbors (Figure 3.3b). Some reaction centers (9-11) have one adjacent neighbor, others have two adjacent neighbors (2-6), and some have three adjacent neighbors (1, 7 and 8). Each reaction center in the list is compared with keys in the Phase I and II look-up tables. In this example, we find matching keys for the 11th row (reaction center: O1a; adjacent neighbor: C8y; distant neighbors: C8x, C8x, O1a). Applying the corresponding value from the Phase I look-up table generates N-acetyl-p-benzoquinone imine. Applying the values from the Phase II look-up table generates two additional products, acetaminophen glucuronide and acetaminophen sulfate. All three predicted products have been experimentally confirmed in published studies [123, 124].



Atom no.	Reaction center	Adjacent neighbor	Distant neighbors	Adjacent neighbor	Distant neighbors	Adjacent neighbor	Distant neighbors
1	C8y	C8x	C8x, C8y	C8x	C1a, C8y	N1b	C5a, C8y
2	C8x	C8x	C8x, C8y	C8y	C8x, C8x, N1b		
3	C8x	C8x	C8x, C8y	C8y	C8x, C8x, N1b		
4	N1b	C5a	C1a, N1b, O5a	C8y	C8x, C8x, N1b		
5	C8x	C8x	C8x, C8y	C8y	C8x, C8x, O1a		
6	C8x	C8x	C8x, C8y	C8y	C8x, C8x, O1a		
7	C5a	C1a	C5a	N1b	C5a, C8y	O5a	C5a
8	C8y	C8x	C8x, C8y	C8x	C8x, C8y	O1a	C8y
9	C1a	C5a	C1a, N1b, O5a				
10	O5a	C5a	C1a, N1b, O5a				
11	O1a	C8y	C8x, C8x, O1a				

Figure 3.3: Schematic illustration of generating the transformation products (PROXIMAL step 2). (a) The example drug, acetaminophen, is shown represented in KEGG atom type format. The .mol file was downloaded from KEGG. (b) A list of atoms comprising acetaminophen and their adjacent and distant neighbors. (c) The products predicted to result from modifications of the reaction center at atom number 11 are N-acetyl-p-benzoquinone imine, acetaminophen glucuronide and acetaminophen sulfate.

3.2.3 Step 3: Ranking the Predicted Metabolites

Several factors influence the likelihood that a particular transformation occurs. Two important, related factors are the enzyme's (catalytic) activity and abundance [125]. Another factor is whether multiple enzymes can catalyze the same transformation. In ranking the predicted products, we assume that a transformation is more likely to occur if there are many different enzymes that can catalyze the reaction, and if the enzymes are highly abundant and active. Based on this assumption, we compute the following score for each predicted transformation product.

$$score = \sum_{k} (\text{average activity})(\text{average abundance}) \tag{3.1}$$

Equation (3.1) sums the product of average activity and abundance for each enzyme that can catalyze the formation of the transformation product, with the relevant enzymes (index k) determined from the look-up tables. Values for average activity and abundance were obtained by analyzing published data. For this analysis, we focused on a subset of major Phase I (CYP) enzymes, as they play a quantitatively dominant role in human drug metabolism [126, 127, 128]. Specifically, we collected data on the following 9 CYP enzymes expressed in the human liver: 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, and 3A4. Both abundance [126, 129, 128, 130, 131] and activity [132, 133, 134, 135, 136] data were obtained from multiple studies involving primary hepatocytes from human donors. The studies were carefully selected such that the substrates used to characterize activity were similar in structure (Appendix B Table 1). To afford quantitative comparisons of data taken from different studies, we accounted for missing values and normalized the data as follows. The activity and abundance datasets were each organized into a matrix, with rows corresponding to enzymes and columns corresponding to studies. All activity data were expressed in terms of pmol substrate converted/min/mg protein, and all abundance data were expressed in terms of pmol CYP/mg protein. In the case a study did not report the abundance or activity of an enzyme, the missing value was imputed based on the row average for the enzyme. As the imputed value affects the average, this value was later iteratively recalculated during our normalization routine until it converged within a reasonable tolerance (0.1%). After estimating initial values for the missing data, each column in the activity or abundance matrix was scaled to vary from 0 to 1 by subtracting the minimum value from each column entry and dividing by the maximal range in the column.

$$a_{new} = \frac{a_{ij} - a_{j,min}}{a_{j,max} - a_{j,min}} \tag{3.2}$$

In equation (3.2), anew is the scaled activity or abundance, and the subscripts i and j refer to the enzyme and referenced study, respectively. Next, quantile normalization was performed on both abundance and activity matrices to normalize the distribution of data across studies. An iterative procedure was then applied using the results of the quantile normalization to recalculate the missing values initially estimated by averaging the values reported in the different studies. Once the values converged, a final value for CYP abundance or activity was calculated by taking a row average. These final, averaged values for each CYP were then normalized with respect to the sum of the final activity or abundance values for all 9 CYPs.

CYP	Abundance			Activity		
	Average	min	max	Average	min	max
1A2	0.0895	0.0089	0.1232	0.0414	0.0000	0.0773
2A6	0.1245	0.0906	0.1888	0.1393	0.1189	0.1664
2B6	0.0177	0.0000	0.0906	0.0814	0.0450	0.0976
2C8	0.1003	0.0906	0.1232	0.1441	0.0976	0.1664
2C9	0.2226	0.1888	0.2498	0.1198	0.0000	0.1541
2C19	0.0114	0.0000	0.0154	0.0580	0.0000	0.1541
2D6	0.0069	0.0000	0.0154	0.0985	0.0450	0.1664
2E1	0.1898	0.0994	0.2498	0.1669	0.1389	0.2018
3A4	0.2374	0.1888	0.2498	0.1506	0.0000	0.2018

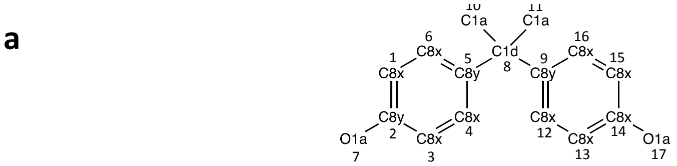
Table 3.1: Abundance and activity values for CYP subfamilies after normalization. Columns 2, 3 and 4 represent the normalized average, minimum and maximum CYP abundance values estimated from experimental data across multiple studies [126, 129, 128, 130, 131]. Columns 5, 6, and 7 represent the normalized average, minimum and maximum CYP activities estimated from experimental data across multiple studies [132, 133, 134, 135, 136]

3.3 Results

To evaluate the effectiveness of PROXIMAL in predicting xenobiotic metabolism, we investigated two test cases involving the environmental chemicals bisphenol A (BPA) and 4-chrolobiphenyl (PCB3). BPA is a synthetic chemical that has been widely used as a plasticizer, and is present in numerous commercial and household products. The primary exposure route for humans is through ingestion of food and drink, as BPA leaches from plastic containers [137].PCB3 is a persistent organic pollutant found in old electronic equipment, paints, plastics, glues and pesticides. Humans can be exposed to PCB3 through air, water or soil since the degradation rate in the environment is slow. Both chemicals can elicit biological effects in mammals that could pose potential health risks [102, 103, 105, 138]. However, it is an open question whether the observed effects are due to the parent chemical or the metabolic derivatives.

3.3.1 BPA Transformations

PROXIMAL identified a total of 17 molecular substructures in BPA (Figure 3.4a). Due to the symmetries present in the molecule, only seven of these substructures are unique (Figure 3.4b). For example, the substructure surrounding atom number 1 up to 2-level nearest neighbors is exactly the same as atom number 3. All seven of these substructures have matching keys in the look-up tables generated using the Phase I and II reaction data from DrugBank and KEGG. Figure 3.5 shows the predicted transformation products. Five of the predicted derivatives (5-hydroxy BPA, BPA glucuronide, BPA sulfate, epoxide BPA, and bisphenol-o-quinone) have been experimentally verified in published reports.



b

Atom no.	Reaction center	Adjacent neighbor	Distant neighbors	Adjacent neighbor	Distant neighbors	Adjacent neighbor	Distant neighbors	Adjacent neighbor	Distant neighbors
1	C8x	C8x	C8x, C8y	C8y	C8x, C8x, O1a				
2	C8y	C8x	C8x, C8y	C8x	C8x, C8y	O1a	C8y		
4	C8x	C8x	C8x, C8y	C8y	C1d, C8x, C8x				
5	C8y	C1d	C1a, C1a, C8y, C8y						
7	O1a	C8y	C8x, C8y, O1a						
8	C1d	C1a	C1d	C1a	C1d	C8y	C1d, C8x, C8x	C8y	C1d, C8x, C8x
10	C1a	C1a	C1a, C8y, C8y						

Figure 3.4: (a) Representation of BPA in KEGG atom type format. Each atom is represented by a number, which corresponds to the atom order in the .mol file (downloaded from KEGG), and its KEGG atom type. (b) Unique BPA atoms and their adjacent and distant neighbors are listed. Each atom of BPA is considered a reaction center, which can have up to four adjacent neighbors.

Four of the seven predicted BPA derivatives result from modifications of the reaction center at atom 1 (first molecular substructure listed in Figure 3.4b). This substructure consists of a reaction center in the aromatic ring (atom type C8x) and its neighbors (atom types C8x and C8y). Applying the matching key and value adds a hydroxyl group to the reaction center (Figure 3.5, $BPA^{I}_{(1)}$). Studies by Schmidt et al. [139] and Jaeg et al. [140] detected the presence of 5-hydroxy BPA in liver microsomes and S9 fractions prepared from mice fed BPA. Applying matching Phase II transformations to the hydroxyl group reaction center resulting from the Phase I modification generates two additional derivatives (Figure 3.5, $BPA^{II}_{(3)}$ and $BPA^{II}_{(4)}$). In addition to hydroxylation, another matching value adds an oxygen atom into the aromatic ring (Figure 3.5, $BPA^{I}_{(3)}$). The resulting arene epoxide is essentially identical to a previously reported BPA derivative [139], except for the position of the epoxide group.

Another molecular substructure recognized as a key is the hydroxyl group (atom number 7) attached to the aromatic ring (5^{th} molecular substructure listed in Figure 3.4b). This key has one value in the Phase I look-up table. The modification specified by this value is to change the hydroxyl group into a carbonyl group (Figure 3.5 $BPA^{I}_{(2)}$). The resulting derivative is similar to a previously detected quinol product [141], only differing by a missing hydroxyl group on the ipso carbon. Applying Phase II transformations directly on the hydroxyl group generates two conjugation products, BPA glucuronide and sulfate (Figure 3.5, $BPA^{II}_{(1)}$ and $BPA^{II}_{(2)}$). Several studies have shown that BPA extensively metabolizes into BPA glucuronide and BPA sulfate in humans, rats, and mice [142, 143, 144, 145].

After identifying the possible transformation products, we rank each predicted metabolite by computing a score that reflects the number of different enzymes that can carry out the predicted transformation as well as published data on the activity and abundance of these enzymes. As published data were more extensive for CYP enzymes compared to conjugation enzymes, we restricted the analysis to ranking only the Phase I products. The enzyme activity and abundance values used for this analysis are shown in Table 3.1. The scores and rankings for the Phase I derivatives of BPA are shown in Table 3.2. The first and second columns of the table show the names of the derivatives using the nomenclature from Figure 3.5 and the CYP enzyme families responsible for the transformation. We demonstrate the score calculation with an example. The transformation to 5-hydroxy BPA ($BPA^{I}_{(1)}$) can be catalyzed by any one of four CYP enzymes, namely 1A1, 1A2, 1B1, and 3A4. The average scores for the enzyme activity of 1A2 and

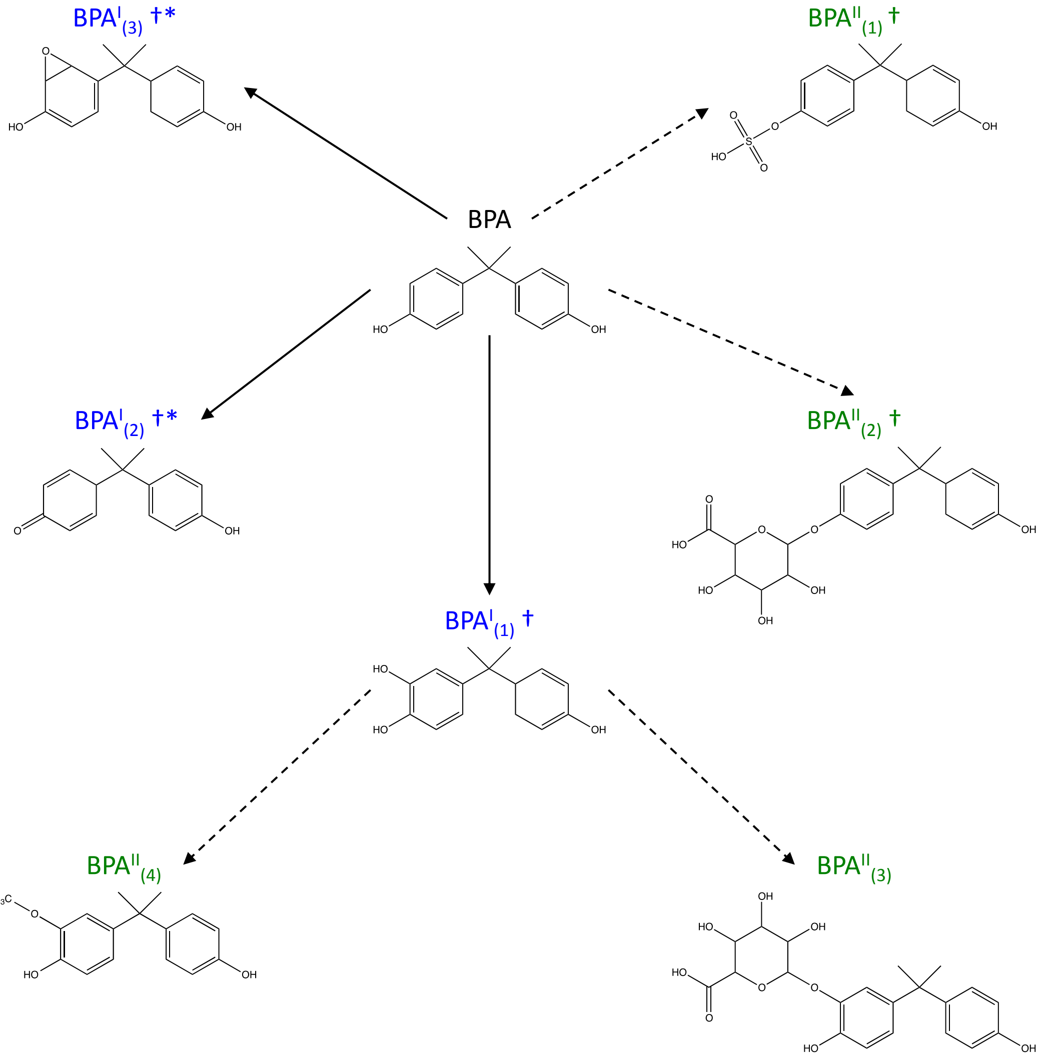


Figure 3.5: Predicted biotransformation products for BPA. The solid and dashed lines represent biotransformation through Phase I and Phase II, respectively. The symbol † indicates that the exact predicted compound has been verified by experimental data, whereas the symbol * indicates the predicted compound is similar to but not exactly the same as the structure reported in literature.

Metabolites	CYPs	Score	Rank	Literature
$BPA^I_{(1)}$	$1A1^x$, 1A2, $1B1^x$,3A4	0.0395	1	[140, 139]
$BPA^I_{(2)}$	1A2, 2D6, 2E1	0.0361	2	[141]*
$BPA^I_{(3)}$	2E1	0.0317	3	[139]*

Table 3.2: Score and ranking for the predicted products of Phase I biotransformation of BPA. The first column shows the predicted compounds resulting from Phase I biotransformation. The notation is same as in Figure 3.5. The second column indicates the CYP families responsible for the biotransformations. The superscript ^{x} is added to enzymes for which activity/abundance data in human liver samples were unavailable, and the enzymes were thus not included in the score calculation. The 3^{rd} and 4^{th} columns show the calculated scores and rank. The last column lists the references reporting the predicted compound. The superscript * indicates that the predicted compound is similar to but not exactly the same as the reported structure.

3A4 are 0.0414 and 0.1506, respectively (Table 3.1). Their corresponding enzyme abundance scores are 0.0895 and 0.2374. We calculated the score for $BPA^I_{(1)}$ as $(0.0414 \times 0.0895 + 0.1506 \times 0.2374) = 0.0395$. We did not include CYP 1A1 and 1B1 in the calculation since data for these enzymes were unavailable. The scores in Table 3.2 indicate that hydroxylated BPA should be the dominant derivative. This ranking is consistent with a human subjects study by Schmidt et al. [139], which found that hydroxylated forms of BPA were the most abundant in liver microsomes.

3.3.2 PCB3 Transformations

The second case study analyzed the modification and conjugation of PCB3 via Phase I and II reactions. Figure 3.6 shows the atom numbers and atom types for PCB3. PROXIMAL identifies a total of 13 molecular substructures in PCB3 where only 8 of these substructures are unique. In total, PROXIMAL predicts 26 derivatives (Figure 3.7).

Four of predicted PCB3 derivatives result from modifications of the substructure around atom number 12 (Figure 3.6), which consists of a reaction cen-

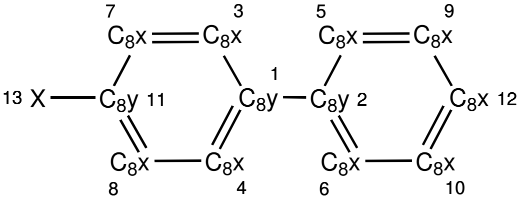


Figure 3.6: PCB3 representation in KEGG atom type format. Each atom is represented by a number, which corresponds to the atom order in the .mol file (from KEGG), and its KEGG atom type.

ter in the aromatic ring (atom type C8x) and its neighbors (atom types C8x and C8x). The derivatives are 4-hydroxy-PCB3 (Figure 3.7, $PCB^{I}_{(2)}$), 3,4-dihydroxy-PCB3 ($PCB^{I}_{(5)}$), epoxide PCB3 ($PCB^{I}_{(6)}$) and cis-3,4-dihydro-3,4-dihydroxy-PCB3 ($PCB^{I}_{(8)}$). Each of these derivatives can be further transformed by applying matching conjugation steps identified from the Phase II look-up table. The hydroxyl group reaction center added via Phase I modification in 4-hydroxy-PCB3 (Figure 3.7 $PCB^{I}_{(2)}$) generates 4-PCB3-sulfate ($PCB^{II}_{(3)}$) and 4-PCB3-glucuronide ($PCB^{II}_{(4)}$). Phase II derivatives of 3,4-diOH-PCB3 ($PCB^{I}_{(5)}$) include its glucuronated ($PCB^{II}_{(9)}$) and methylated conjugates ($PCB^{II}_{(8)}$). The epoxide group on the non-chlorinated aromatic ring of PCB3 ($PCB^{I}_{(6)}$) can be reduced and conjugated through Phase II enzymes to form PCB3 glutathione (Figure 3.7 $PCB^{II}_{(10)}$ and $PCB^{II}_{(11)}$). Applying a Phase II transformation to cis-3,4-dihydro-3,4-dihydroxy-PCB3 (Figure 3.7 $PCB^{I}_{(8)}$) generates a glucuronic acid conjugate.

Another major reaction center is one of the aromatic ring carbons (atom number 9). Applying Phase I modifications on this substructure generates an arene oxide PCB3 ($PCB^{I}_{(1)}$) and 3-hydroxy-PCB3 (Figure 3.7 $PCB^{I}_{(4)}$). Like 4-hydroxy-PCB3 ($PCB^{I}_{(2)}$), 3-hydroxy-PCB3 can be further conjugated with a sulfate ($PCB^{II}_{(6)}$) or glucuronide group ($PCB^{II}_{(7)}$). Similarly, the arene oxide product ($PCB^{I}_{(1)}$) can also be conjugated with glutathione ($PCB^{II}_{(1)}$ and $PCB^{II}_{(2)}$) through Phase II enzymes.

The third Phase I reaction center is located at atom number 7. Modification of the corresponding substructure again produces a hydroxylated PCB3 ($PCB^{I}_{(7)}$) as well as 3,4-dichlorobiphenyl ($PCB^{I}_{(9)}$). Phase II transformation of the hydroxylated derivative can generate a methylated ($PCB^{II}_{(12)}$) or mono-glucuronide conjugate ($PCB^{II}_{(13)}$). The remaining Phase I transformation products, an epoxide ($PCB^{I}_{(3)}$) and 2-hydroxy-PCB3 ($PCB^{I}_{(10)}$), derive from modifications of reaction centers at atom numbers 3 and 5. The corresponding Phase II derivatives include a glutathione ($PCB^{II}_{(5)}$), glucuronide ($PCB^{II}_{(15)}$) and methylated conjugate ($PCB^{II}_{(16)}$).

