Evaluating Enzymatic Synthesis of Small Molecule Drugs

Matthew Moura*, Justin Finkle†, Sarah Stainbrook†, Jennifer Greene*, Linda J Broadbelt*†‡, Keith EJ Tyo*†‡

* Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, 60208
† Interdisciplinary Biological Sciences Graduate Program, Northwestern University, Evanston, IL, 60208
‡ Corresponding authors

© 2015. This manuscript version is made available under the Elsevier user license
http://www.elsevier.com/open-access/userlicense/1.0/
Abstract

There have been many achievements in applying biochemical synthetic routes to the synthesis of commodity chemicals. However, most of these endeavors have focused on optimizing and increasing the yields of naturally existing pathways. We sought to evaluate the potential for biosynthesis beyond the limits of known biochemistry towards the production of small molecule drugs that do not exist in nature. Because of the potential for improved yields compared to total synthesis, and therefore lower manufacturing costs, we focused on drugs for diseases endemic to many resource-poor regions, like tuberculosis and HIV. Using generalized biochemical reaction rules, we were able to design biochemical pathways for the production of eight small molecule drugs or drug precursors and identify potential enzyme-substrate pairs for nearly every predicted reaction. All pathways begin from native metabolites, abrogating the need for specialized precursors. The simulated pathways showed several trends with the sequential ordering of reactions as well as the types of chemistries used. For some compounds, the main obstacles to finding feasible biochemical pathways were the lack of appropriate, natural starting compounds and a low diversity of biochemical coupling reactions necessary to synthesize molecules with larger molecular size.

Keywords: Metabolic engineering, computational biology, predictive biochemistry, pharmaceutical production, drug synthesis
1. Introduction

While organic synthesis is the preferred route for production of small molecule drugs, engineered biosynthesis schemes have shown great promise in changing how drugs can be made. Paclitaxel\(^1\) and artemisinin\(^2\) are two examples in which biochemical pathways were heterologously expressed to dramatically increase available yields. In addition to these two endeavors, enzymes have shown great promise as biocatalysts for single-step reactions in drug synthesis schemes for selective reductions\(^3\), transaminations\(^4\), enantioselective purifications\(^5\), and others.\(^6\)

Enzyme-driven catalytic schemes for small molecules benefit from many advantages. Enzymes natively operate in aqueous conditions, ambient temperatures, and at moderate pH, which can simplify manufacturing and reduce waste treatment requirements. Additionally, the specificity and selectivity of enzymes are typically higher than that achieved through traditional chemical synthesis routes. Enzymes thus have the potential to reduce byproducts, especially in the case of \textit{in vitro} reaction systems. Thus far, however, biosynthetic applications to drug synthesis have focused on either using enzymes for single steps, as previously mentioned, or for the production of natural product drugs. To date, there has been no evaluation of biosynthesis for the full synthesis of non-natural small molecule drugs.

As non-natural biosynthesis pathways will inherently depend on enzymatic activity against non-native substrates, the concept of enzyme promiscuity will be vital for successful novel pathway development. Enzyme promiscuity can be broadly classified as either catalytic or substrate promiscuity\(^7\), where an enzyme either catalyzes a distinct chemistry from that which it typically performs (catalytic) or catalyzes the same basic chemical reaction but on a variety of substrates (substrate). Studies have shown that promiscuity is much more permissive than has
been previously thought. As advances in genomic sequencing and enzyme discovery progress, the coverage of permissible enzymatic chemistry will only grow; however, so too will the difficulty in abstracting this information. Computational tools will therefore be necessary to successfully find enzyme candidates capable of performing these promiscuous reactions.

Predictive biochemistry approaches are required for the in silico design of these types of pathways. There are many tools for the computational design of biochemical pathways, as have been reviewed by our group and others, but the majority of these techniques optimize preexisting metabolic networks for maximizing production levels of known metabolites. Greater use of the tools available for metabolic engineering is necessary to continue the growth of this field towards the novel bioproduction of chemicals. Some examples of techniques that predict biochemical reactions include XTMS, MINE, the KEGG PathPred system, Gem-Path, and the tools utilized in this paper – the Biochemical Network Integrated Computational Explorer (BNICE) and SimZyme. In these cases, biochemistry is distilled into generalized reaction rules which are applied to putative substrates to predict potential enzymatic chemistry. The BNICE software package has been well-documented as a predictor of experimentally validated biochemistry.

In this work, biosynthetic routes for non-natural drug molecules with confirmed clinical relevance were developed. Using BNICE, reaction networks that define the biochemical reactions of each step were designed for several drug molecules. SimZyme, which identifies specific isoenzymes for a particular biochemistry, was then used to identify enzyme candidates for nearly every reaction step. The general process is summarized in Figure 1. To ensure that pathways were completely biosynthetic, successful networks all begin from molecules either
known to exist within the metabolome, or from molecules with previously demonstrated heterologous production routes.

2. Methods

2.1 Assembling an Initial List of Candidate Molecules

To develop a list of candidate small molecules that would benefit from reduced manufacturing costs that may be afforded by biochemical synthesis, multiple sources were utilized. First, the World Health Organization’s (WHO) Global Price Reporting Mechanism\textsuperscript{22} was used to identify drugs that consumed large portions of the WHO’s expenditures in resource poor regions. To identify targets with costs infeasible for mass distribution systems, experts within the Northwestern Center for Global Health, the Bill and Melinda Gates Foundation, the Clinton Global Initiative, and the TB Alliance were also consulted. From this search, a total of 13 molecules were chosen as simulation targets (\textit{Supplementary Table A}).\textsuperscript{17} All targets were selected based on preexistent pharmaceutical relevance, and on having no known methods for synthesizing the chemicals through fully biosynthetic means.

2.2 The BNICE Framework

Reaction networks were generated with BNICE as has been described previously\textsuperscript{9,17,19,23}. In brief, the system uses a set of hand-curated reaction rules, termed “operators”, based on the Enzyme Commission (EC) nomenclature for enzymatic reactions. Operators recognize molecular substructures with demonstrated activity within defined EC classes and search for these patterns on putative substrates. These substructures can include reacting as well as spectator atoms. The operator performs the biochemical reaction common to that EC class on the defined substructure, including the use of any required cofactor molecules, like ATP or NADH, as dictated by the EC entries used to originally define the operator. Using all operators available within the lab from
previous research\textsuperscript{19–21}, these reaction rules were applied iteratively to starting compounds in either forward or reverse (retrosynthetic) directions. Additionally, SimIndex\textsuperscript{18} was used to allow for more efficient and extensive network exploration by pruning networks based on the chemical similarity of pathway intermediates to the desired product.

2.3 Identifying Starting Molecules from Native Metabolism

It was desirable for networks and pathways to connect directly to native metabolites. This would provide a template for future experimental works, where drug molecules could be synthesized without the need for potentially expensive chemical additives to cell culture, or the need for transporting additives into a cellular environment.

Two methods were used to identify biological starting substrates for the drug pathway simulations. Target drug products were analyzed with Kyoto Encyclopedia of Genes and Genomes (KEGG) SimComp\textsuperscript{24} to identify chemically similar naturally occurring metabolites. Alternatively, a retrosynthetic BNICE simulation, with the operators altered to encode the reverse reaction from the originally stated direction, was run to identify hypothetical precursors to the target compound. Compounds predicted in the retrosynthetic network were compared against metabolites found in the iAF1260 \textit{Escherichia coli} metabolic model\textsuperscript{25} or the full KEGG database\textsuperscript{26}.

2.4 Identifying