In corroborating our predictions on PCB transformations with published reports, we expanded the literature search to include studies involving rodents, as there have been only few studies involving primary human liver cells. A recent study by Dhakal et al. [107] examined the metabolism and toxicity of PCB3 in male rats by analyzing urine samples collected following a bolus intra-peritoneal injection of the chemical. Using MS analysis, the authors identified several Phase I and Phase II products, including 2-, 3-, 4-hydroxy-PCB3, and their corresponding sulfate and glucuronide conjugates. With the exception of 2-PCB3-sulfate, these derivatives were also identified by our prediction method (Figure 3.7). The same study [107] also reported the amount of 4-hydroxy-PCB3 (Figure 3.7, $PCB^{I}_{(2)}$) in the urine samples was approximately 10 times greater than 3-hydroxy PCB3 ($PCB^{I}_{(4)}$). A separate study analyzing the distribution of hydroxylated PCB3 derivatives in rat liver microsomes found that the most abundant forms were, in decreasing order, 4-OH-PCB3, 3-OH-PCB3, and 2-OH-PCB3 [146]. This is in good agreement with the relative ranking of these three derivatives computed from enzyme activity and

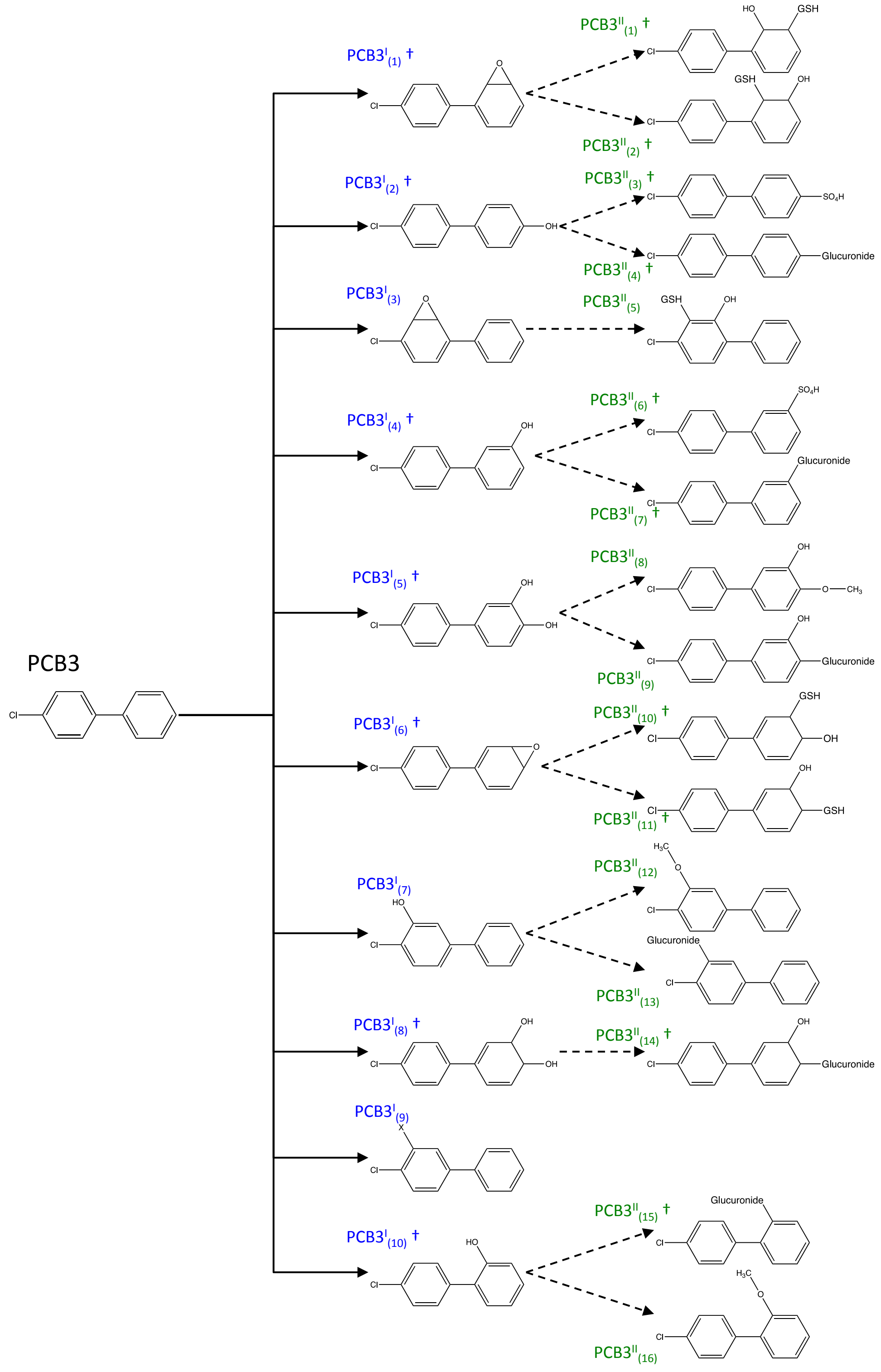


Figure 3.7: Predicted biotransformation products for PCB3. The solid and dashed lines represent the biotransformation reactions mediated by Phase I and Phase II enzymes, respectively. The symbol † indicates that the exact compound has been experimentally observed.

abundance data (Table 3.3).

Additional derivatives predicted by PROXIMAL and experimentally confirmed in the study by Dhakal et al. include dihydrodiol (Figure 3.7, $PCB^{I}_{(8)}$), dihydrodiol glucuronide ($PCB^{II}_{(14)}$) and 3,4-dihydroxy-PCB3 ($PCB^{I}_{(5)}$). Dhakal et al. also reported detecting molecules with mass signatures that correspond to the arene epoxide derivatives (Figure 3.7, $PCB^{I}_{(1)}$ and $PCB^{I}_{(6)}$) and their glutathione conjugates ($PCB^{II}_{(1)}$, $PCB^{II}_{(2)}$, $PCB^{II}_{(10)}$ and $PCB^{II}_{(11)}$) predicted by PROXIMAL; however, these compounds could not be confirmed due to lack of pure chemical standards. In a separate study involving rat livers, Lehmann et al. [147] showed that PCB3 can be transformed into several active electrophiles, including arene oxides, which may bind to DNA, RNA and/or hemoglobin to cause cellular damage and increase the frequency of mutations. Transformation of PCB3 into arene oxide derivatives has also been observed in a study by McLean et al. [146] involving rat liver microsomes. Altogether, we were able to confirm 17 out of the 26 predicted PCB3 derivatives based on experimental data published in other studies.

3.4 Discussion

In this work, we present a computational method, termed PROXIMAL, for predicting the transformation of xenobiotic chemicals by human CYP oxidoreductases and transferases. We evaluated the predictive power of the method by investigating case studies involving two prevalent environmental contaminants, BPA and PCB3, which are increasingly associated with developmental disorders and metabolic diseases. Overall, we found strong corroborating evidences in the literature for the predicted transformations of these two chemicals. In the case of BPA, we could

Metabolites	CYPs	Score	Rank	Literature
$PCB^I_{(1)}$	1A1^x, 1B1^x, 1A2, 2A6, 2B6, 2C9, 2C19, 2E1, 3A4	0.1172	1	[107, 147, 146]
$PCB^I_{(2)}$	1B1^x, 2B6, 2C8, 2C9, 2C18^x, 2C19, 2D6, 2E1, 3A4, 3A5^x	0.1113	2	[107, 146]
$PCB^I_{(3)}$	1A1^x, 1B1^x, 2C8, 2C9, 2E1, 3A4	0.1086	3	
$PCB^I_{(4)}$	2B6, 3A4	0.0372	4	[107, 146]
$PCB^I_{(5)}$	3A4	0.0358	5	[107]
$PCB^I_{(6)}$	2E1	0.0317	6	[107, 147, 146]
$PCB^I_{(7)}$	2E1	0.0317	7	
$PCB^I_{(8)}$	2C9, 2C19	0.0273	8	[107]
$PCB^I_{(9)}$	1A2	0.0037	9	
$PCB^I_{(10)}$	2D6	0.0007	10	[107, 146]

Table 3.3: Score and ranking for the predicted products of Phase I biotransformation of PCB3. The first column shows the predicted compounds resulting from Phase I biotransformation. The notation is the same as in Figure 3.7. The second column indicates the CYP families responsible for the biotransformations. The superscript ^{x} is added to enzymes for which activity/abundance data in human liver samples were unavailable, and the enzymes were thus not included in the score calculation. The 3^{rd} and 4^{th} columns show the calculated scores and rank. The last column lists the references reporting the predicted compound.

confirm five of the seven predicted derivatives. In the case of PCB3, we confirmed 17 out of the 26 predictive derivatives, although we should note that the literature comparisons were based on studies that used animal models. Additional studies on PCB3 transformations in humans would be needed to further validate the predictions.

It is important to point out that the lack of literature evidence does not necessarily imply that a prediction is false. It is possible that certain metabolic transformations of BPA and PCB3 have not yet been observed due to the instability of the products or some other difficulty in detecting these derivatives. Another reason could be that these products were outside the scope of a targeted analysis. One way to more comprehensively validate computational predictions on xenobiotic transformation is to perform untargeted metabolomics studies, for example using high-resolution MS. However, assigning a chemical identity to every ion detected in a full-scan MS experiment remains a difficult task. In this regard, pairing experimental investigation of xenobiotic transformation with computational exploration would be extremely useful, especially for processing complex MS data generated through untargeted analysis. The predicted chemical structures can be used to calculate the expected mass signatures (accurate masses), which then can be queried against spectral data from a high-resolution MS experiment. This type of informed search streamlines the data processing, and could help avoid false negatives that result from relying on metabolite databases that likely contain only a small, known subset of xenobiotic transformation products. The predicted chemical structures can also be used to calculate isotope patterns and MS/MS fragmentation patterns, which are crucial in confirming the identity of detected ions, particularly when high-purity standards are unavailable for the chemicals of interest. Furthermore, having a priori knowledge of the expected derivatives can guide the experimental workflow, including the choice of solvent for sample extraction and the method for chromatographic separation.

The present study focused on environmental pollutants to illustrate and evaluate our prediction method. In addition to organic pollutants, PROXIMAL could also be used for predicting transformations of other types of chemicals that contain substructures recognizable by Phase I and Phase II enzymes. Examples include drugs as well as various phytochemicals, e.g. phenolic compounds. Drug toxicity often arises from metabolic activation; *i.e.* the derivatives of a drug can be more toxic than the drug [148]. In this light, PROXIMAL could be used in conjunction with toxicity prediction software such as ADMET Predictor and Derek Nexus to assess the toxic potential of a drug compound's possible derivatives that could form endogenously following the drug's administration. Prospectively, this type of analysis could become part of an in silico screen designed to ultimately reduce the chance of drug-induced liver injury, which is a leading concern during drug development and testing [149, 150].

In addition to predicting the chemical structures of potential biotransformation products, PROXIMAL also provides a relative ranking of these products in terms of their likely occurrence. The ability to predict quantitatively dominant derivatives for a given chemical could complement experimental approaches, for example by informing the selection of metabolites for targeted analysis. For the two chemicals examined in our study, the predicted products with the highest ranking were also the derivatives that were frequently reported in the literature. This suggests that our ranking scheme could be biologically relevant. The caveat here is that the literature reports could reflect not only the prevalence of these derivatives, but also the level of interest in these compounds by the researchers. Clearly, a more thorough validation, for example using untargeted analysis, will be needed to evaluate the accuracy of the rankings.

An implicit assumption of our ranking scheme is that the likelihood of forming a particular derivative depends primarily on the kinetics of the biotransformation. We further assumed that the kinetics depends on the total abundance and activity of the enzymes involved. For example, we predicted that the hydroxylated forms of BPA are more likely to occur compared to other derivative forms. This result reflects the relatively large number (hence total abundance and activity) of CYP enzymes that can mediate the hydroxylation reaction. This prediction is consistent with experimental measurements on microsomes collected from pooled human liver samples [139].

Our ranking scheme does not take into account whether a particular transformation reaction is energetically more favorable, *i.e.* yields a more negative change in free energy, than other possible transformation reactions. Consequently, derivatives that are formed by the same set of enzymes will have the same rank. In contrast, derivatives formed by different sets of enzymes, including structural isomers, will have different rank. For example, our ranking scheme predicted that different hydroxyl isomers of PCB3 would be formed with different likelihoods. Specifically, we predicted that 4-OH-PCB3 is the most abundant derivative, followed by 3-OH-PCB3 and 2-OH-PCB3, in order. Interestingly, this prediction is consistent with experimental data [107, 146]. As shown in Table 3.3, CYP1B1 can catalyze the formation of 4-OH-PCB3 ($PCB^{I}_{(2)}$) but not the formation of 3-OH-PCB3 ($PCB^{I}_{(4)}$), whereas CYP3A4 can catalyze the formation of both isoforms. These differences between the two CYP enzymes reflect the reaction pattern information in the lookup tables, and imply that the extent of substrate flexibility varies from one CYP enzymes to another. It has been shown that CYP3A4 exhibits a large degree of flexibility, and can add a hydroxyl group to several different carbon atoms in a substrate molecule [151].

A limitation of our ranking method is that it does not include Phase II products. This was primarily due to insufficient information regarding Phase II enzymes. After an extensive literature search, we found only one study [133] that reported the specific activities of all six conjugation enzymes considered in the present work. As data become available, a ranking analysis could be performed based on relative enzyme activity and abundance similar to the analysis of Phase I enzymes. An additional factor to consider is that formation of the conjugation products depends on Phase I modification, as the atoms introduced in this step form the reaction centers for Phase II conjugation. As such, Phase I rankings may also need to be considered in ranking conjugation products. Yet another factor to consider is the availability of cofactors and conjugation substrates. For example, glutathione is a major antioxidant in the liver, and can become a limiting reactant under oxidative stress conditions.

Lastly, our ranking scheme did not include regulatory effects such as induction or inhibition of CYPs by the xenobiotic chemicals. It is well known that BPA can selectively induce or inhibit metabolic activity of certain CYPs. For example, Cannon et al. [110] reported on both inhibition and induction of BPA on CYP activity in human liver S9 fractions. One way to improve our ranking scheme is to include enzyme regulation would be to introduce an activity adjustment factor. For a given xenobiotic of interest, this factor would adjust the baseline CYP activities used in the present study to account for the inhibitory or inducing effects of the xenobiotic and its predicted derivatives. Clearly, this approach would require a substantial amount of additional information on the regulatory effects of xenobiotic chemicals. For the present study, we did not find sufficient data to confidently determine the inhibitory or inducing effects of BPA or PCB3 on the relevant CYP enzymes. As is the case for predicting the structural modifications resulting from biotransformation, a purely experimental approach will likely be intractable to determine the regulatory effects of xenobiotic chemicals on CYP enzymes, which may be mediated by ligand-activated nuclear receptors [152] controlling the expression of the enzymes. In this regard, computational approaches are warranted, for example to identify patterns in the chemical structures of known ligands for the regulatory molecules.

In summary, we present in this chapter a method to predict xenobiotic metabolism. To demonstrate the method, we applied the method to two case studies of endocrine disrupting environmental chemicals, and successfully predicted biotransformation products reported in the literature. We found experimental evidence for 71 and 65% of the predicted BPA and PCB3 metabolites, respectively. Our method uses known chemical modifications found in reaction databases such as DrugBank and KEGG in conjunction with SIMCOMP to predict xenobiotic transformations for a given compound of interest. A novel aspect is the ability to rank the predicted metabolites based on available information regarding the activity and abundance of the transformation enzymes. While the scope of this ranking was limited, we found good agreement between the predictions and findings reported in the published literature. In the discussion, we identify several limitations of the present prediction method that reflect the relative scarcity of information on Phase II enzymes and the regulation of xenobiotic transformation enzymes. Further studies, preferably in human cells, are warranted to further improve the predictive power and physiological relevance of PROXIMAL.

3.5 Conclusion

We presented in this chapter a new method, PROXIMAL, for predicting the biotransformation of xenobiotics by human enzymes. PROXIMAL analyzes Phase I and Phase II xenobiotic transformations that are cataloged in public databases (*i.e.* DrugBank [96, 97, 98] and KEGG [15, 99]), and builds look-up tables linking specific molecular substructures with matching biotransformation operations that modify these substructures. To achieve specificity, the look-up tables take into account molecular substructures that consist of a reaction center and its two-level nearest neighbors. Given a query compound, PROXIMAL applies a select set of transformations from the look-up tables at one or more matching sites, or reaction centers, of the query compound. PROXIMAL then ranks the transformation results based on the activity and abundance of the enzymes involved in the transformations. To evaluate the predictive power of PROXIMAL, we investigated two case studies involving bisphenol A (BPA) and 4-chlorobiphenyl (PCB3), two environmental chemicals with suspected endocrine disrupting activity. Our findings were corroborated with evidence from the literature. Importantly, the approach underlying PROXIMAL is general, and can be applied to construct synthesis pathways for other classes of metabolites that are not catalogued in current databases.

Chapter 4

Probabilistic Strain Optimization under Constraint Uncertainty

While the methods in previous chapter concern the construction of synthesis pathways, this chapter focuses on tuning the host to maximize the production of a desired native or non-native metabolite. An important step in this process is to identify reactions whose activities should be modified to achieve the desired cellular objective. Preferably, these reactions are identified systematically, as the number of possible combinations of reaction modifications could be very large. Over the last several years, a number of computational methods have been described for identifying combinations of reaction modifications. However, none of these methods explicitly address uncertainties in implementing the reaction activity modifications. In this chapter, we model the uncertainties as probability distributions in the flux carrying capacities of reactions. Based on this model, we develop an optimization method that identifies reactions for flux capacity modifications to predict outcomes with high statistical likelihood.

4.1 Methods

We investigate three computational methods to address uncertainty in strain optimization. Specifically, we compare two probabilistic methods, CCP based optimization (CCOpt) and sampling based optimization (MCOpt), against deterministic optimization (DetOpt). The performance of each method is tested on two metabolic models for which enzyme level changes and corresponding flux capacity distributions are estimated either from kinetic parameters or steady-state flux data. The performance of the solutions, *i.e.* predicted target fluxes and corresponding intervention sets, is evaluated using Monte Carlo simulations (MCEval) designed to simulate the variable outcomes resulting from experimental implementation of the modifications specified by the optimization solutions.

4.1.1 Chance-Constrained Optimization (CCOpt)

Figure 4.1 illustrates the difference between a deterministic and probabilistic interpretation of an uncertain upper-bound constraint on the flux of reaction j. In the deterministic interpretation, the value of flux v_j of any feasible solution is enforced to be strictly less than all of the values in the upper-bound (flux-capacity) distribution Cap_j^u. This yields the constraint:

$$Prob\{v_j < Cap_j^u\} = 1 \tag{4.1}$$

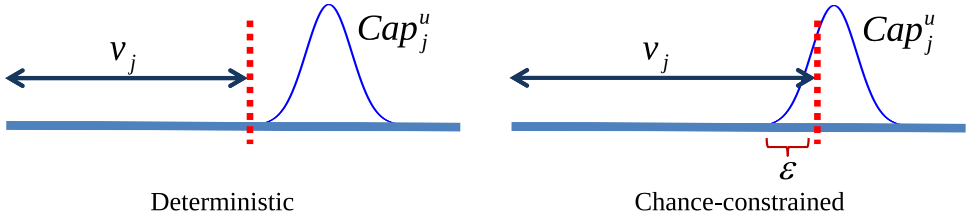


Figure 4.1: Deterministic and chance-constrained interpretation of an upper bound on reaction flux. The dotted lines represent the upper bound for the flux of a reaction j in a deterministic (left panel) and chance-constrained interpretation (right panel). The arrows show the flux ranges. If the upper bound is a random variable, the deterministic interpretation forces the flux v_j below the lowest value in the upper bound distribution. The chance-constrained interpretation allows v_j to exceed the lowest value in the upper bound distribution by some probability specified by the parameter ϵ.

In the probabilistic interpretation, the constraint is not always satisfied, *i.e.* there is a nonzero probability that flux v_j will be equal to or larger than some of the values in the distribution Cap_j^u. In the case of CCP, the constraint is relaxed by introducing a parameter ϵ, which reflects the confidence level for the probability that the solution satisfies the constraint:

$$Prob\{v_j < Cap_j^u\} \geq 1 - \epsilon \tag{4.2}$$

To generalize the previous inequality to also consider the effects of up- or down-regulating the activity of an enzyme (e.g. through an adjustment in the expression of the gene that encodes the enzyme), we introduce two sets of binary decision variables y_j^u and y_j^d. In this chapter, we use the phrasing "up- or down-regulation" to describe engineering modifications that result in expression level changes of enzymes or groups of enzymes regardless of the method. A value of 1 indicates that the corresponding enzyme is up- or down-regulated, whereas a value of 0 indicates the corresponding enzyme expression is unchanged.

$$Prob\{v_j \leq (1-y_j^u).(1-y_j^d).SSU_j+y_j^u.(1-y_j^d).Cap_j^u+y_j^d.(1-y_j^u).Cap_j^d\} \geq 1-\epsilon \quad (4.3)$$

In equation (4.3), SSU_j denotes the reference (unmodified) state upper bound for reaction j. The fact that there are two random variables (Cap_j^u and Cap_j^d) does not pose a challenge in solving such an inequality, as at most one of them will have a nonzero coefficient at a time. Mathematically, the sum of the two decision variables must be less than or equal to one ($y_j^u + y_j^d \leq 1$), which simplifies the above inequality into the following:

$$Prob\{v_j \leq SSU_j + y_j^u.(Cap_j^u - SSU_j) + y_j^d.(Cap_j^d - SSU_j)\} \geq 1 - \epsilon \qquad (4.4)$$

A graphical illustration of the probabilistic constraints is shown in Figure 4.2. Down-regulating a reaction decreases the upper bound, or the flux capacity. It could also decrease the lower bound to zero. The capacity change could leave the flux unchanged or decrease it below the level of the reference (unmodified) state lower bound. Up-regulating a reaction increases the flux capacity, but does not affect the lower bound. The flux value could remain the same or rise above the reference state upper bound. In this study, we model the capacity change resulting from a gene expression modification as a probabilistic (rather than deterministic) event, which leads to a flux capacity distribution (dashed red lines).

Various approaches have been developed to solve CCP problems based on properties such as the distribution of random variables, linearity, and type (individual or joint) of the chance constraints [71]. One method to solve a CCP problem is

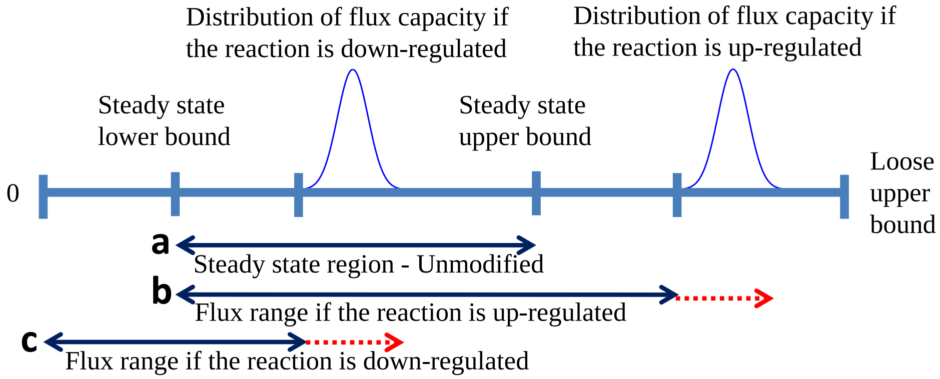

Distribution of flux capacity if the reaction is down-regulated
Distribution of flux capacity if the reaction is up-regulated
Steady state lower bound
Steady state upper bound
0
Loose upper bound
a
Steady state region - Unmodified
b
Flux range if the reaction is up-regulated
c
Flux range if the reaction is down-regulated

Figure 4.2: Chance-constrained reaction flux bounds with or without enzyme level changes. (a) When there is no modification in the enzyme level, the flux for reaction j lies within the reference state range. (b) When the reaction is up-regulated, the upper bound distribution shifts above the reference state upper bound. (c) When the reaction is down-regulated, the upper bound distribution shifts below the reference state upper bound and the new (modified state) lower bound may also shift below the reference state lower bound. The red dashed lines in (b) and (c) show the range of possible flux capacity values equaling the spread of the capacity distributions.

to convert the probabilistic constraints (here, equation (4.4)) into their deterministic equivalents at their specified confidence level ϵ. This approach requires that the random variables of the problem are independent, and appear only in an exclusive linear form, such that the coefficients of all but one are always zero [153]. Our formulation meets all of these conditions; therefore, the chance constraints can be converted into their deterministic equivalents. Using the inverse of the cumulative distribution functions (CDF) for Cap_j^u and Cap_j^d, inequality (4.4) can be reformulated as:

$$v_j \leq SSU_j + y_j^u.(F_{j,u}^{-1}(\epsilon) - SSU_j) + y_j^d.(F_{j,d}^{-1}(\epsilon) - SSU_j) \tag{4.5}$$

where $F_{j,u}^{-1}(\epsilon)$ and $F_{j,d}^{-1}(\epsilon)$ denote the inverse CDFs of Cap_j^u and Cap_j^d respectively, which can be numerically calculated if needed.

Recasting the chance constraints into the equivalent deterministic constraints, the uncertain optimization problem of maximizing the flux of a desired product

Variable name	Definition
v_{target}	Flux variable for the target reaction
v_j	Flux variable for reaction j
$v_{biomass}^{max}$	Theoretical maximum growth of the wild-type (unmodified) organism
y_j^u	Binary variable for reaction j (1 if up-regulated, 0 otherwise)
y_j^d	Binary variable for reaction j (1 if down-regulated, 0 otherwise)
SSU_j	Reference (unmodified) state upper bound for reaction j
SSL_j	Reference (unmodified) state lower bound for reaction j
$F_{j,u}^{-1}$	Inverse CDF of Cap_j^u
$F_{j,d}^{-1}$	Inverse CDF of Cap_j^d
Cap_j^u	Random variable for flux capacity of reaction j when up-regulated
Cap_j^d	Random variable for flux capacity of reaction j when down-regulated
ϵ	Confidence level desired by the user
α	Small penalty for each added intervention
L	Maximum number of allowed interventions
S_{ij}	Stoichiometric coefficient for metabolite i and reaction j
M	Reaction set
N	Metabolite set

Table 4.1: CCOpt variables and their definitions.

through gene up/down-regulation operations can be formulated for a system of arbitrary size consisting of N metabolites and M reactions. Without loss of generality, reversible reactions are split into forward and backward components such that the reaction set comprises only irreversible reactions. In this formulation, the inputs are the stoichiometric matrix, S, a small penalty for each added intervention, α, theoretical maximum biomass of the wild-type organism, steady-state boundaries, and inverse CDF of capacity distributions for each reaction. The outputs are a maximum target flux value and an intervention set. Table 4.1 defines the optimization variables. The chance-constrained cell optimization problem has the following constraints:

$$maximize(v_{target} - \alpha . \sum_{j=1}^{|M|}(y_j^u + y_j^d)) \tag{4.6}$$

s.t.

$$\sum_{j=1}^{|M|} S_{ij} v_j = 0, \forall i \in N \tag{4.7}$$

$$v_{biomass} \geq (0.01).v_{biomass}^{max} \tag{4.8}$$

$$v_j \leq SSU_j + y_j^u.(F_{j,u}^{-1}(\epsilon) - SSU_j) + y_j^d.(F_{j,d}^{-1}(\epsilon) - SSU_j), \forall j \in M \tag{4.9}$$

$$v_j \geq SSL_j.(1 - y_j^d), \forall j \in M \tag{4.10}$$

$$\sum_{j=1}^{|M|} (y_j^u + y_j^d) \leq L \tag{4.11}$$

$$y_j^u + y_j^d \leq 1, \forall j \in M \tag{4.12}$$

$$y_j^d + y_k^d \leq 1, y_j^u + y_k^u \leq 1, \forall j \in M, k = j\text{'s backward counterpart} \tag{4.13}$$

$$y_j^u \in \{0,1\}, y_j^d \in \{0,1\}, \forall j \in M \tag{4.14}$$

The main objective of the problem is to maximize the target reaction flux v_{target}. It is expected that the optimal value of v_{target} will increase monotonically with the number of allowed interventions (enzyme up/down-regulation operations) L. On the other hand, the engineering cost is also expected to increase with the number of interventions. Therefore, the objective function in (4.6) also includes the term $-\alpha. \sum_{j=1}^{|M|} (y_j^u + y_j^d)$, which imposes a small penalty for each added intervention, and balances the optimal flux of the target reaction against the number of required interventions. Matrix S in (4.7) is a stoichiometric matrix of a biochemical network where each entry in the matrix represents the stoichiometric coefficient of a compound participating in a particular reaction. Each column describes a reaction. A column entry is zero if the compound does not participate in the reaction, positive if the compound is a product and negative if the compound is a reactant. Each row in S specifies the mass balance relationship for a particular metabolite. fConstraint (4.7) represents the steady state assumption that the rate of production of each intracellular metabolite is equal to its rate of consumption. Constraint (4.8) guarantees a minimal growth rate equaling at least 1% of the theoretical maximum of the wild-type (unmodified) organism. A minimal growth rate constraint is required to guarantee that the cell remains viable. This parameter can be adjusted by the user based on the metabolic model, available data and expectations for cell viability, which does not alter the optimization algorithm. To maximize the growth rate while simultaneously maximizing a certain target metabolite, a bi-level optimization with two objectives (maximizing biomass and a target flux) can be applied in place of the constraints (4.6) and (4.8). However, linear bi-level programs are NP-hard [154] and there are no efficient algorithms to solve large-scale problems [155]. Constraint (4.9) sets the upper bound flux capacity for each reaction j. SSU_j shows the reference state upper bound for reaction j. $F_{j,u}^{-1}$ and $F_{j,d}^{-1}$ denote the inverse CDFs of Cap_j^u (the capacity distribution for reaction j if up-regulated) and Cap_j^d (the capacity distribution for reaction j if down-regulated) respectively. Constraint (4.10) sets the lower bound flux for each reaction j to SSL_j (an observed reference state lower bound, if the observation data is available) or zero, based on the value of the binary variable y_j^d. Constraint (4.11) sets an upper bound on the number of allowed interventions to L, a constant value. Inequality (4.12) ensures that enzyme manipulations are exclusive, *i.e.* a reaction can be either up- or down-regulated in a solution, but not both. Similarly, constraint (4.13) guarantees that the forward and backward directions of a reversible reaction are not both up- and down-regulated at the same time. Constraint (4.14) specifies that the decision variables y_j^u and y_j^d can only be 0 or 1.

4.1.2 Deterministic Optimization (DetOpt)

The deterministic formulation (DetOpt) can be derived from the CCP formulation by setting $\epsilon = 0$ in (4.9), *i.e.* v_j is strictly less than all possible values the random variables Cap_j^u or Cap_j^d can take.

4.1.3 Monte Carlo-based Optimization (MCOpt)

Chance-constrained optimization can be emulated by repeatedly solving the fixed constraint (deterministic) optimization problem in which the constraint parameters (Cap_j^u or Cap_j^d) are set to randomly drawn values using a MC sampling procedure for each instance of the problem. The MC sampling requires *a priori* knowledge of the distributions for the flux capacities (Cap_j^u/Cap_j^d). The procedure for computing the distributions is described below. Using the randomly drawn set of flux capacities, the capacity constraints become fixed constraints. Effectively, we replace the inequality in (4.9) with the constraint below:

$$v_j \leq SSU_j + y_j^u.(X_j^u - SSU_j) + y_j^d.(X_j^d - SSU_j), \forall j \in M \tag{4.15}$$

where X_j^u and X_j^d are the randomly drawn set of flux capacities. Each MC sample, *i.e.* set of randomly drawn flux capacities, defines an instance of an optimization problem. The solution to this optimization problem is a set of interventions and a corresponding optimal flux value for the target reaction. Repeating the process (sampling and optimization) many times, we obtain a distribution of optimal target flux values.

4.1.4 Computing Capacity Distributions

Traditionally, a gene up/down-regulation operation has been modeled as a deterministic event leading to a fold-change in the level of the corresponding enzyme, and hence a fold-change in the flux capacity of the reaction catalyzed by the enzyme. Here, we model enzyme level modification as an uncertain event using a probability distribution. We assume a normal distribution [156] with an average fold-change of $\mu = 6$ following gene up-regulation and a spread of $\delta = 6\sigma = 8$, where σ denotes the standard deviation. The average fold-change value reflects experimental data reported in gene over-expression studies involving mammalian cells, specifically adipocytes [157]. We note that the average fold-change value is a user-specified parameter that can be adjusted to reflect different cell types and experimental data, and thus does not lead to loss of generality. The spread δ is chosen so that $\mu - \delta/2 > 1$, which ensures that the flux capacity after up-regulating the enzyme level is higher than the unmodified state. A decrease in enzyme level, and hence reaction flux capacity, is modeled by a normal distribution $N_d(\mu, \sigma^2)$ with an average fold-change of $\mu = 0.5$ and a spread of $\delta = 1$.

Based on the probabilistic interpretation of fold-changes in enzyme levels resulting from gene modifications, we also estimate the resulting reaction flux capacities as probability distributions. We use two different estimation methods depending on whether the model is kinetic or stoichiometric. In the case of a kinetic model, a fold-change in enzyme level is assumed to directly correlate with a fold-change in the maximal reaction velocity ($v_{j,max}$). Here, the maximal reaction velocity has the same units as reaction flux. Therefore, flux capacity distributions were calculated by simply multiplying the enzyme fold-change distributions with $v_{j,max}$. In the case of a stoichiometric model, the distributions of flux capacities are approximated using enzyme control flux (ECF) analysis [77]. Briefly, ECF analysis calculates the effect of enzyme level changes on flux distribution based on elementary mode analysis [158] and a power law model for the relationship between reaction flux and enzyme activity. Typically, the ECF problem is underdetermined, and the solution is obtained as a range of minimal and maximal flux for each reaction. We use the maximal flux value as the corresponding reaction flux capacity. The maximal flux values, calculated using sample points from the distributions of enzyme level modifications ($N_u(\mu, \sigma^2)$ and $N_d(\mu, \sigma^2)$), form a capacity distribution.

4.1.5 Monte Carlo-based Evaluation (MCEval) Framework

We evaluate CCOpt, DetOpt, and MCOpt using Monte Carlo (MCEval) simulations designed to mimic the expected variations in outcomes when the intervention sets identified by the three different optimization methods are experimentally implemented. For CCOpt and DetOpt, each solution is a single optimal flux of the target reaction and a corresponding set of interventions. The MCOpt solution comprises a distribution of maximal fluxes and their corresponding sets of interventions. To compare these solutions, we perform separate MCEval simulations using the interventions obtained from CCOpt, DetOpt, and MCOpt, and apply flux balance analysis (FBA) [78] with the objective function of maximizing the target flux.

$$maximize(v_{target}) \tag{4.16}$$

s.t.

$$\sum_{j=1}^{|M|} S_{ij} v_j = 0, \forall i \in N \tag{4.17}$$

$$v_{biomass} \geq (0.01).v_{biomass}^{max} \tag{4.18}$$

$$v_j \leq \begin{cases} SSU_j & \text{if reaction } j \text{ is unmodified} \\ X_j^u & \text{if reaction } j \text{ is up-regulated} \\ X_j^d & \text{if reaction } j \text{ is down-regulated} \end{cases} \quad \forall j \in M \tag{4.19}$$

$$v_j \geq \begin{cases} SSL_j & \text{if reaction } j \text{ is up-regulated or unmodified} \\ 0 & \text{if reaction } j \text{ is down-regulated} \end{cases} \quad \forall j \in M \tag{4.20}$$

In the FBA problem, the flux capacity constraints are drawn from the capacity distributions (X_j^u and X_j^d in equation (4.19)) if the corresponding reaction (enzyme) belongs to the optimized set of interventions. Otherwise, the capacity constraints are set to maximal steady state value (SSU_j) calculated for the unmodified reference state. Similar to MCOpt, MCEval repeatedly solves a series of optimization problems to generate a distribution of optimal target flux values. Unlike MCOpt, MCEval does not seek to identify an intervention set reflecting decisions on enzyme activity modification. Rather, each instance of MCEval simply solves for the optimal flux and the corresponding flux distribution based on capacity constraints specified by the CCOpt, DetOpt, or MCOpt solution that is to be evaluated.

4.2 Results and Discussion

To assess the benefits and limitations of the optimization methods, we compare their performance using test cases involving both a kinetic and a stoichiometric model. The kinetic model describes the metabolism of Chinese hamster ovary (CHO) cells in fed-batch culture [159]. The stoichiometric model describes the metabolism of adipocytes undergoing differentiation and growth [160].

4.2.1 CHO Cell Model

The CHO cell model comprises 24 metabolites and 47 irreversible reactions. The kinetic parameters of the model were previously estimated by fitting the model equations to experimentally obtained metabolite time course data [159]. These parameters are used to estimate the effects of enzyme activity increases and decreases on the corresponding reaction flux capacity distributions. The flux capacity distributions for the adipocyte model are estimated from steady state metabolic flux data obtained in previous studies [161]. Additional details of the model including reaction definitions are provided in Appendix C. The test objective is the synthesis of a recombinant protein product, a therapeutic antibody.

We first estimate the steady state flux values of a nominal reference state and the corresponding capacity distributions. The reference state fluxes (SSU, SSL) are estimated through a linear programming formulation that maximizes/minimizes each reaction flux subject to

$$S.V = 0, 0 \leq v_j \leq v_{j,max}, v_j = v_j^{meas}, j \in MeasuredData \tag{4.21}$$

where $v_{j,max}$ is the maximal velocity of reaction j and $MeasuredData$ is a set of measured exchange flux values for glucose, glutamine, glycine, glutamate and ammonia. The maximal velocities ($v_{j,max}$) are reported in [159] for only 16 of the 47 reactions in the model that explicitly defined with rate expressions. To calculate the $v_{j,max}$ values for the remaining reactions, we solve a series of flux maximization problems subject to the 16 pre-defined maximum velocities. The capacities reflecting up/down-regulations of enzyme activities Cap_j^u/Cap_j^d are obtained by multiplying the maximum velocities with the assumed enzyme activity distributions:

$$Cap_j^{u/d} = v_{j,max}.N_{u/d}(\mu, \sigma^2) \tag{4.22}$$

We compare the intervention sets obtained from CCOpt with $\epsilon = 0.1$ and $\epsilon = 0.25$ (representing two choices of conservative and relaxed confidence levels respectively) and those from DetOpt, and evaluate the intervention sets using Monte Carlo simulations (MCEval). In Figure 4.3, the intervention sets (U for the up-regulation set) identified by each optimization method are shown above their corresponding optimal target flux values. Empty sets represent no identified interventions. For $L = 1$, DetOpt and CCOpt at $\epsilon = 0.1$ and $\epsilon = 0.25$ all select reaction 17, which is the lumped antibody synthesis reaction. For $L = 2$, CCOpt adds reaction 13 to form an intervention set of $\{13, 17\}$. Up-regulating reaction 13 increases the synthesis of cysteine, which could be a limiting reactant. As reported in [162], one of the rate-limiting steps of antibody production in CHO cells is the folding and assembly of polypeptides in the endoplasmic reticulum, which requires cysteine residues. For $L = 3$, CCOpt further adds reaction 1, which lumps together several steps in glycolysis. Up-regulating the flux through glycolysis increase the supply of pyruvate for oxidation in the tricarboxylic acid (TCA) cycle, which in turn could provide additional energy for antibody synthesis [163]. For $L = 4$, CCOpt adds reaction 2, which acts to balance the cytosolic redox by oxidizing NADH and possibly relieves feedback inhibition of glycolysis.

Compared to CCOpt, DetOpt predicts smaller maximal antibody synthesis rates ($\sim 1000nmol/10^6cells/day$) due to the conservative choice of reaction flux capacities. The maximal synthesis rate predicted using CCOpt is more than twice the flux predicted by DetOpt ($\sim 2200nmol/10^6cells/day$). The intervention set identified by DetOpt consists of only a single reaction even when the maximal number allowed interventions is raised, indicating that the deterministic method does fully utilize the degree of freedom available in the problem.

Figure 4.4 shows the distribution of maximum antibody production rates obtained using MCEval for the intervention sets reported in Figure 4.3. In all cases, the maximum flux predicted by DetOpt falls outside the probable (5^{th} to 95^{th} percentile) range calculated by MCEval, whereas the maximal flux predicted by CCOpt falls within this range. When only one intervention is allowed ($L = 1$), the selected reaction is the same for CCOpt and DetOpt. However, the flux predicted by CCOpt is higher, and is also more reliable in a probabilistic sense. When the degree of freedom is higher ($L = 2, 3$ and 4), and different intervention sets are selected, MCEval calculates higher probable ranges for the intervention sets identified by CCOpt compared to DetOpt. For example, for $L = 4$, the probable range for CCOpt lies between 1805 and $2870nmol/10^6cells/day$ whereas both the 5^{th} and 95^{th} percentile values for DetOpt are at 1079.

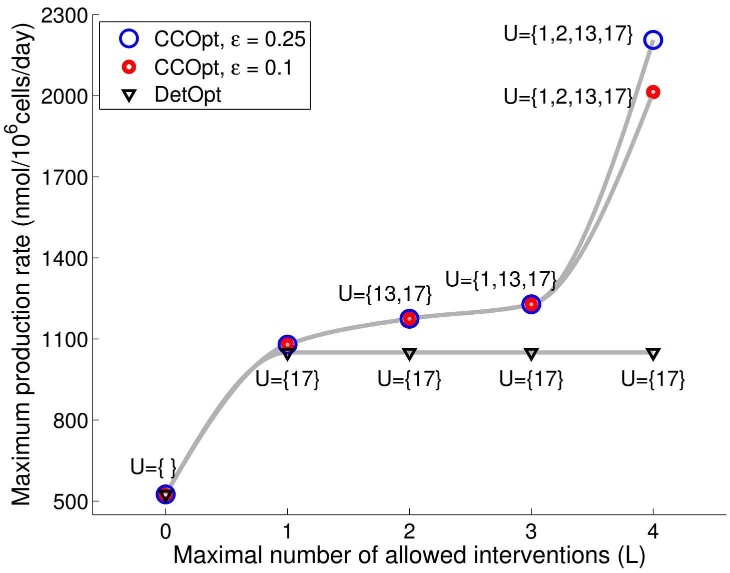


Figure 4.3: Maximum antibody production rate and intervention sets obtained by CCOpt and DetOpt using the CHO cell model. The reactions selected for modification for each intervention set are shown above each data point. The maximum production rates obtained by CCOpt with, $\epsilon = 0.25$, CCOpt with $\epsilon = 0.1$, and DetOpt are shown as blue and red circles and black triangles, respectively. Set U refers to the reactions that need to be up-regulated.

Figure 4.5 shows the distribution of solutions resulting from 10^6 iterations of the Monte Carlo optimization method (MCOpt). For $L = 1$ and $L = 2$, MCOpt generates the same solutions as CCOpt and DetOpt. For $L = 3$, MCOpt identifies four sets of interventions: {1, 13, 17}, {5, 13, 17}, {13, 17}, and {17}. The first set is dominant at a frequency of 99.86%, and matches the CCOpt solution. For $L = 4$, the trend is the same as $L = 3$, with one dominant solution (frequency $> 99\%$) that matches the corresponding CCOpt solution. This set also corresponds to the highest predicted target flux among all intervention sets comprising four reactions.

In the case of $L = 4$, the aggregate effect of uncertainties in flux capacities is to result in a normally distributed target flux. However, this is not the case for $L < 4$, where the dominant target flux values generated by MCOpt distribute narrowly with nearly zero spread. Moreover, the mean target flux values rise only incrementally

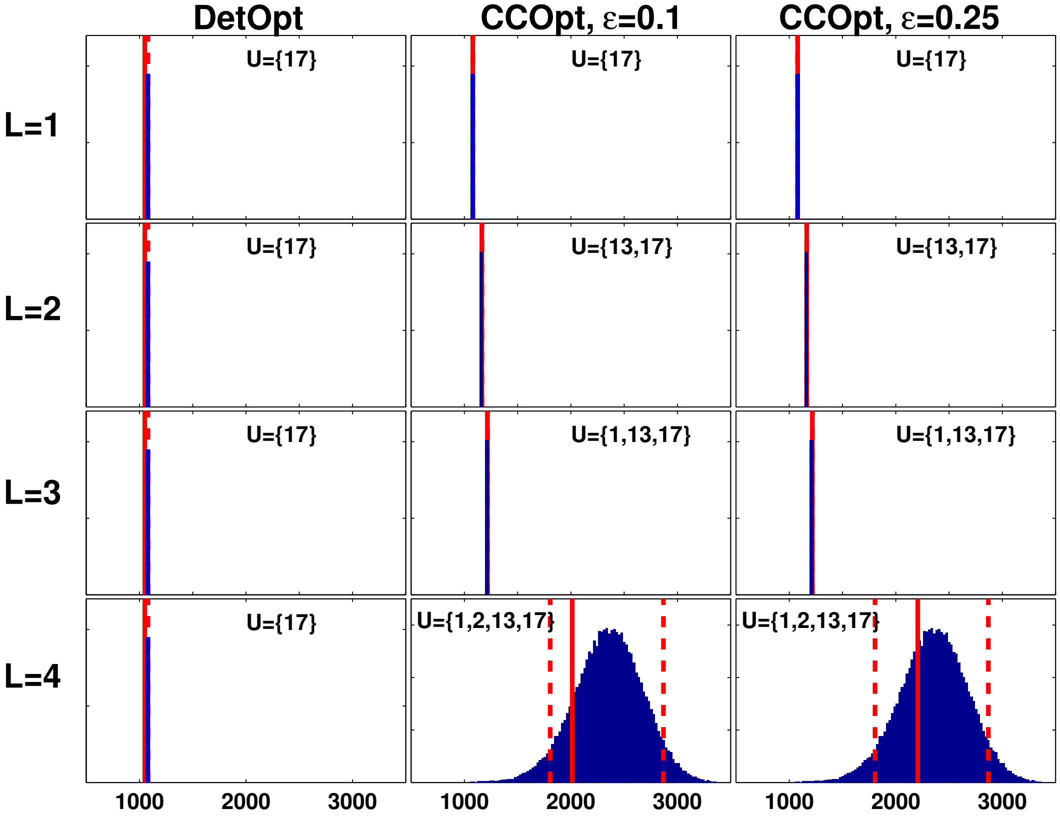


Figure 4.4: Monte Carlo sampling based flux balance analysis (FBA) simulations of the intervention sets identified by CCOpt and DetOpt for antibody production using the CHO cell model. Each panel shows a Monte Carlo distribution of FBA optimized target flux values, with the rows and columns corresponding to different caps on the number of interventions (L) and different optimization methods/settings, respectively. The x-axis represents the maximum antibody production rate in units of $nmol/10^6 cells/day$. The y-axis represents the sampled frequency of an FBA solution. The dashed lines denote the 5^{th} and 95^{th} percentile values, defined as the values below and above which 5% of the data fall, respectively. A single dashed line indicates that these two percentile values are the same. The solid lines indicate the maximum production rates obtained using CCOpt or DetOpt.

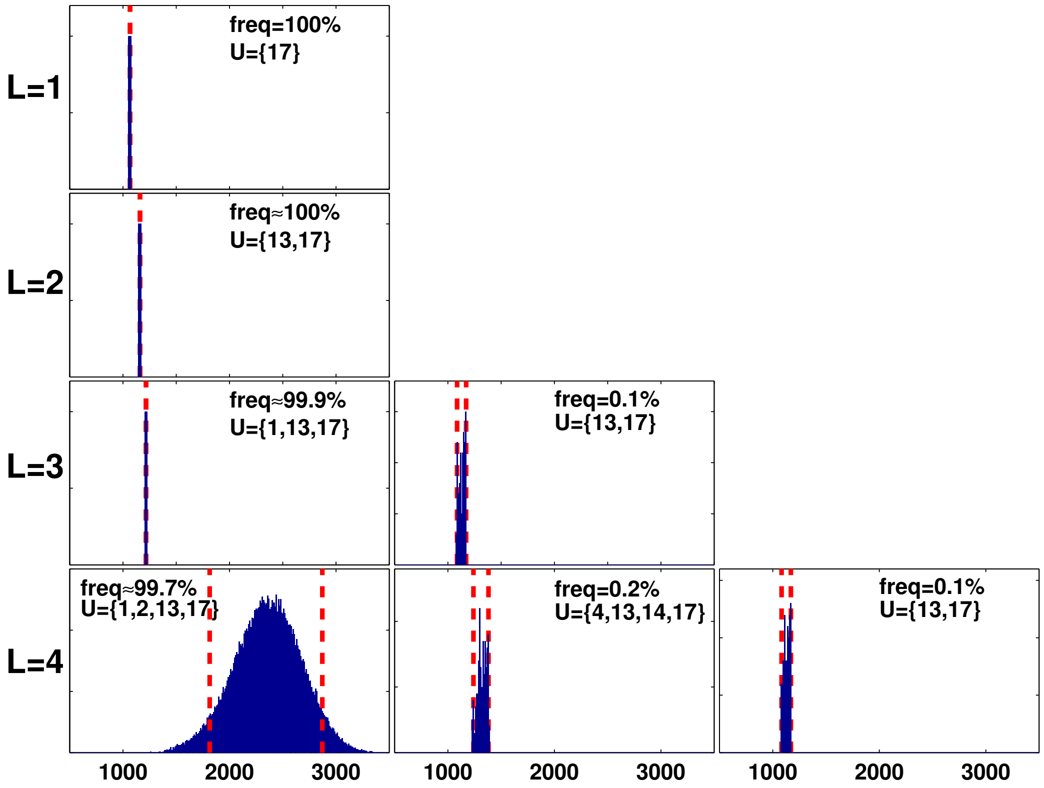


Figure 4.5: Monte Carlo sampling based optimization (MCOpt) of antibody production using the CHO cell model. Each panel shows a MCOpt calculated distribution of target flux values, with the rows and columns corresponding to different caps on the number of interventions (L) and different intervention sets, respectively. For $L = 1$ or 2, MCOpt identified only one intervention set. The x-axis represents the maximum antibody production rate in units of $nmol/10^6 cells/day$. The dashed lines denote the 5^{th} and 95^{th} percentile values. The selection frequency of an intervention set as a fraction of the total pool of MCOpt solutions for a given L is shown as a percentile value at the top of each panel.

from $L = 1$ to 3, suggesting that the probabilistic outcomes accumulate at the lower bound of the probable range due to one or more bottlenecks in the network that are not relieved until all 4 reaction flux capacity modifications are introduced.

Similar to the CCOpt and DetOpt solutions, the MCOpt solutions are evaluated using MCEval (Figure 4.6). The MCEval results for $L = 1$ and 2 are identical to the MCOpt results for $L = 1$ and 2 shown in Figure 4.5, respectively. For $L = 3$, MCOpt generates two sets of interventions, where one dominant set is identified with

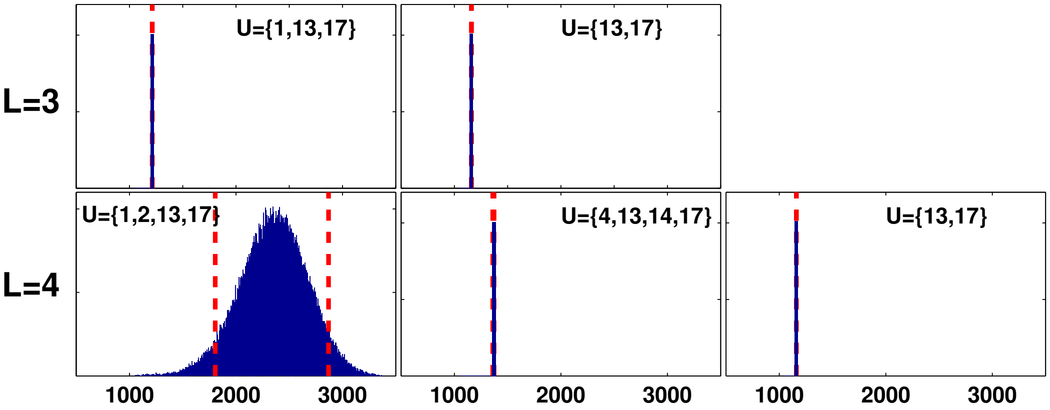


Figure 4.6: Monte Carlo sampling based flux balance analysis (FBA) simulations identified by MCOpt for antibody production using the CHO cell model. Each panel shows a Monte Carlo distribution of FBA optimized target flux values, with the rows and columns corresponding to different caps on the number of interventions (L) and different intervention sets, respectively. Results are shown only for $L =$ 3 and 4. The x-axis represents the maximum antibody production rate in units of $nmol/10^6 cells/day$. The y-axis represents the sampled frequency of an FBA solution. The dashed lines denote the 5th and 95th percentile values. A single dashed line indicates that these two percentile values are the same.

99.9% frequency. Results of MCEval confirm that this solution ({1, 13, 17}) indeed has a higher probable target flux value. A similar trend is observed for $L = 4$. The set with the highest probable target flux values is identical to the CCOpt solution and the dominant (most frequently identified) MCOpt solution. The probable ranges (5th and 95th percentile values) calculated by MCEval for the MCOpt intervention sets {1, 2, 13, 17}, {4, 13, 14, 17} and {13, 17} are (1805, 2870), (1375, 1389) and (1175, 1175) $nmol/10^6 cells/day$, respectively. The MCEval simulations produce a normal distribution of target fluxes only for the solution {1, 2, 13, 17}, presumably because only this set of interventions sufficiently relieves the flux capacity bottlenecks in the network. The results of these evaluations indicate that CCOpt and MCOpt essentially identify the same best intervention sets, where CCOpt arrives at the results without requiring the sampling run-time cost of MCOpt.

4.2.2 Adipocyte Model

In the second case study on the adipocyte model [160], we maximize the production of tripalmitoylglycerol as a representative triacylglycerol (TAG) in adipocyte lipid droplets [164]. This model includes 66 irreversible reactions and 38 metabolites. The details of the model are provided in Appendix C. Unlike the CHO cell case study, we did not use $v_{j,max}$ values to estimate the flux capacities and reference state fluxes. Instead, the reference state flux values are calculated by maximizing each reaction subject to a set of measured untreated control data reported in [161]. To estimate the flux capacity distributions, enzyme control flux (ECF) analysis [77] is used, where the analysis calculates the impact of a change in an enzyme's activity on the steady state flux distribution of the metabolic network. The first step in calculating the distributions is to generate all elementary modes (EMs). For the base adipocyte model, 16,818 EMs were identified using efmtool [165]. In the second step, EM coefficients (EMCs) are calculated through an iterative process. The third step is to estimate the EMCs for a change in enzyme activity. An increase or decrease in enzyme activity is modeled by a normal distribution $N_u(\mu, \sigma^2)$ or $N_d(\mu, \sigma^2)$ as described in Methods (see Section 4.1.4). The fourth step is to calculate the flux distributions using the adjusted EMC vectors. Since the enzyme activity change is described by a distribution, multiple flux distributions are calculated. For each reaction in the network, the reaction flux capacity is the set to the maximal flux value of the reaction from the flux distributions. Repeating the third and fourth steps for all reactions generates a statistical distribution of flux capacities for the network. The maximum TG production rate and intervention sets obtained from CCOpt and DetOpt are shown in Figure 4.7. For both CCOpt and DetOpt, the maximal predicted target flux increases with the number of allowed interventions. As was the case for the CHO cell model, CCOpt predicts a larger maximal flux and generates a more diverse set of solutions compared to DetOpt. In general, DetOpt underutilizes the degrees of freedom available at larger L values. For example, the DetOpt solution comprises only 2 interventions when up to 3 interventions are allowed, whereas the CCOpt solution utilizes all 3 allowed interventions. A second general trend is that the smaller sets of interventions are subsets of the larger sets. An interesting observation is that a single intervention ($L = 1$) yields no change in the predicted maximal flux. This is expected, as reactions 17 and 24 are in series, and both are required for TG synthesis. A change in one without a change in the other merely shifts the limiting capacity to the unchanged reaction.

Reaction 17 is a part of the TCA cycle. Reactions 24 and 26 are palmitate biosynthesis and tripalmitoylglycerol biosynthesis, respectively. All three reactions directly impact synthesis of TG, which is formed from esterification of palmitate with glycerol phosphate, with the latter derived from glycerone phosphate. Previous reports [166], including our own work [161], have shown that the addition of long-chain fatty acids stimulates cellular TG accumulation. At first glance, the intervention targets selected by CCOpt appear trivially intuitive. However, other, equally intuitive alternatives also exist, which were not selected. For example, another intuitive intervention to increase net TG accumulation is to down-regulate lipolysis (reaction 27). This intervention was not selected, because the reference (unmodified) state lower bound for reaction 27 is already zero, and a further reduction would have no impact on TG production. In this regard, the optimization results depend not only on the model, but also on the observed reference state.

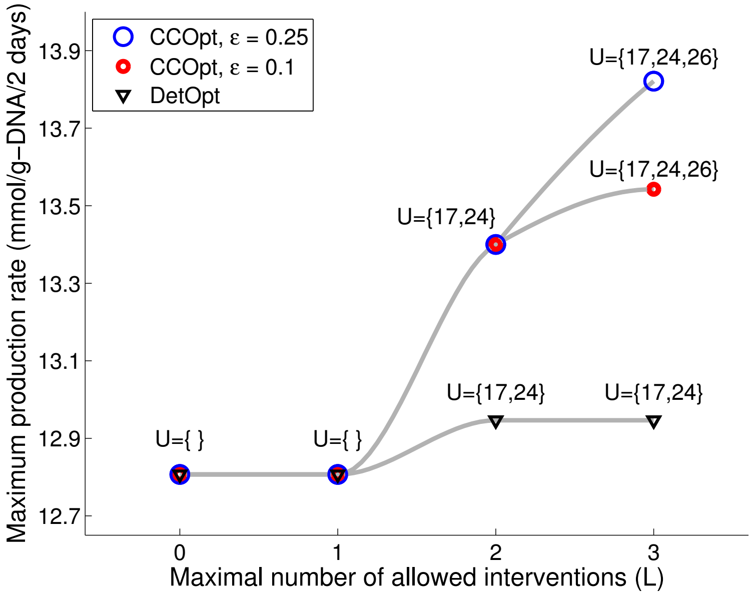


Figure 4.7: Maximum tripalmitoylglycerol production rate and intervention sets obtained by CCOpt and DetOpt using the adipocyte model. The reactions selected for modification for each intervention set are shown above each data point. The maximum production rates obtained by CCOpt with $\epsilon = 0.25$, CCOpt with $\epsilon = 0.1$, and DetOpt are shown as blue and red circles and black triangles, respectively.

As was the case for the CHO cell model, the results of CCOpt more closely match the results of MCEval simulations compared to DetOpt (Figure 4.8). Since neither DetOpt nor CCOpt identified any solutions for $L = 1$, MCEval simulations are not shown. For $L = 2$ and 3, the maximal fluxes predicted by DetOpt (13 $mmol/g - DNA/2days$, shown as solid lines) lie at the lower end of the distributions generated by MCEval. In contrast, the maximal fluxes predicted by CCOpt consistently fall in the probable ($5^{th} - 95^{th}$ percentile) range (shown as dashed lines) of the MCEval distributions. For $L = 3$, the 95^{th} percentile value obtained from MCEval simulations of the CCOpt intervention set is significantly larger than the 95^{th} percentile value obtained from MCEval simulations of the DetOpt intervention set. Additionally, the flux values predicted by CCOpt with and are both in the probable range as calculated by MCEval.

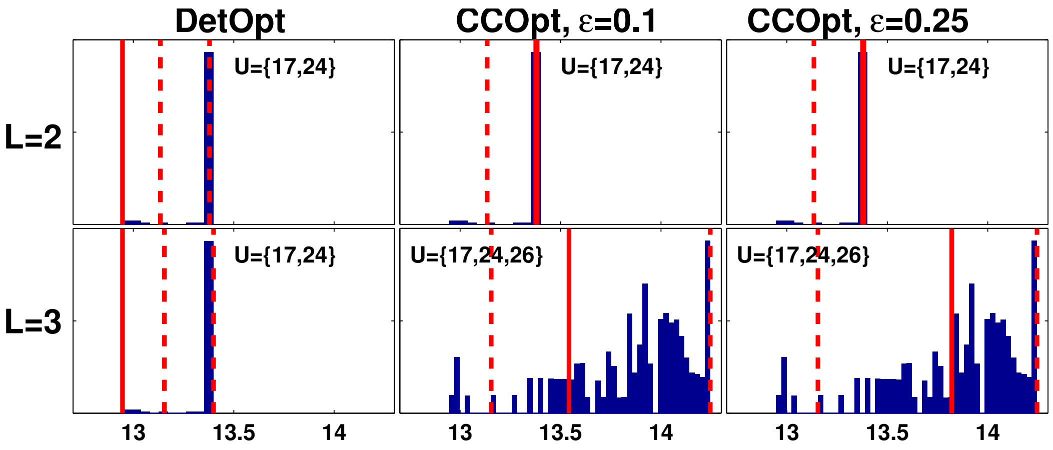


Figure 4.8: Monte Carlo sampling based flux balance analysis (FBA) simulations of the intervention sets identified by CCOpt and DetOpt for tripalmitoylglycerol production using the adipocyte model. Each panel shows a Monte Carlo distribution of FBA optimized target flux values, with the rows and columns corresponding to different caps on the number of interventions (L) and different optimization methods/settings, respectively. Results are shown only for $L = 2$ and 3, as setting $L = 1$ failed to produce any solutions (empty sets in Figure 4.7). The x-axis represents the maximum production rate in units of $mmol/g - DNA/2days$. The y-axis represents the sampled frequency of an FBA solution. The dashed lines denote the 5th and 95th percentile values. The solid lines indicate the maximum production rates obtained using CCOpt or DetOpt.

Applying MCOpt to the adipocyte model generates one solution for $L = 1$ and 2 and two solutions for $L = 3$ (Figure 4.9). The solutions with the highest frequency are identical to the CCOpt solutions. These solutions are {}, {17, 24} and {17, 24, 26} for $L = 1$, 2, and 3, respectively, and occur with 100%, 100% and 89.2% frequency. Of the two MCOpt solutions for $L = 3$, the dominant solution has the higher probable target flux values, which is consistent with the results of MCEval simulations (Figure 4.10).

4.2.3 Computational Complexity and Scalability of Methods

Our optimization problems (CCOpt, MCOpt and DetOpt) are formulated as mixed integer linear programming (MILP). A MILP problem requires a subset of variables to take on integer values, while the other variables can take on non-integer values.

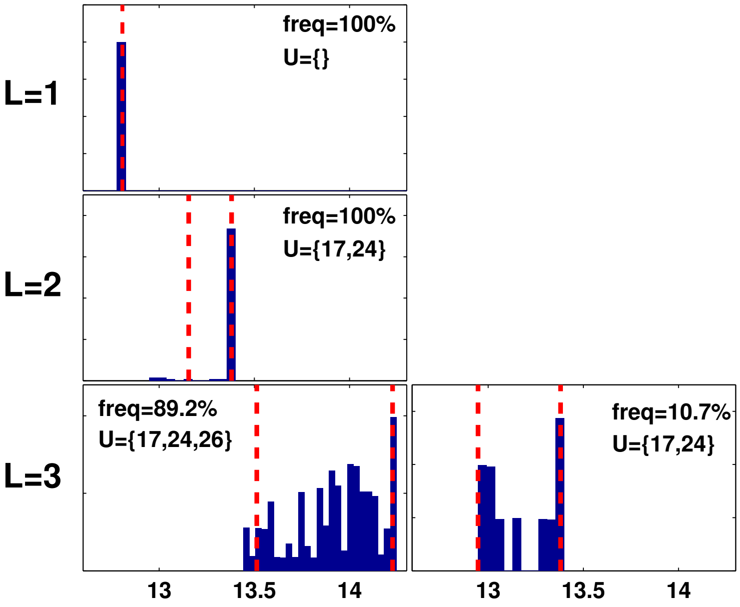


Figure 4.9: Monte Carlo sampling based optimization (MCOpt) of tripalmitoylglycerol (TG) synthesis using the adipocyte model. Each panel shows a MCOpt calculated distribution of target flux values, with the rows and columns corresponding to different caps on the number of interventions (L) and different intervention sets, respectively. For $L = 1$ or 2, MCOpt identified only one intervention set. The x-axis represents the maximum TG synthesis rate. The dashed lines denote the 5^{th} and 95^{th} percentile values. The selection frequency of an intervention set as a fraction of the total pool of MCOpt solutions for a given L is shown as a percentile value at the top of each panel.

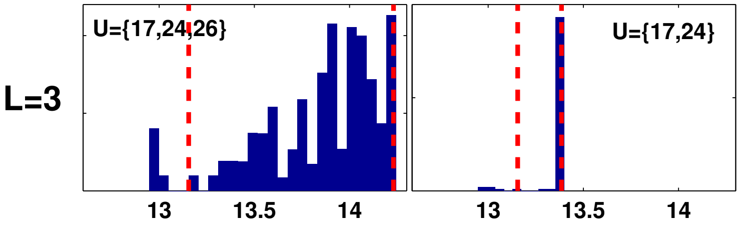


Figure 4.10: Monte Carlo sampling based flux balance analysis (FBA) simulations of the intervention sets identified by MCOpt for tripalmitoylglycerol production using the adipocyte model. Each panel shows a Monte Carlo distribution of FBA optimized target flux values, with the rows and columns corresponding to different caps on the number of interventions (L) and different intervention sets, respectively. The x-axis represents the maximum production rate in units of $mmol/g - DNA/2days$. The y-axis represents the sampled frequency of an FBA solution.

This problem is NP-hard [167], and thus it is unlikely that there exists an efficient (polynomial-time in the size of the model) algorithm to obtain a globally optimal solution. In the present study, we implemented our optimization methods (CCOpt, MCOpt and DetOpt) using the GNU Linear Programming Kit (GLPK) [168] in MATLAB. The runtime of our computational experiments solving the MILP problems was on the order of a few seconds on a Core i5 2.53 GHz CPU for both CHO and adipocyte models.

In addition to the scalability issue inherent to MILP problems, another computational challenge lies in estimating the flux capacity distributions. For the stoichiometric model of this study, we used enzyme control flux analysis (ECF) [77] to obtain these distributions. The ECF method in turn relies on elementary mode (EM) analysis, which can be applied to metabolic models comprising $<\sim 100$ reactions, but remains intractable for genome-scale models. An alternative strategy is to model the fold-change in flux capacity, *i.e.* enzyme activity, resulting from a gene expression modification using a probability distribution, e.g. a normal distribution. This strategy requires knowledge of maximal enzyme velocities (v_{max}). If these parameters are not known, they may be estimated from FBA, which has been demonstrated on genome-scale models.

These types of limitations, while not trivial, are comparable to other computational strain design methods. For example, bi-level optimization, used in OptKnock [47], is also NP-hard [169], and thus can be intractable for large-scale problems. As an NP-hard problem, the runtime grows exponentially with the number of allowed reaction modifications [48]. Methods that rely on EM analysis [74, 75, 76, 170] face a similar limitation as our capacity estimation problem, as the analysis is generally only practical for small to mid-scale models. Methods based on local search [48] or metaheuristics [73, 171] are computationally less prohibitive than MILP, and likely offer the best alternative for large-scale problems. On the other hand, these methods cannot guarantee global solution optimality, and may arrive at solutions that are far from exact.

4.3 Conclusion

This study investigates three distinct ways of capturing uncertainty about parameter values when formulating an optimization problem with the objective of identifying targets for enzyme activity adjustments that maximize the production of a desired molecule. The three approaches are chance-constrained programming (CCOpt), Monte Carlo sampling-based solution of the uncertain problem (MCOpt), and deterministic optimization based on worst-case assumptions (DetOpt). Evaluation of the approaches for two test cases (CHO cell and adipocyte models) using Monte Carlo simulations (MCEval) shows that a more sophisticated probabilistic approach such as CCOpt has several advantages compared to a conservative conventional approach like DetOpt. Chance-constrained programming explores a larger portion of the solution space and is able to find a more diverse set of options. Additionally, CCOpt consistently outperforms DetOpt in terms of predicting the more likely maximum of the objective function value. Comparisons of the intervention sets from CCOpt and DetOpt using MCEval shows that the maximal fluxes predicted by CCOpt was always in the probable ($5^{th} - 95^{th}$ percentile) range calculated by MCEval, whereas the maximal fluxes predicted by DetOpt typically lies outside of this range. When compared to the sampling-based optimization approach (MCOpt), CCOpt consistently finds the solution most frequently selected by MCOpt, but at a fraction of the computational cost (seconds vs. days).

The CCOpt formulation can be readily extended to capture other types of uncertainties, such as biological variability in measured data and cell transfection efficiency, making CCOpt an effective technique for probabilistic strain optimization.

Chapter 5

Conclusions and Future Directions

Current metabolic engineering practices rely heavily on domain knowledge and laboratory experimentation. Computational methods in the form of design automation tools promise to play an important role in guiding discovery and speeding experimentation. Such tools have proven effective across many engineering disciplines including electronics and automotive design. In the future, it may be possible to redesign an organism *in silico* without resorting to repeated experimentations. This thesis presented three algorithmic approaches that can play a critical role in identifying optimal synthesis routes and in tuning microbial hosts.

A central theme in this thesis is the use of probabilistic approaches during optimization and discovery. In realizing synthesis pathways, *ProbPath* probabilistically selects reactions among available reactions in the KEGG database. Exploring various probabilistic selection approaches allowed the important realization that pathway selection algorithms should not bias the solution space towards low- or high-connected metabolites. PROXIMAL provided transformation predictions and ranked them based on available biological data. Enzyme abundance and activity provided weightings for possible outcomes. Our robust optimization method, CCOpt, to identify gene interventions was probabilistic in guaranteeing that the constraints are met with some desired confidence. Our results showed that chance-constrained programming expanded the solution space, identifying a more diverse set of design options. Additionally, we showed that this approach consistently outperformed a deterministic approach in terms of predicting the more likely maximum of the objective function value.

5.1 Research Summary

We presented a pathway construction algorithm, *ProbPath*, for identifying viable synthesis pathways compatible with balanced cell growth. Rather than exhaustive exploration, we utilized probabilistic selection of reactions to construct the pathways. Three different schemes were investigated for the selection of reactions: high metabolite connectivity, low connectivity and uniformly random. For all case studies, which involved a diverse set of target metabolites, the uniformly random selection scheme predicted the highest average maximum yield. When compared to an exhaustive search enumerating all possible reaction routes, our probabilistic algorithm returned nearly identical distributions of yields, while requiring far less computing time (minutes vs. years). The pathways identified by our algorithm have previously been confirmed in the literature as viable, high-yield synthesis routes. Prospectively, our algorithm could facilitate the design of novel, non-native synthesis routes by efficiently exploring the diversity of biochemical transformations in nature.

Our algorithm, *ProbPath*, is useful in identifying biosynthesis pathways when the target metabolite and its corresponding reactions are catalogued in databases. If this information is unavailable, however, other approaches are needed. We presented a new method for predicting possible transformation routes for uncatalogued target metabolites. The method was specifically developed in the context of discovering biotransformations routes for xenobiotics under human liver enzymes, but can be generalized to discover synthesis pathways that utilize metabolites and reactions not catalogued in databases. The method can be further generalized to other problems such as constructing selection pathways for the directed evolution of enzymes. Using reaction data from DrugBank and KEGG, our tool, PROXIMAL, builds look-up tables that catalogue the sites and types of substructural modifications possible for a given group of enzymes. For a compound of interest, PROXIMAL searches for substructures that match the sites catalogued in the look-up tables, applies the corresponding modifications to generate a panel of possible transformation products, and ranks the products based on the activity and abundance of the enzymes involved. Unlike other prediction methods that apply generic rules, PROXIMAL generates transformations that are specific for the compound of interest by analyzing the compound's substructures. We assessed our tool by predicting the possible transformations of xenobiotic chemicals through phase I and phase II human enzymes. We evaluated the accuracy of PROXIMAL's predictions through case studies on two environmental chemicals with suspected endocrine disrupting activity, bisphenol A (BPA) and 4-chlorobiphenyl (PCB3). Comparisons with published reports confirm 5 out of 7 and 17 out of 26 of the predicted derivatives for BPA and

PCB3, respectively.

In addition to identifying synthesis pathways, strain optimization methods allow tuning of design variable to meet engineering objectives. We addressed in this thesis uncertainties in achieving targeted enzyme values when up or down regulating reactions. We modeled the uncertainties as probability distributions in the flux carrying capacities of reactions. Based on this model, we developed an optimization method that identifies reactions for flux capacity modifications to predict outcomes with high statistical likelihood. We compared three optimization methods that select an intervention set comprising up- or down- regulation of reaction flux capacity: CCOpt (Chance constrained optimization), DetOpt (Deterministic optimization), and MCOpt (Monte Carlo-based optimization). We evaluated the methods using a Monte Carlo simulation-based method, MCEval (Monte Carlo Evaluations). We presented two case studies analyzing a CHO cell and an adipocyte model. The flux capacity distributions required for our methods were estimated from maximal reaction velocities or elementary mode analysis. The intervention set selected by CCOpt consistently outperforms the intervention set selected by DetOpt in terms of tolerance to flux capacity variations. MCEval showed that the optimal flux predicted based on the CCOpt intervention set is more likely to be obtained, in a probabilistic sense, than the flux predicted by DetOpt. The intervention sets identified by CCOpt and MCOpt were similar; however, the exhaustive sampling required by MCOpt incurred significantly greater computational cost.

5.2 Future Research Directions

While this thesis advanced computational methods for pathway synthesis and strain optimization, there are still many improvements required to bridge the gap between computational predictions and experimental outcomes. Computational methods must provide support for a holistic design flow based on accurate and more detailed modeling. In this section, we describe some future directions.

ProbPath currently ranks pathways solely on metabolic objectives (e.g. yield, length of synthesis pathway). *ProbPath* does not take into consideration the potential genetic compatibility of the genes along the synthesis pathway with the host. Integration of a heterologous gene with the host can be investigated in two ways: how likely is it for the non-native genes to be expressed in the host and whether it disrupts the host function. Feature research can consider the gene expression and solubility of proteins as a ranking metric. Each gene in the pathway can be classified with high/low expression/solubility with features such as the DNA/protein sequences, protein secondary structures, molecular weights, hydrophilicity, and hydrophobicity. To detect the disruption potential of the non-native pathways, the toxicity of intermediate metabolites can be investigated.

In this thesis, PROXIMAL is presented as a special case for constructing biotransformation routes in the context of a specific set of human liver enzymes for xenobiotic transformations. Based on enzyme promiscuity, this approach can be generalized to take as input an arbitrary set of enzymes and generate all biotransformations for any target molecule. PROXIMAL can also be used to recursively generate all possible pathways starting with the target metabolite and ending with a metabolite in the host. *ProbPath* can be used to sample this space to identify viable synthesis pathways.

PROXIMAL ranks the predicted xenobiotic transformations based on the activity and abundance of involved enzymes. Although biologically relevant, this scheme may be limiting in practice as this information might be unavailable for some enzymes. The current ranking scheme can be augmented with a confidence level, where each transformation in the look-up table is associated with a confidence level calculated reflecting confidence in the biological data. The PROXIMAL ranking method can be improved by also taking into account the energetic favorability of transformation reactions.

CCOpt identifies a set of interventions to over-produce metabolites in a host while considering uncertainties in the parameters. However, CCOpt only considers up or down regulation as possible genetic modifications, but does not determine an optimal fold change in the gene expression. Future research can apply simulated annealing and mixed-integer linear programming for identifying optimal fold changes, or fold ranges, while considering practical limitations on realizing gene modifications. The CCOpt formulation can also be extended to capture other types of uncertainties such as biological variability in measured data.

Our current work does not take into account the impact of gene capacity modifications on steady-state boundaries of fluxes in the network, which may result in conservative estimates of flux calculations. Dynamically recalculating the steady-state boundaries based on the selected intervention set can fix this problem. Candidate methods that allow for dynamic constraint considerations include iterative heuristics such as simulated annealing.

Appendix A

Supplementary Material for Chapter 2

In this appendix, we present yield distribution plots for all test cases for different weighting schemes (uniform, low and high connectivity, and exhaustive). We then show the yield results of different weighting schemes (uniform, low and high connectivity). Finally, we provide a table comparing the result of uniform probabilistic and exhaustive search with the same length limit.

A.1 Yield Distributions

The following figures show supplementary yield distributions for different test case metabolites comparing the results of uniform, high-connectivity, low-connectivity probabilistic and exhaustive methods. As evident from the figures, the yield distributions obtained from the uniform probabilistic method have similar patterns to those obtained using the exhaustive method, which suggests that the uniform prob-

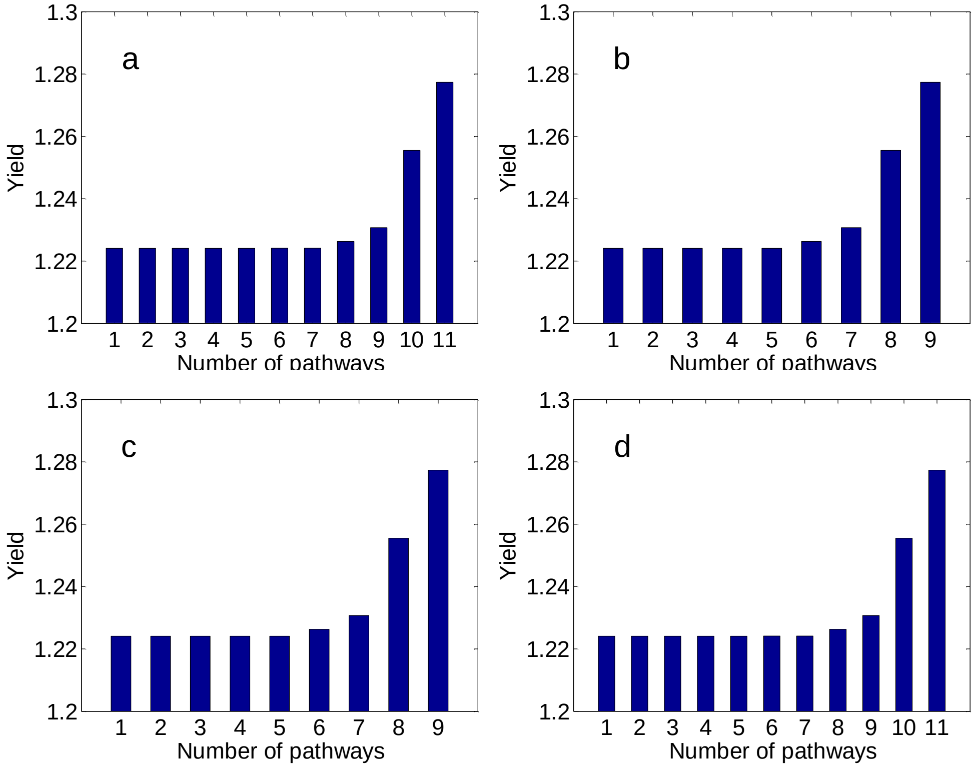


Figure A.1: Yield distribution histograms for isopentenyl diphosphate obtained using uniform (a), exhaustive (b), high-connectivity (c) and low-connectivity probabilistic (d) methods. The x-axis shows the number of pathways obtained by a single run of the exhaustive method with maximum 10 reactions in the pathway (b), 1000 iterations of the probabilistic methods (a, c, d).

abilistic method is capable of producing different potential high-yield pathways that other weighting schemes ignore. The similarity between generated patterns using different weighting schemes suggests that there is no correlation between connectivity and yield distribution.

A.2 Uniform vs. Connectivity Weightings

In the following figures we compare the search performance between the uniform, high-connectivity and low-connectivity weightings for different test case metabolites. We ran each method 50 times for different number of iterations and saved the

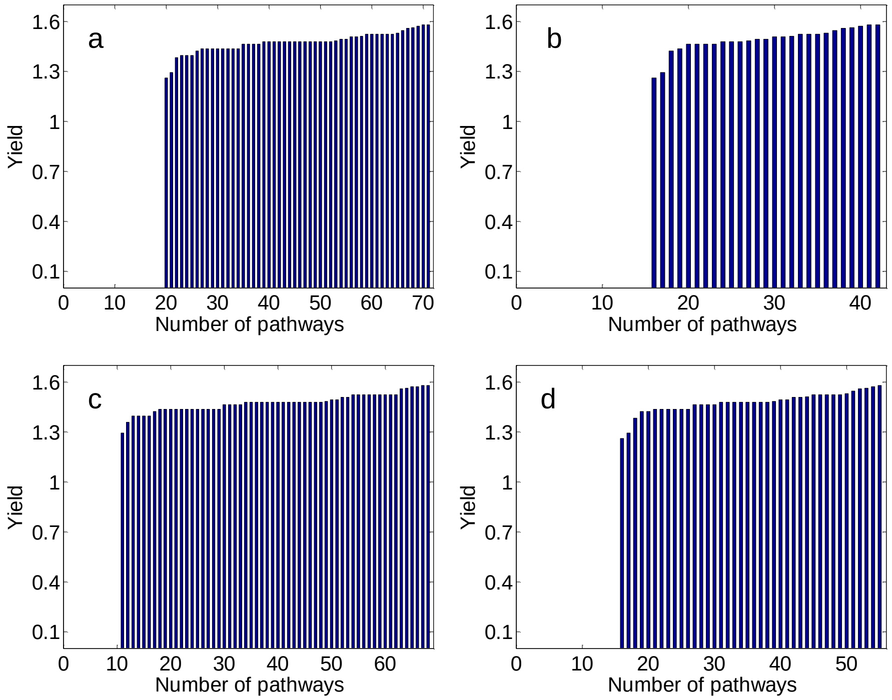


Figure A.2: Yield distribution histograms for *myo*-Inositol obtained using uniform (a), exhaustive (b), high-connectivity (c) and low-connectivity probabilistic (d) methods.

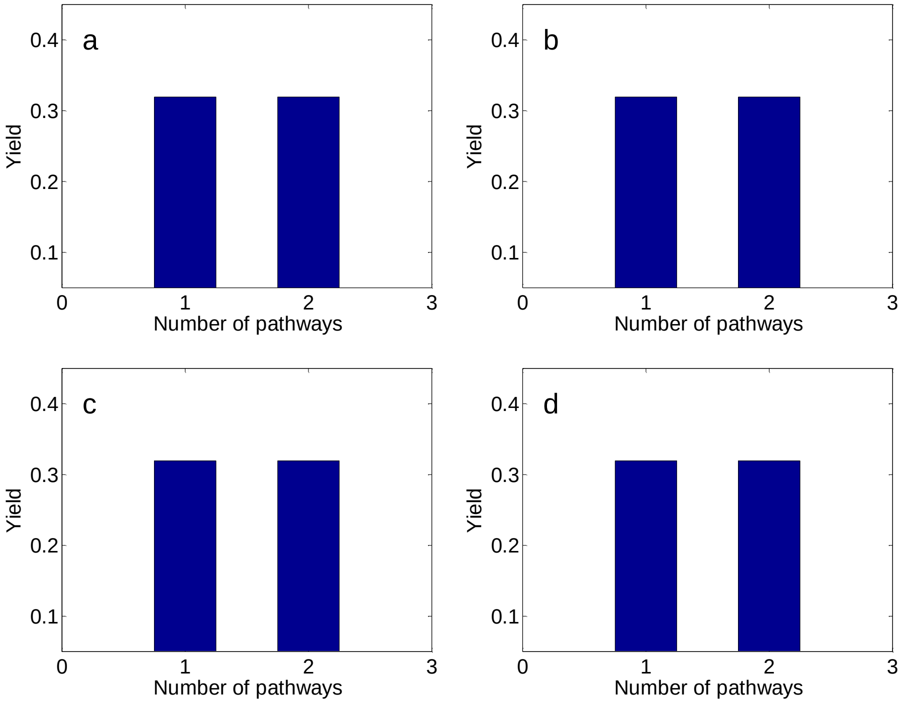


Figure A.3: Yield distribution histograms for taxadiene obtained using uniform (a), exhaustive (b), high-connectivity (c) and low-connectivity probabilistic (d) methods.

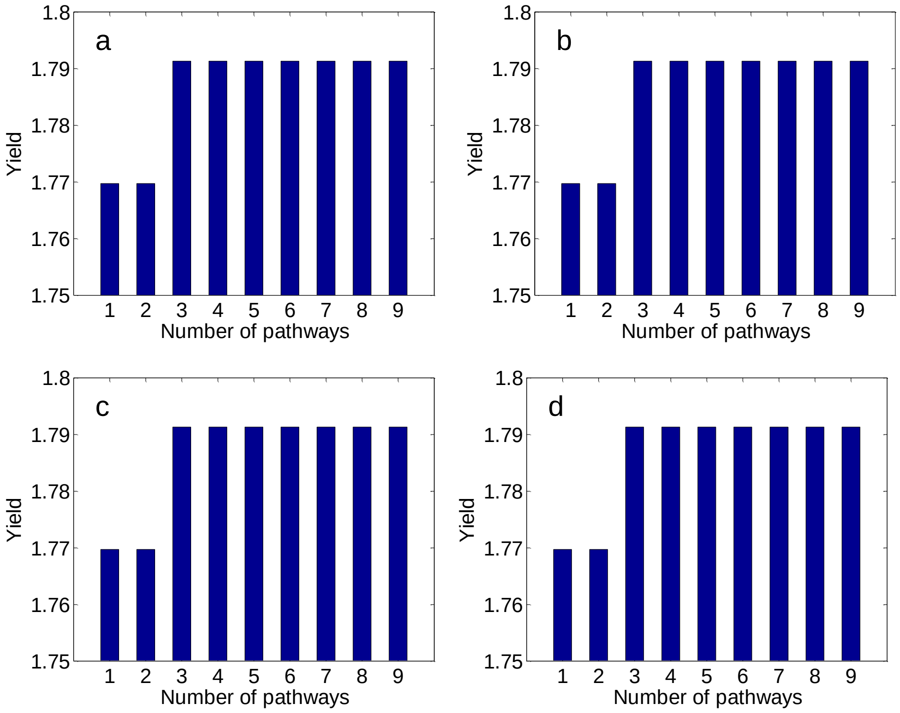


Figure A.4: Yield distribution histograms for (R,R),2-3, butanediol obtained using uniform (a), exhaustive (b), high-connectivity (c) and low-connectivity probabilistic (d) methods.

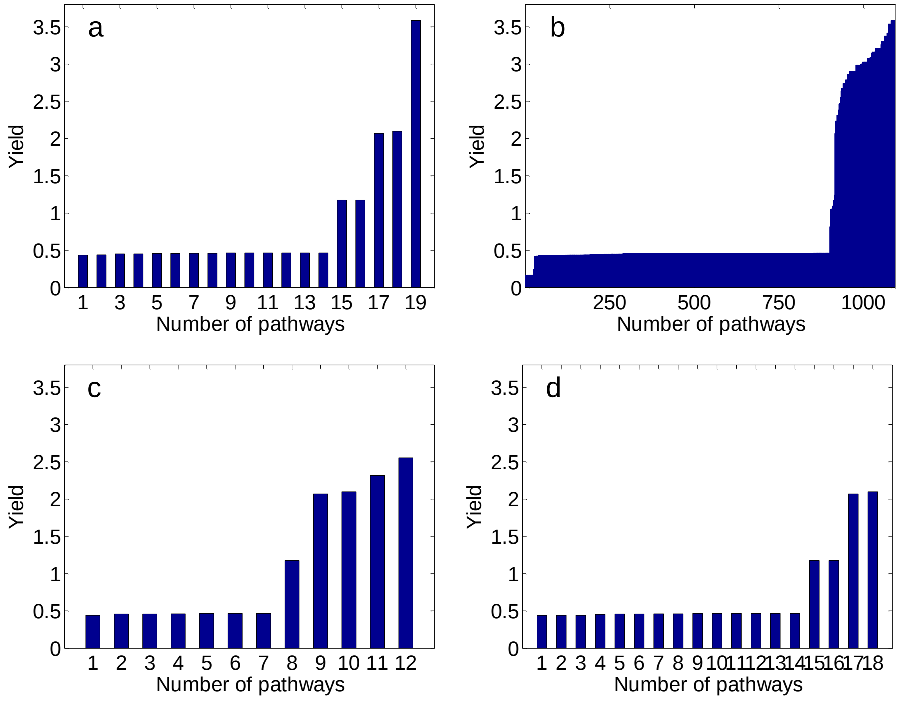


Figure A.5: Distribution histograms for fatty acid ethyl esters (FAEEs) obtained using uniform (a), exhaustive (b), high-connectivity (c) and low-connectivity probabilistic (d) methods.

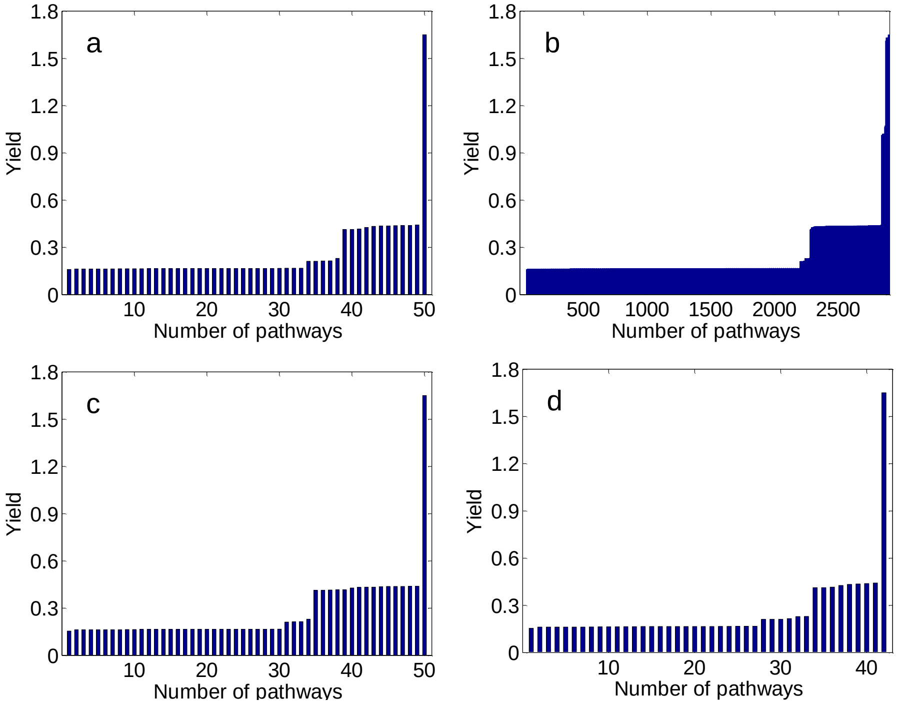


Figure A.6: Distribution histograms for triacylglycerol obtained using uniform (a), exhaustive (b), high-connectivity (c) and low-connectivity probabilistic (d) methods.

maximum obtained yield for each run. We plotted the average and maximum of the stored yields.

A.3 Uniform *ProbPath* vs. Exhaustive Search

The following table summarizes the maximum yields and the number of pathways obtained using 1,000 iterations of the probabilistic search and one single run of the exhaustive search with the same length limit.

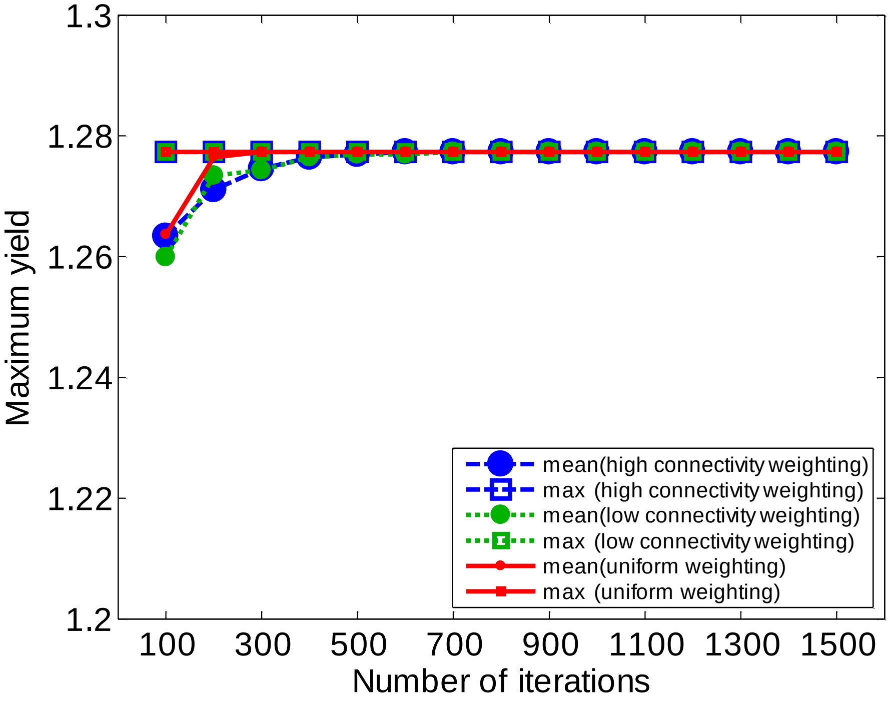


Figure A.7: Comparing uniform, high-connectivity and low-connectivity weighting schemes for IPP using different number of iterations ranging from 100 to 1500. The x-axis shows number of iterations and y-axis shows the maximum obtained yield. The blue, green and red curves indicate the results of high-connectivity, low-connectivity and uniform weightings respectively. The maximums are shown with squares and the means are indicated by circles.

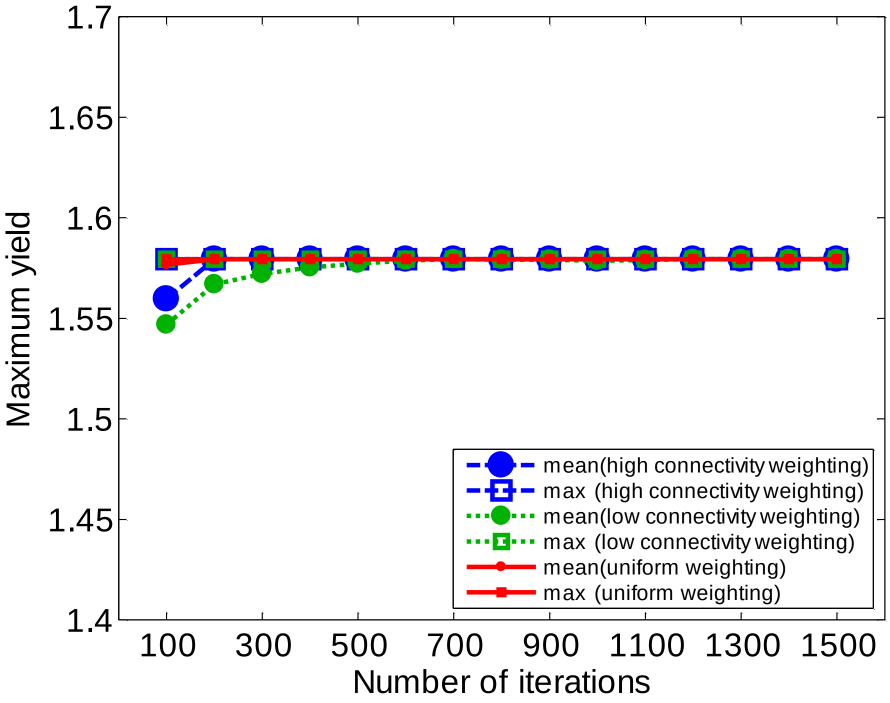


Figure A.8: Comparing uniform, high-connectivity and low-connectivity weighting schemes for *myo*-Inositol.

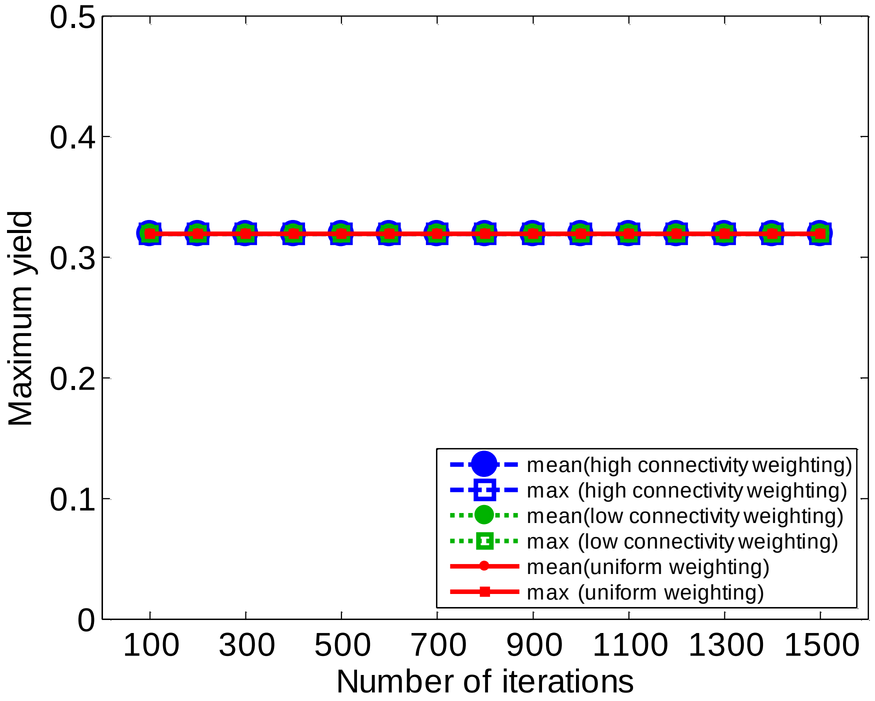


Figure A.9: Comparing uniform, high-connectivity and low-connectivity weighting schemes for taxadiene.

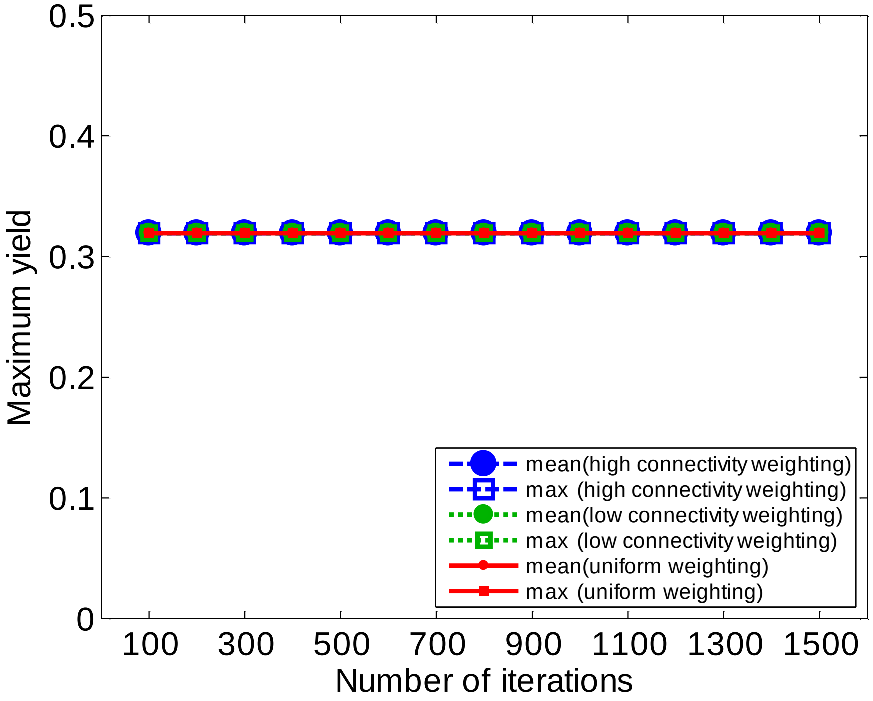


Figure A.10: Comparing uniform, high-connectivity and low-connectivity weighting schemes for 1,3-propanediol.

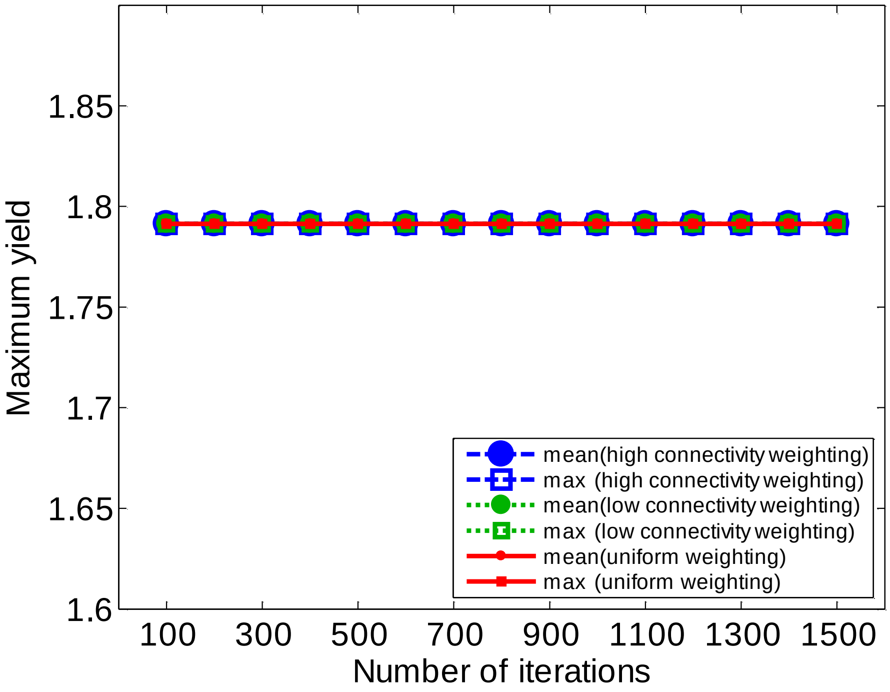


Figure A.11: Comparing uniform, high-connectivity and low-connectivity weighting schemes for (R,R),2-3, butanediol.

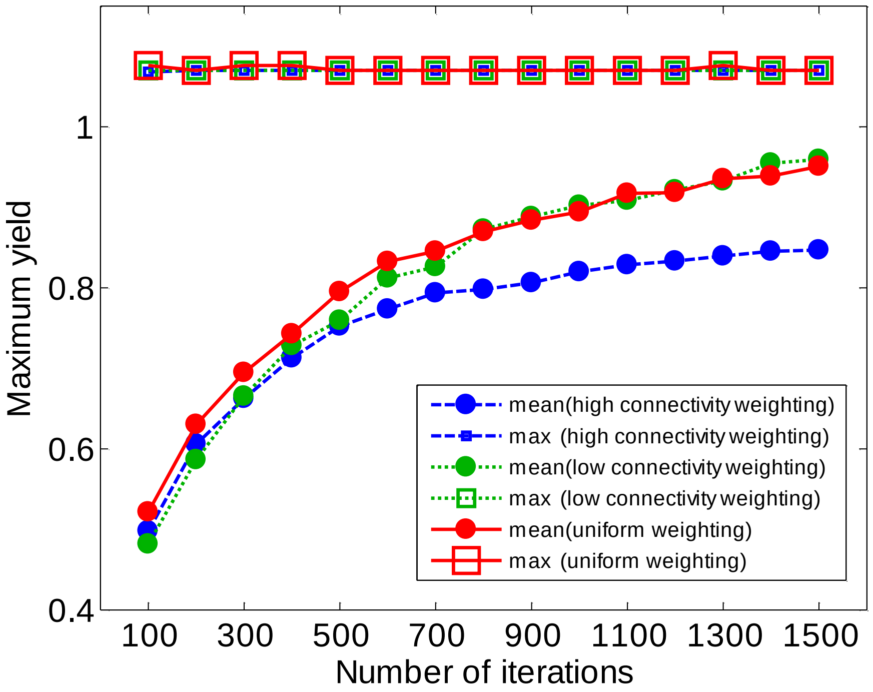


Figure A.12: Comparing uniform, high-connectivity and low-connectivity weighting schemes for fatty acid methyl esters (FAMEs).

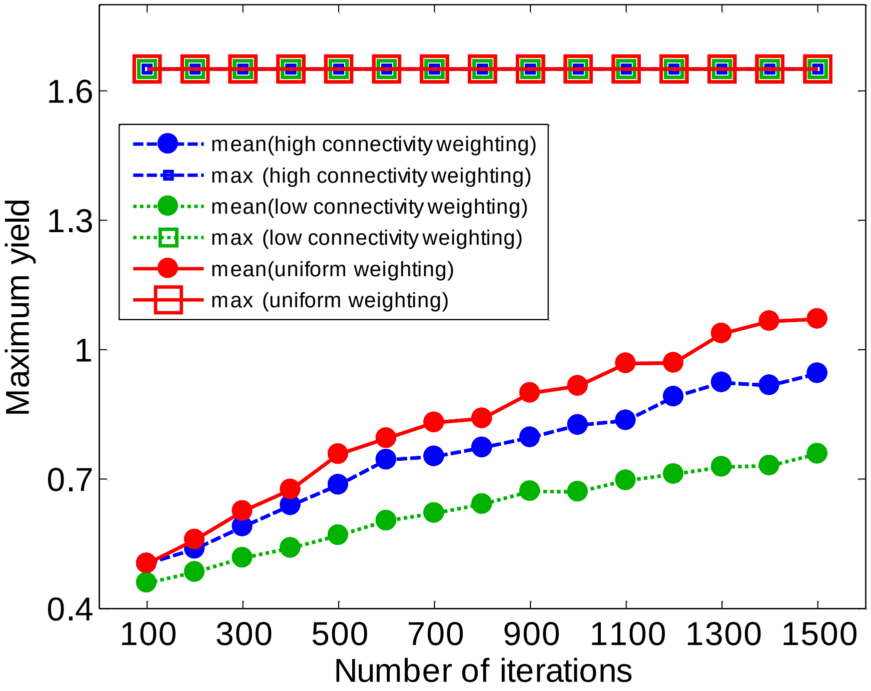


Figure A.13: Comparing uniform, high-connectivity and low-connectivity weighting schemes for triacylglycerol.

Metabolite name	Method	# pathways	Maximum Flux $mmol/gDW/h$
Isopentenyl diphosphate	Uniform *ProbPath*	9	1.28
	Exhaustive search	9	1.28
Myo-Inositol	Uniform *ProbPath*	31	1.58
	Exhaustive search	42	1.58
Taxadiene	Uniform *ProbPath*	2	0.32
	Exhaustive search	2	0.32
1,3-Propanediol	Uniform *ProbPath*	1	2.19
	Exhaustive search	1	2.19
2,3-Butanediol	Uniform *ProbPath*	9	1.79
	Exhaustive search	9	1.79
Fatty acid ethyl esters	Uniform *ProbPath*	16	3.58
	Exhaustive search	1092	3.58
Fatty acid methyl esters	Uniform *ProbPath*	38	1.07
	Exhaustive search	1353	1.08
Triacylglycerol	Uniform *ProbPath*	56	1.65
	Exhaustive search	2900	1.65

Table A.1: Summary of search results obtained using uniform probabilistic and exhaustive methods with the length limit of 10.

Appendix B

Supplementary Material for Chapter 3

CYP	substrate						
1A2	Phenacetin		Phenacetin	Phenacetin	Phenacetin, Resazurin	7-ethoxycoumarin	Caffeine
2A6			Coumarin	Coumarin	Coumarin	Coumarin	
2B6			Bupropion	Bupropion	Bupropion, Mephenytoin	7-ethoxycoumarin	
2C8			Amdiaquine	Amdiaquine	Paclitaxel	Paclitaxel	
2C9	Tolbutamide		Diclofenac	Diclofenac	Diclofenac	S-warfarin	Tolbutamide
2C19	S-mephenytoin		S-mephenytoin	S-mephenytoin	S-mephenytoin	S-mephenytoin	S-mephenytoin
2D6	Dextromethorphan		Dextromethorphan	Dextromethorphan	Dextromethorphan	Bufuralol	Bufuralol
2E1	Chlorozoxazone		Chlorozoxazone	Chlorozoxazone	Chlorozoxazone	7-ethoxycoumarin	Paranitrophenol
3A4	Midazolan	Testosterone	Testosterone, Midazolan	Testosterone, Midazolan	Testosterone, Midazolan	Testosterone	Testosterone
Referenced Study	[132]	[134]	[135](200 donor pool)	[135](50 donor pool)	[135](16 donor pool)	[136]	[133]

Table B.1: Substrates included in estimating cytochrome P450 activity.

Appendix C

Supplementary Material for Chapter 4

C.1 CHO Cell Model

RXN	Reactions
1	$G6P \rightarrow 2PYR + 3ATP + 2NADHc$
2	$PYR + NADHc \rightarrow LAC$
3	$PYR + GLU \rightarrow ALA + AKG$
4	$PYR + OXA \rightarrow AKG + 2CO_2 + 2NADH$
5	$AKG \rightarrow MAL + CO_2 + NADH + FADH2 + ATP$
6	$MAL \rightarrow OXA + NADH$
7	$MAL \rightarrow PYR + CO_2 + NADPH$
8	$GLN \rightarrow GLU + NH_3 + ATP$
9	$AKG + NH_3 + NADH \rightarrow GLU$
10	$ASN \rightarrow ASP + NH_3$

11 $ASP + AKG \rightarrow OXA + GLU$

12 $SER + CO_2 + NH_3 + NADHc \rightarrow 2GLY$

13 $C\_C + NADHc \rightarrow 2CYS$

14 $NADH + 0.5O_2 \rightarrow 2.5ATP$

15 $FADH2 + 0.5O_2 \rightarrow 1.5ATP$

16 $0.084ALA + 0.041ASN + 0.080ASP + 3.755ATP + 0.026CYS + 0.452G6P + 0.087GLN + 0.056GLY + 0.427OXA + 0.096SER \rightarrow BIOM + 0.004FADH2 + 0.008GLU + 0.445MAL + 0.639NADH + 0.209PYR$

17 $0.061ALA + 0.034ASN + 0.039ASP + 4.000ATP + 0.024CYS + 0.048GLU + 0.045GLN + 0.072GLY + 0.126SER \rightarrow ANTI$

18 $BIOM \rightarrow \mathbf{BIOM}$

19 $ANTI \rightarrow \mathbf{ANTI}$

20 $\mathbf{GLC} + ATP \rightarrow G6P$

21 $LAC \rightarrow \mathbf{LAC}$

22 $ALA \rightarrow \mathbf{ALA}$

23 $\mathbf{ASN} \rightarrow ASN$

24 $ASP \rightarrow \mathbf{ASP}$

25 $\mathbf{C\_C} + GLU \rightarrow C\_C + \mathbf{GLU}$

26 $\mathbf{GLN} \rightarrow GLN$

27 $GLY \rightarrow \mathbf{GLY}$

28 $\mathbf{SER} \rightarrow SER$

29 $NH_3 \rightarrow \mathbf{NH_3}$

30 $\mathbf{O_2} \rightarrow O_2$

31	$CO_2 \rightarrow \mathbf{CO_2}$
32	$2CYS + O_2 \rightarrow C\_C$
33	$GLU \rightarrow \mathbf{GLU}$
34	$NADHc \rightarrow 0.5NADH + 0.5FADH2$
35	$LAC \rightarrow PYR + NADHc$
36	$ALA + AKG \rightarrow PYR + GLU$
37	$GLU + NH_3 + ATP \rightarrow GLN$
38	$GLU \rightarrow AKG + NH_3 + NADH$
39	$ASP + NH_3 \rightarrow ASN$
40	$2GLY \rightarrow SER + CO_2 + NH_3 + NADHc$
41	$\mathbf{LAC} \rightarrow LAC$
42	$\mathbf{ALA} \rightarrow ALA$
43	$\mathbf{ASP} \rightarrow ASP$
44	$GLN \rightarrow \mathbf{GLN}$
45	$\mathbf{GLY} \rightarrow GLY$
46	$\mathbf{NH_3} \rightarrow NH_3$
47	$\mathbf{GLU} \rightarrow GLU$

*Extracellular metabolites are shown in bold.

C.2 Adipocyte Model

RXN	Reactions
1	$\mathbf{Glucose} + ATP \rightarrow Glucose6\_P + ADP$
2	$Glucose6\_P \rightarrow Fructose6\_P$

3 $Fructose6\_P + ATP \rightarrow Glyceraldehyde3\_P + Glycerone\_P + ADP$

4 $Glycerone\_P \rightarrow Glyceraldehyde3\_P$

5 $Glyceraldehyde3\_P + NAD^{+} + ADP + Pi \rightarrow P\_Enolpyruvate + ATP + H_2O$

6 $P\_Enolpyruvate + ADP + H+ \rightarrow Pyruvate + ATP$

7 $Pyruvate + NADH + H^{+} \rightarrow \textbf{Lactate} + NAD^{+}$

8 $Glucose6\_P + 2NADP^{+} + H_2O \rightarrow Ribulose5\_P + CO_2 + 2NADPH + 2H^{+}$

9 $3Ribulose5\_P \rightarrow 2Fructose6\_P + Glyceraldehyde3\_P$

10 $\boldsymbol{Pyruvate} + \boldsymbol{Oxaloacetate} + NAD^{+} + H_2O \rightarrow \boldsymbol{Citrate} + CO_2 + NADH + H^{+}$

11 $\boldsymbol{Pyruvate} + HCO_3^{-} + ATP \rightarrow \boldsymbol{Oxaloacetate} + ADP + Pi$

12 $\boldsymbol{Citrate} + NAD^{+} \rightarrow \boldsymbol{2\_Oxoglutarate} + CO_2 + NADH + H^{+}$

13 $\boldsymbol{2\_Oxoglutarate} + NAD + +CoA \rightarrow \boldsymbol{Succinyl\_CoA} + CO_2 + NADH$

14 $\boldsymbol{Succinyl\_CoA} + FAD + Pi + ADP \rightarrow \boldsymbol{Fumarate} + FADH2 + ATP + CoA$

15 $\boldsymbol{Fumarate} + H_2O \rightarrow \boldsymbol{Malate}$

16 $\boldsymbol{Malate} + NAD^{+} \rightarrow \boldsymbol{Oxaloacetate} + NADH + H^{+}$

17 $Citrate + CoA + ATP \rightarrow Acetyl\_CoA + Oxaloacetate + ADP + Pi$

18 $Oxaloacetate + NADH + H^{+} \rightarrow Malate + NAD^{+}$

19 $Malate + NADP^{+} \rightarrow Pyruvate + CO_2 + NADPH$

20 $Citrate + NADP^{+} \rightarrow 2\_Oxoglutarate + CO_2 + NADPH + H^{+}$

21 $Oxaloacetate + ATP \rightarrow P\_Enolpyruvate + CO_2 + ADP$

22 $NADH + 0.5O_2 + 3ADP + 3Pi + 4H+ \rightarrow NAD + +3ATP + 4H_2O$

23 $FADH2 + 0.5O_2 + 2ADP + 2Pi + 3H+ \rightarrow FAD + 2ATP + 3H_2O$

24 $8Acetyl\_CoA + 14NADPH + 7ATP + 7HCO_3^- + 14H+ \rightarrow Palmitate + 14NADP + +8CoA + 7ADP + 7Pi + 7CO_2 + 6H_2O$

25 $Tripalmitoylglycerol \rightarrow \textbf{Tripalmitoylglycerol}$

26 $Glycerone\_P + 3Palmitate + NADH + 3ATP + H_2O + H^+ \rightarrow Tripalmitoylglycerol + NAD^+ + Pi + 3AMP + 3PPi$

27 $Tripalmitoylglycerol + 3H_2O \rightarrow \textbf{Glycerol} + 3Palmitate$

28 $2\boldsymbol{Acetyl\_CoA} \rightarrow \boldsymbol{Acetoacetate} + 2CoA$

29 $\boldsymbol{Acetoacetyl\_CoA} \rightarrow \boldsymbol{Acetoacetate} + CoA$

30 $\boldsymbol{Acetoacetate} + NADH \rightarrow \textbf{3\_Hydroxybutyrate}$

31 $Pyruvate + NH_4^+ + NADPH \rightarrow Alanine$

32 $Aspartate + NH_4^+ \rightarrow \textbf{Asparagine}$

33 $Aspartate \rightarrow Oxaloacetate + NH_4^+ + NADH$

34 $Cysteine \rightarrow Pyruvate + NH_4^+ + NADH$

35 $Glutamate \rightarrow 2\_Oxoglutarate + NH_4^+ + NADH$

36 $\textbf{Glutamine} \rightarrow Glutamate + NH_4^+ + ATP$

37 $Serine + THF \rightarrow Glycine$

38 $\textbf{Histidine} + THF \rightarrow Glutamate + NH_4^+$

39 $\textbf{Isoleucine} + 2CoA \rightarrow \boldsymbol{Succinyl\_CoA} + \boldsymbol{Acetyl\_CoA} + NH_4^+ + FADH2 + 2NADH$

40 $\textbf{Leucine} + CoA + CO_2 + ATP \rightarrow \boldsymbol{Acetoacetate} + \boldsymbol{Acetyl\_CoA} + NH_4^+ + FADH2 + 2NADH$

41 $\textbf{Lysine} \rightarrow 2\_Oxoadipate + 2NH_4^+ + 3NADH$

42 $2\_Oxoadipate + CoA \rightarrow \boldsymbol{Acetoacetyl\_CoA} + 2CO_2 + FADH2 + 2NADH$

43 $\textbf{Methionine} + Serine + ATP + CoA + THF \rightarrow \boldsymbol{Succinyl\_CoA} + Cysteine + NH_4^+ + NADH$

44 $\textbf{Phenylalanine} + O_2 + NADH \rightarrow Tyrosine$

45 $Glutamate + ATP + 2NADPH \rightarrow \textbf{Proline}$

46 $Serine \rightarrow Pyruvate + NH_4^+$

47 $\textbf{Threonine} + CoA \rightarrow Glycine + \boldsymbol{Acetyl\_CoA} + NADH$

48 $\textbf{Tryptophan} + 3O_2 + NADPH \rightarrow 2\_Oxoadipate + Alanine + CO_2 + NH_4^+$

49 $Tyrosine + 2O_2 \rightarrow \boldsymbol{Acetoacetate} + \boldsymbol{Fumarate} + CO_2 + NH_4^+ + NADH$

50 $\textbf{Valine} + CoA \rightarrow \boldsymbol{Succinyl\_CoA} + CO_2 + 4NADH + FADH2 + NH_4^+$

51 $Palmitate \rightarrow \textbf{Palmitate}$

52 $\boldsymbol{Acetoacetate} \rightarrow \textbf{Acetoacetate}$

53 $Alanine \rightarrow \textbf{Alanine}$

54 $\textbf{Aspartate} \rightarrow Aspartate$

55 $\textbf{Cysteine} \rightarrow Cysteine$

56 $\textbf{Glutamate} \rightarrow Glutamate$

57 $Glycine \rightarrow \textbf{Glycine}$

58 $\textbf{Serine} \rightarrow Serine$

59 $\textbf{Tyrosine} \rightarrow Tyrosine$

60 $\mathbf{O_2} \rightarrow O_2$

61 $CO_2 \rightarrow \mathbf{CO_2}$

62 $NH_4^+ \rightarrow \mathbf{NH_4^+}$

63 $Pyruvate \rightarrow \boldsymbol{Pyruvate}$

64 $\boldsymbol{Citrate} + Malate \rightarrow Citrate + \boldsymbol{Malate}$

65 $2\_Oxoglutarate + \boldsymbol{Malate} \rightarrow \boldsymbol{2\_Oxoglutarate} + Malate$

66 $\boldsymbol{Malate} + Pi \rightarrow Malate + \boldsymbol{Pi}$

*Extracellular metabolites are shown in bold. Mitochondrial metabolites are indicated in bold italics.

Bibliography

[1] Ilana S. Aldor and Jay D. Keasling. Process design for microbial plastic factories: metabolic engineering of polyhydroxyalkanoates. *Current opinion in biotechnology*, 14(5):475–483, October 2003.

[2] Charles E. Nakamura and Gregory M. Whited. Metabolic engineering for the microbial production of 1,3-propanediol. *Current opinion in biotechnology*, 14(5):454–459, October 2003.

[3] Eric J. Steen, Yisheng Kang, Gregory Bokinsky, Zhihao Hu, Andreas Schirmer, Amy McClure, Stephen B. del Cardayre, and Jay D. Keasling. Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature*, 463(7280):559–562, January 2010.

[4] Vincent J. J. Martin, Douglas J. Piteral, Sydnor T. Withers, Jack D. Newman, and Jay D. Keasling. Engineering a mevalonate pathway in escherichia coli for production of terpenoids. *Nature biotechnology*, 21(7):796–802, 2003.

[5] Douglas J. Pitera, Chris J. Paddon, Jack D. Newman, and Jay D. Keasling. Balancing a heterologous mevalonate pathway for improved isoprenoid production in escherichia coli. *Metabolic engineering*, 9(2):193–207, 2007.

[6] Kevin T. Watts, Benjamin N. Mijts, and Claudia Schmidt-Dannert. Current and emerging approaches for natural product biosynthesis in microbial cells. 2005.

[7] Salvador Peiru, Hugo G. Menzella, Eduardo Rodriguez, John Carney, and Hugo Gramajo. Production of the potent antibacterial polyketide erythromycin c in escherichia coli. *Applied and Environmental Microbiology*, 71(5):2539–2547, May 2005.

[8] Blaine A. Pfeifer, Suzanne J. Admiraal, Hugo Gramajo, David E. Cane, and Chaitan Khosla. Biosynthesis of complex polyketides in a metabolically engineered strain of e. coli. *Science*, 291(5509):1790, 2001.

[9] D. K. Ro, E. M. Paradise, M. Ouellet, K. J. Fisher, K. L. Newman, J. M. Ndungu, K. A. Ho, R. A. Eachus, T. S. Ham, J. Kirby, M. C. Chang, S. T. Withers, Y. Shiba, R. Sarpong, and J. D. Keasling. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–3, April 2006.

[10] L. Yarris. An age-old microbe may hold the key to curing an age-old affliction., 2006.

[11] R. Sanders. Launch of antimalarial drug a triumph for uc berkeley, synthetic biology., 2013.

[12] C. A. Mack. Fifty years of moore's law. *Semiconductor Manufacturing, IEEE Transactions on*, 24(2):202–207, 2011.

[13] R. K. Cavin, P. Lugli, and V. V. Zhirnov. Science and engineering beyond moore's law. *Proceedings of the IEEE*, 100(Special Centennial Issue):1720–1749, 2012.

[14] Berin Szoka and Adam Marcus. *The Next Digital Decade: Essays on the Future of the Internet.* TechFreedom, 2011.

[15] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, January 2000.

[16] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38:D355–360, January 2010.

[17] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F. Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic acids research*, 34:D354–357, January 2006.

[18] Ingrid M. Keseler, Amanda Mackie, Martin Peralta-Gil, Alberto Santos-Zavaleta, Socorro Gama-Castro, Cesar Bonavides-Martinez, Carol Fulcher, Araceli M. Huerta, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muniz-Rascado, Quang Ong, Suzanne Paley, Imke Schroder, Alexander G. Shearer, Pallavi Subhraveti, Mike Travers, Deepika Weerasinghe, Verena Weiss, Julio Collado-Vides, Robert P. Gunsalus, Ian Paulsen, and Peter D. Karp. Ecocyc: fusing model organism databases with systems biology. *Nucleic acids research*, 41(D1):D605–D612, January 2013.

[19] Adam M. Feist, Christopher S. Henry, Jennifer L. Reed, Markus Krummenacker, Andrew R. Joyce, Peter D. Karp, Linda J. Broadbelt, Vassily Hatzimanikatis, and Bernhard O. Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol Syst Biol*, 3, June 2007.

[20] Ines Thiele and Bernhard O. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat.Protocols*, 5(1):93–121, 2010.

[21] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[22] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc, New York, NY, USA, 1 edition, 1997.

[23] Ines Thiele, Neema Jamshidi, Ronan MT Fleming, and Bernhard Palsson. Genome-scale reconstruction of escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS computational biology*, 5(3):e1000312, 2009.

[24] Christopher S. Henry, Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982, 2010.

[25] Mona Yousofshahi, Ehsan Ullah, Russell Stern, and Soha Hassoun. Mc3: a steady-state model and constraint consistency checker for biochemical net-

works. *BMC Systems Biology*, 7(1):129, 2013.

[26] Athel Cornish-Bowden. *Fundamentals of enzyme kinetics.* John Wiley and Sons, 2013.

[27] Eberhard O. Voit. *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists.* Cambridge University Press, 2000.

[28] M. Peschel and W. Mende. *The predator-prey model: do we live in a Volterra world?* Springer, 1986.

[29] Vassily Hatzimanikatis and James E. Bailey. Mca has more to say. *Journal of theoretical biology*, 182(3):233–242, 1996.

[30] Diana Visser and Joseph J. Heijnen. The mathematics of metabolic control analysis revisited. *Metabolic engineering*, 4(2):114–123, 2002.

[31] W. Liebermeister and E. Klipp. Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theor Biol Med Model*, 3:42, 2006.

[32] Orland R. Gonzalez, Christoph Kper, Kirsten Jung, Prospero C. Naval, and Eduardo Mendoza. Parameter estimation using simulated annealing for s-system models of biochemical networks. *Bioinformatics*, 23(4):480–486, February 2007.

[33] Linh M. Tran, Matthew L. Rizk, and James C. Liao. Ensemble modeling of metabolic networks. *Biophysical journal*, 95(12):5606–5617, Dec 2008.

[34] Harley H. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 94(3):814–819, 1996.

[35] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2014.

[36] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of physical chemistry*, 81(25):2340–2361, 1977.

[37] Andre S. Ribeiro. Stochastic and delayed stochastic models of gene expression and regulation. *Mathematical biosciences*, 223(1):1–11, Jan 2010.

[38] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–3, Nov 2002.

[39] R. Urbanczik and C. Wagner. An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, 21(7):1203–10, April 2005.

[40] N. Vijayasankaran, R. Carlson, and F. Srienc. Metabolic pathway structures for recombinant protein synthesis in escherichia coli. *Appl Microbiol Biotechnol*, 68(6):737–46, Oct 2005.

[41] A. P. Burgard, E. V. Nikolaev, C. H. Schilling, and C. D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome research*, 14(2):301–312, Feb 2004.

[42] R. P. Carlson. Decomposition of complex microbial behaviors into resource-based stress responses. *Bioinformatics (Oxford, England)*, 25(1):90–97, Jan 2009.

[43] C. T. Trinh, P. Unrean, and F. Srienc. Minimal escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Applied and Environmental Microbiology*, 74(12):3634–3643, Jun 2008.

[44] Jorg Schwender, Fernando Goffman, John B. Ohlrogge, and Yair Shachar-Hill. Rubisco without the calvin cycle improves the carbon efficiency of developing green seeds. *Nature*, 432(7018):779–782, 2004.

[45] Christian Barrett, Markus Herrgard, and Bernhard Palsson. Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation. *BMC Systems Biology*, 3(1):30, 2009.

[46] C. Kaleta, L. F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts? pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome research*, 19(10):1872–1883, Oct 2009.

[47] Anthony Burgard, Priti Pharkya, and Costas Maranas. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657, 2003.

[48] Desmond S. Lun, Graham Rockwell, Nicholas J. Guido, Michael Baym, Jonathan A. Kelner, Bonnie Berger, James E. Galagan, and George M. Church. Large-scale identification of genetic design strategies using local search. *Mol Syst Biol*, 5, August 2009.

[49] Ron Caspi, Hartmut Foerster, Carol A. Fulcher, Pallavi Kaipa, Markus Krummenacker, Mario Latendresse, Suzanne Paley, Seung Y. Rhee, Alexander G. Shearer, Christophe Tissier, Thomas C. Walk, Peifen Zhang, and Peter D. Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 36:D623–631, January 2008.

[50] Ross Overbeek, Tadhg Begley, Ralph M. Butler, Jomuna V. Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valerie de Crecy-Lagard, Naryttza Diaz, Terry Disz, Robert Edwards, Michael Fonstein, Ed D. Frank, Svetlana Gerdes, Elizabeth M. Glass, Alexander Goesmann, Andrew Hanson, Dirk Iwata-Reuyl, Roy Jensen, Neema Jamshidi, Lutz Krause, Michael Kubal, Niels Larsen, Burkhard Linke, Alice C. McHardy, Folker Meyer, Heiko Neuweger, Gary Olsen, Robert Olson, Andrei Osterman, Vasiliy Portnoy, Gordon D. Pusch, Dmitry A. Rodionov, Christian Ruckert, Jason Steiner, Rick Stevens, Ines Thiele, Olga Vassieva, Yuzhen Ye, Olga Zagnitko, and Veronika Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, 33(17):5691–5702, October 2005.

[51] Adam M. Feist, Daniel C. Zielinski, Jeffrey D. Orth, Jan Schellenberger, Markus J. Herrgard, and Bernhard . Palsson. Model-driven evaluation of the production potential for growth-coupled products of escherichia coli. *Metabolic engineering*, 12(3):173–186, May 2010.

[52] Yuki Moriya, Daichi Shigemizu, Masahiro Hattori, Toshiaki Tokimatsu, Masaaki Kotera, Susumu Goto, and Minoru Kanehisa. Pathpred: an

enzyme-catalyzed metabolic pathway prediction server. *Nucleic acids research*, 38(suppl 2):W138–W143, July 2010.

[53] D. C. McShan, S. Rao, and I. Shah. Pathminer: predicting metabolic pathways by heuristic search. *Bioinformatics*, 19(13):1692–1698, September 2003.

[54] Priti Pharkya, Anthony P. Burgard, and Costas D. Maranas. Optstrain: A computational framework for redesign of microbial production systems. 2004.

[55] Esa Pitkanen, Paula Jouhten, and Juho Rousu. Inferring branching pathways in genome-scale metabolic networks. *BMC Systems Biology*, 3(1):103, 2009.

[56] Collin H. Martin, David R. Nielsen, Kevin V. Solomon, and Kristala L. Jones Prather. Synthetic metabolism: Engineering biology at the protein and pathway scales. *Chemistry & biology*, 16(3):277–286, March 2009.

[57] Guillermo Rodrigo, Javier Carrera, Kristala Jones Prather, and Alfonso Jaramillo. Desharky: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics*, 24(21):2554–2556, November 2008.

[58] Vassily Hatzimanikatis, Chunhui Li, Justin A. Ionita, Christopher S. Henry, Matthew D. Jankowski, and Linda J. Broadbelt. Exploring the diversity of complex metabolic networks. *Bioinformatics*, 21(8):1603–1609, April 2005.

[59] Todd Werpy, Gene Petersen, A. Aden, J. Bozell, J. Holladay, J. White, Amy Manheim, D. Eliot, L. Lasure, and S. Jones. *Top value added chemicals from biomass.Volume 1-Results of screening for potential candidates from sugars and synthesis gas*, 2004.

[60] Shelley D. Copley. An evolutionary biochemist's perspective on promiscuity. *Trends in biochemical sciences*, 2015/01. doi: 10.1016/j.tibs.2014.12.004; 22.

[61] Maria Svedendahl Humble and Per Berglund. Biocatalytic promiscuity. *European Journal of Organic Chemistry*, 2011(19):3391–3401, 2011.

[62] H. Nam, N. E. Lewis, J. A. Lerman, D. H. Lee, R. L. Chang, D. Kim, and B. O. Palsson. Network context and selection in the evolution to enzyme specificity. *Science (New York, N.Y.)*, 337(6098):1101–1104, Aug 2012.

[63] A. Hamm, N. Krott, I. Breibach, R. Blindt, and AK Bosserhoff. Efficient transfection method for primary cells. *Tissue Eng*, 8(2):235–245, April 2002.

[64] BI Florea, C. Meaney, HE Junginger, and G. Borchard. Transfection efficiency and toxicity of polyethylenimine in differentiated calu-3 and nondifferentiated cos-1 cell cultures. *AAPS PharmSci*, 4(3):E12, 2002.

[65] Priti Pharkya and Costas D. Maranas. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic engineering*, 8(1):1–13, Jan. 2006.

[66] Shabbir Ahmed and Alexander Shapiro. *Solving Chance-Constrained Stochastic Programs via Sampling and Integer Programming*, pages 261–269. Tutorials in Operations Research. INFORMS, 2008.

[67] A. Charnes and W. W. Cooper. Chance-constrained programming. *Management Science*, 6(1):pp. 73–79, Oct. 1959.

[68] M. Mani and M. Orshansky. A new statistical optimization algorithm for gate sizing, 2004.

[69] Minkang Zhu, Daniel B. Taylor, Subhash C. Sarin, and Randall Kramer. Chance constrained programming models for risk-based economic and policy analysis of soil conservation. *Agricultural and Resource Economics Review*, 23(1), 1994.

[70] A. M. Yeou-Koung Tung. Groundwater management by chance-constrained model. *Journal of Water Resources Planning and Management*, 112:1, 1986.

[71] Wim van Ackooij, Riadh Zorgati, Ren Henrion, and Andris Mller. *Chance Constrained Programming and Its Applications to Energy Management, Stochastic Optimization.* InTech, 2011.

[72] Costas D. Maranas. Optimal molecular design under property prediction uncertainty. *AIChE Journal*, 43(5):1250–1264, 1997.

[73] Kiran Patil, Isabel Rocha, Jochen Forster, and Jens Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6(1):308, 2005.

[74] Oliver Hadicke and Steffen Klamt. Casop: A computational approach for strain optimization aiming at high productivity. *Journal of Biotechnology*, 147(2):88–101, May 2010.

[75] Guido Melzer, Manely Esfandabadi, Ezequiel Franco-Lara, and Christoph Wittmann. Flux design: In silico design of cell factories based on correlation of pathway fluxes to desired properties. *BMC Systems Biology*, 3(1):120, 2009.

[76] Habib Driouch, Guido Melzer, and Christoph Wittmann. Integration of in vivo and in silico metabolic fluxes for improvement of recombinant protein production. *Metabolic engineering*, 14(1):47–58, Jan. 2012.

[77] Hiroyuki Kurata, Quanyu Zhao, Ryuichi Okuda, and Kazuyuki Shimizu. Integration of enzyme activities into metabolic flux distributions by elementary mode analysis. *BMC Systems Biology*, 1(1):31, 2007.

[78] Amit Varma and Bernhard O. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Nat Biotech*, 12(10):994–998, October 1994.

[79] Solomon W. Golomb and Leonard D. Baumert. Backtrack programming. *J.ACM*, 12(4):516–524, oct 1965.

[80] Ioannis Vlahavas and Dimitris Vrakas. *Artificial intelligence for advanced problem solving techniques*. Information Science Reference, 2008.

[81] Torsten Blum and Oliver Kohlbacher. Metaroute: fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics*, 24(18):2108–2109, September 2008.

[82] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[83] Albert-Laszlo Barabasi. Scale-free networks: A decade and beyond. *Science*, 325(5939):412–413, July 2009.

[84] Didier Croes, Fabian Couche, Shoshana J. Wodak, and Jacques van Helden. Metabolic pathfinding: inferring relevant pathways in biochemical networks. *Nucleic acids research*, 33($suppl_2$):W326–330, July 2005.

[85] P. Gerlee, L. Lizana, and K. Sneppen. Pathway identification by network pruning in the metabolic network of escherichia coli. *Bioinformatics*, 25(24):3282–3288, December 2009.

[86] Scott A. Becker, Adam M. Feist, Monica L. Mo, Gregory Hannum, Bernhard O. Palsson, and Markus J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox. *Nat.Protocols*, 2(3):727–738, March 2007.

[87] Ayoun Cho, Hongseok Yun, Jin Park, Sang Lee, and Sunwon Park. Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Systems Biology*, 4(1):35, 2010.

[88] Jack D. Newman, Jessica Marshall, Michelle Chang, Farnaz Nowroozi, Eric Paradise, Douglas Pitera, Karyn L. Newman, and Jay D. Keasling. High-level production of amorpha-4,11-diene in a two-phase partitioning bioreactor of metabolically engineered escherichia coli. *Biotechnology and bioengineering*, 95(4):684–691, 2006.

[89] Tae Seok Moon, Sang-Hwal Yoon, Amanda M. Lanza, Joseph D. Roy-Mayhew, and Kristala L. Jones Prather. Production of glucaric acid from a synthetic pathway in recombinant escherichia coli. *Applied and Environmental Microbiology*, 75(3):589–595, February 2009.

[90] Qiulong Huang, Charles A. Roessner, Rodney Croteau, and A. Ian Scott. Engineering escherichia coli for the synthesis of taxadiene, a key intermediate in the biosynthesis of taxol. *Bioorganic & medicinal chemistry*, 9(9):2237–2242, September 2001.

[91] Yajun Yan, Chia-Chi Lee, and James C. Liao. Enantioselective synthesis of pure (r,r)-2,3-butanediol in escherichia coli with stereospecific secondary alcohol dehydrogenases. *Organic and Biomolecular Chemistry*, 7(19):3914–3917, 2009.

[92] Anita Jakobsen, Inga Aasen, Kjell Josefsen, and Arne Strom. Accumulation of docosahexaenoic acid-rich lipid in thraustochytrid aurantiochytrium sp. strain t66: effects of n and p starvation and o2 limitation. *Applied Microbiology and Biotechnology*, 80(2):297–306, August 2008.

[93] Imandokht Famili and Christophe H. Schilling. Multicellular metabolic models and methods, 2006.

[94] Dokyun Na, Tae Yong Kim, and Sang Yup Lee. Construction and optimization of synthetic pathways in metabolic engineering. *Current opinion in microbiology*, 13(3):363–370, June 2010.

[95] Saikat Saha, Balaji Enugutti, Sona Rajakumari, and Ram Rajasekharan. Cytosolic triacylglycerol biosynthetic pathway in oilseeds. molecular cloning and expression of peanut cytosolic diacylglycerol acyltransferase. *Plant Physiology*, 141(4):1533–1543, August 2006.

[96] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S. Wishart. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041, January 2011.

[97] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, January 2008.

[98] David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration, January 2006.

[99] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205, January 2014.

[100] F. P. Guengerich. Common and uncommon cytochrome p450 reactions related to metabolism and chemical toxicity. *Chemical research in toxicology*, 14(6):611–650, 2001.

[101] J. N. Feige, L. Gelman, D. Rossi, V. Zoete, R. Metivier, C. Tudor, S. I. Anghel, A. Grosdidier, C. Lathion, Y. Engelborghs, O. Michielin, W. Wahli, and B. Desvergne. The endocrine disruptor monoethyl-hexyl-phthalate is a selective peroxisome proliferator-activated receptor gamma modulator that promotes adipogenesis. *The Journal of biological chemistry*, 282(26):19152–19166, Jun 2007.

[102] A. Gerlienke Schuur, Abraham Brouwer, ke Bergman, Michael W. H. Coughtrie, and Theo J. Visser. Inhibition of thyroid hormone sulfation by

hydroxylated metabolites of polychlorinated biphenyls. *Chemico-biological interactions*, 109(1-3):293–297, Feb. 1998.

[103] Monique H. A. Kester, Sema Bulduk, Dick Tibboel, Walter Meinl, Hansruedi Glatt, Charles N. Falany, Michael W. H. Coughtrie, Ake Bergman, Stephen H. Safe, George G. J. M. Kuiper, A. Gerlienke Schuur, Abraham Brouwer, and Theo J. Visser. Potent inhibition of estrogen sulfotransferase by hydroxylated pcb metabolites: A novel pathway explaining the estrogenic activity of pcbs. *Endocrinology*, 141(5):1897–1900, 2000.

[104] Li-Quan Wang, Hans-Joachim Lehmler, Larry W. Robertson, and Margaret O. James. Polychlorobiphenylols are selective inhibitors of human phenol sulfotransferase 1a1 with 4-nitrophenol as a substrate. *Chemico-biological interactions*, 159(3):235–246, 2006.

[105] Gregory G. Oakley, Udaya sankar Devanaboyina, Larry W. Robertson, and Ramesh C. Gupta. Oxidative dna damage induced by activation of polychlorinated biphenyls (pcbs): Implications for pcb-induced oxidative stress in breast cancer. *Chemical research in toxicology*, 9(8):1285–1292, 1996.

[106] Milou ML Dingemans, Aart de Groot, RGDM Van Kleef, Ake Bergman, Martin van den Berg, Henk PM Vijverberg, and Remco HS Westerink. Hydroxylation increases the neurotoxic potential of bde-47 to affect exocytosis and calcium homeostasis in pc12 cells. *Environmental health perspectives*, 116(5):637, 2008.

[107] Kiran Dhakal, Xianran He, Hans-Joachim Lehmler, Lynn M. Teesch, Michael W. Duffel, and Larry W. Robertson. Identification of sulfated metabo-

lites of 4-chlorobiphenyl (pcb3) in the serum and urine of male rats. *Chemical research in toxicology*, 25(12):2796–2804, 2012.

[108] Lynda B. M. Ellis, Junfeng Gao, Kathrin Fenner, and Lawrence P. Wackett. The university of minnesota pathway prediction system: predicting metabolic logic. *Nucleic acids research*, 36(suppl 2):W427–W432, July 2008.

[109] Junfeng Gao, Lynda B. M. Ellis, and Lawrence P. Wackett. The university of minnesota pathway prediction system: multi-level prediction and visualization. *Nucleic acids research*, April 2011.

[110] Bo Kyeng Hou, Lawrence P. Wackett, and Lynda B. M. Ellis. Microbial pathway prediction: A functional group approach. *Journal of chemical information and computer sciences*, 43(3):1051–1057, 2003.

[111] Gilles Klopman, Mario Dimayuga, and Joseph Talafous. Meta. 1. a program for the evaluation of metabolic transformation of chemicals. *Journal of chemical information and computer sciences*, 34(6):1320–1325, 1994.

[112] Gilles Klopman, Meihua Tu, and Joseph Talafous. Meta. 3. a genetic algorithm for metabolic transform priorities optimization. *Journal of chemical information and computer sciences*, 37(2):329–334, 1997.

[113] Joseph Talafous, Lawrence M. Sayre, John J. Mieyal, and Gilles Klopman. Meta. 2. a dictionary model of mammalian xenobiotic metabolism. *Journal of chemical information and computer sciences*, 34(6):1326–1333, 1994.

[114] William G. Button, Philip N. Judson, Anthony Long, and Jonathan D. Vessey. Using absolute and relative reasoning in the prediction of the poten-

tial metabolism of xenobiotics. *Journal of chemical information and computer sciences*, 43(5):1371–1377, 2003.

[115] N. Greene, P. N. Judson, J. J. Langowski, and C. A. Marchant. Knowledge-based expert systems for toxicity and metabolism prediction: Derek, star and meteor. *SAR and QSAR in environmental research*, 10(2-3):299–314, 1999.

[116] Carol A. Marchant, Katharine A. Briggs, and Anthony Long. In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows, meteor, and vitic. *Toxicology Mechanisms and Methods*, 18(2-3):177–187, 2008.

[117] Junfeng Gao, Lynda B. M. Ellis, and Lawrence P. Wackett. The university of minnesota biocatalysis/biodegradation database: improving public access. *Nucleic acids research*, 38(suppl 1):D488–D491, January 2010.

[118] Johannes Kirchmair, Mark J. Williamson, Jonathan D. Tyzack, Lu Tan, Peter J. Bond, Andreas Bender, and Robert C. Glen. Computational prediction of metabolism: Sites, products, sar, p450 enzyme dynamics, and mechanisms. *Journal of Chemical Information and Modeling*, 52(3):617–648, 2012.

[119] P. Jancova, P. Anzenbacher, and E. Anzenbacherova. Phase ii drug metabolizing enzymes. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*, 154(2):103–116, June 2010.

[120] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, January 2000.

[121] Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125(39):11853–11865, 2003.

[122] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic acids research*, 32(suppl 1):D277–D280, January 2004.

[123] Huw Jones. Xenobiotic metabolism and zebrafish (danio rerio) larvae, Dec. 2010.

[124] J. E. Laine, S. Auriola, M. Pasanen, and R. O. Juvonen. Acetaminophen bioactivation by human cytochrome p450 enzymes and animal microsomes. *Xenobiotica*, 39(1):11–21, 2009.

[125] Gabriela Chavarria-Soley, Heinrich Sticht, Eleni Aklillu, Magnus Ingelman-Sundberg, Francesca Pasutto, Andr Reis, and Bernd Rautenstrauss. Mutations in cyp1b1 cause primary congenital glaucoma by reduction of either activity or abundance of the enzyme. *Human mutation*, 29(9):1147–1153, 2008.

[126] Hirotaka Kawakami, Sumio Ohtsuki, Junichi Kamiie, Takashi Suzuki, Takaaki Abe, and Tetsuya Terasaki. Simultaneous absolute quantification of 11 cytochrome p450 isoforms in human liver microsomes by liquid chromatography tandem mass spectrometry with in silico target peptide selection. *Journal of pharmaceutical sciences*, 100(1):341–352, 2011.

[127] Slobodan Rendic and Frederick J. Di Carlo. Human cytochrome p450 enzymes: A status report summarizing their reactions, substrates, inducers, and inhibitors. *Drug metabolism reviews*, 29(1-2):413–580, 1997.

[128] T. Shimada, H. Yamazaki, M. Mimura, Y. Inui, and F. P. Guengerich. Interindividual variations in human liver cytochrome p-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 japanese and 30 caucasians. *Journal of Pharmacology and Experimental Therapeutics*, 270(1):414–423, July 1994.

[129] K. Rowland Yeo, A. Rostami-Hodjegan, and G. T. Tucker. Abundance of cytochromes p450 in human liver: a meta-analysis. *The British Pharmacological Society*, 51:687–688, 2004.

[130] Karthik Venkatakrishnan, Lisa L. von Moltke, Michael H. Court, Jerold S. Harmatz, Charles L. Crespi, and David J. Greenblatt. Comparison between cytochrome p450 (cyp) content and relative activity approaches to scaling from cdna-expressed cyps to human liver microsomes: Ratios of accessory proteins as sources of discrepancies between the approaches. *Drug Metabolism and Disposition*, 28(12):1493–1504, December 2000.

[131] Ulrich M. Zanger and Matthias Schwab. Cytochrome p450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics*, 138(1):103–141, April 2013.

[132] Lies De Bock, Koen Boussery, Pieter Colin, Julie De Smet, Huybrecht T'Jollyn, and Jan Van Bocxlaer. Development and validation of a fast and sensitive uplc-ms/ms method for the quantification of six probe metabolites

for the in vitro determination of cytochrome p450 activity. *Talanta*, 89(0):209–216, Jan. 2012.

[133] Krishna R. Iyer and Michael W. Sinz. Characterization of phase i and phase ii hepatic drug metabolism activities in a panel of human liver preparations. *Chemico-biological interactions*, 118(2):151–169, April 1999.

[134] Vinod Ramachandran, Vsevolod E. Kostrubsky, Bernard J. Komoroski, Shimin Zhang, Kenneth Dorko, James E. Esplen, Stephen C. Strom, and Raman Venkataramanan. Troglitazone increases cytochrome p-450 3a protein and activity in primary cultures of human hepatocytes. *Drug Metabolism and Disposition*, 27(10):1194–1199, October 1999.

[135] Kamlesh Shrivas, Samuel T. Mindaye, Melkamu Getie-Kebtie, and Michail A. Alterman. Mass spectrometry-based proteomic analysis of human liver cytochrome(s) p450. *Toxicology and applied pharmacology*, 267(1):125–136, Feb. 2013.

[136] M.Suzuki K.Tane N.Shimada M.Nakajima T. Yokoi, H.Yamazaki. In vitro inhibitory effects of troglitazone and its metabolites on drug oxidation activities of human cytochrome p450 enzymes: comparison with pioglitazone and rosiglitazone. *Xenobiotica*, 30(1):61–70, 2000.

[137] J A Brotons, M F Olea-Serrano, M Villalobos, V Pedraza, and N Olea. Xenoestrogens released from lacquer coatings in food cans, Jun 1995.

[138] Jackye Peretz, Lisa Vrooman, William A. Ricke, Patricia A. Hunt, Shelley Ehrlich, Russ Hauser, Vasantha Padmanabhan, Hugh S. Taylor, Shanna H.

Swan, and Catherine A. VandeVoort. Bisphenol a and reproductive health: Update of experimental and human evidence, 2007-2013. *Environmental Health Perspectives*, August 2014.

[139] Jan Schmidt, Petra Kotnik, Jurij Trontelj, eljko Knez, and Lucija Peterlin Mai. Bioactivation of bisphenol a and its analogs (bpf, bpaf, bpz and dmbpa) in human liver microsomes. *Toxicology in Vitro*, 27(4):1267–1276, June 2013.

[140] Jean Philippe Jaeg, Elisabeth Perdu, Laurence Dolo, Laurent Debrauwer, Jean-Pierre Cravedi, and Daniel Zalko. Characterization of new bisphenol a metabolites produced by cd1 mice liver microsomes and s9 fractions. *Journal of Agricultural and Food Chemistry*, 52(15):4935–4942, 2004.

[141] Shigeo Nakamura, Yoshito Tezuka, Atsuko Ushiyama, Chiaki Kawashima, Yumina Kitagawara, Kyoko Takahashi, Shigeru Ohta, and Tadahiko Mashino. Ipso substitution of bisphenol a catalyzed by microsomal cytochrome p450 and enhancement of estrogenic activity. *Toxicology letters*, 203(1):92–95, 2011.

[142] Hideo Kurebayashi, Kazuho Okudaira, and Yasuo Ohno. Species difference of metabolic clearance of bisphenol a using cryopreserved hepatocytes from rats, monkeys and humans. *Toxicology letters*, 198(2):210–215, Oct. 2010.

[143] J. J. Pritchett, R. K. Kuester, and I. G. Sipes. Metabolism of bisphenol a in primary cultured hepatocytes from mice, rats, and humans. *Drug Metabolism and Disposition*, 30(11):1180–1185, November 2002.

[144] JA Taylor, FS Vom Saal, WV Welshons, B. Drury, G. Rottinghaus, PA Hunt, PL Toutain, CM Laffont, and CA VandeVoort. Similarity of bisphenol a

pharmacokinetics in rhesus monkeys and mice: relevance for human exposure. *Environ Health Perspect*, 119(4):422–430, April 2011.

[145] Tina Trdan Lusin, Robert Roskar, and Ales Mrhar. Evaluation of bisphenol a glucuronidation according to ugt1a1*28 polymorphism by a new lc-ms/ms assay. *Toxicology*, 292(1):33–41, 2012.

[146] Mitch R. McLean, Udo Bauer, Anthony R. Amaro, and Larry W. Robertson. Identification of catechol and hydroquinone metabolites of 4-monochlorobiphenyl. *Chemical research in toxicology*, 9(1):158–164, 1996.

[147] Leane Lehmann, Harald L.Esch, Patricia A.Kirby, Larry W.Robertson, and Gabriele Ludewig. 4-monochlorobiphenyl (pcb3) induces mutations in the livers of transgenic fisher 344 rats. *Carcinogenesis*, 28(2):471–478, August 2006.

[148] Kevin Park, Dominic P. Williams, Dean J. Naisbitt, Neil R. Kitteringham, and Munir Pirmohamed. Investigation of toxic metabolites during drug development. *Toxicology and applied pharmacology*, 207(2):425–434, 2005.

[149] Minjun Chen, Vikrant Vijay, Qiang Shi, Zhichao Liu, Hong Fang, and Weida Tong. Fda-approved drug labeling for the study of drug-induced liver injury. *Drug discovery today*, 16(15):697–703, 2011.

[150] William M. Lee. Drug-induced hepatotoxicity. *New England Journal of Medicine*, 349(5):474–485, 2003.

[151] U. Diczfalusy and I. Bjorkhem. Still another activity by the highly promiscuous enzyme cyp3a4: 25-hydroxylation of cholesterol. *Journal of lipid research*, 52(8):1447–1449, Aug 2011.

[152] Keisuke Watanabe, Kaori Sakurai, Yuri Tsuchiya, Yasushi Yamazoe, and Kouichi Yoshinari. Dual roles of nuclear receptor liver x receptor (lxr) in the cyp3a4 expression in human hepatocytes as a positive and negative regulator. *Biochemical pharmacology*, 86(3):428–436, 2013.

[153] Baoding Liu. *Theory and Practice of Uncertain Programming.* Springer Publishing Company, Incorporated, 2nd edition, 2009.

[154] Patrice Marcotte and Gilles Savard. Bilevel programming: A combinatorial perspective, 2005.

[155] Benoit Colson, Patrice Marcotte, and Gilles Savard. Bilevel programming: A survey. *4OR: A Quarterly Journal of Operations Research*, 3(2):87–107, 2005.

[156] Xutao Deng, Jun Xu, James Hui, and Charles Wang. Probability fold change: A robust computational approach for identifying differentially expressed gene lists. *Computer methods and programs in biomedicine*, 93(2):124–139, Feb. 2009.

[157] H. C. Wang, Y. H. Ko, H. J. Mersmann, C. L. Chen, and S. T. Ding. The expression of genes related to adipocyte differentiation in pigs. *Journal of animal science*, 84(5):1059–1066, May 2006.

[158] Stefan Schuster, Thomas Dandekar, and David A. Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in biotechnology*, 17(2):53–60, Feb. 1999.

[159] Ryan P. Nolan and Kyongbum Lee. Dynamic model for cho cell engineering. *Journal of Biotechnology*, 158(12):24 – 33, 2012.

[160] Yaguang Si, Jeongah Yoon, and Kyongbum Lee. Flux profile and modularity analysis of time-dependent metabolic changes of de novo adipocyte formation. *American Journal of Physiology - Endocrinology And Metabolism*, 292(6):E1637–E1646, June 2007.

[161] Yaguang Si, Hai Shi, and Kyongbum Lee. Impact of perturbed pyruvate metabolism on adipocyte triglyceride accumulation. *Metabolic engineering*, 11(6):382–390, Nov. 2009.

[162] Sarah L. Davies and David C. James. Engineering mammalian cells for recombinant monoclonal antibody production, 2009.

[163] Moon Sue Lee, Kyoung Wook Kim, Young Hwan Kim, and Gyun Min Lee. Proteome analysis of antibody-expressing cho cells in response to hyperosmotic pressure. *Biotechnology progress*, 19(6):1734–1741, 2003.

[164] Toyoshi Fujimoto, Yuki Ohsaki, Jinglei Cheng, Michitaka Suzuki, and Yuki Shinohara. Lipid droplets: a classic organelle with new outfits, 2008.

[165] Marco Terzer and Joerg Stelling. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–2235, October 2008.

[166] Chang Yeop Han, Atil Y. Kargi, Mohamed Omer, Christina K. Chan, Martin Wabitsch, Kevin D. O'Brien, Thomas N. Wight, and Alan Chait. Differential effect of saturated and unsaturated free fatty acids on the generation of monocyte adhesion and chemotactic factors by adipocytes. *Diabetes*, 59(2):386–396, February 2010.

[167] M. Conforti, G. Cornujols, and G. Zambelli. Polyhedral approaches to mixed integer linear programming. *50 Years of Integer Programming 1958-2008*, pages 343–385, 2010.

[168] A. Makhorin. Glpk (gnu linear programming kit), 2006.

[169] X. Deng. Complexity issues in bilevel linear programming. *Multilevel optimization: algorithms and applications*, 20:149–164, 1998.

[170] Andreas Neuner and Elmar Heinzle. Mixed glucose and lactate uptake by corynebacterium glutamicum through metabolic engineering. *Biotechnology Journal*, 6(3):318–329, 2011.

[171] Miguel Rocha, Paulo Maia, Rui Mendes, Jose Pinto, Eugenio Ferreira, Jens Nielsen, Kiran Patil, and Isabel Rocha. Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinformatics*, 9(1):499, 2008.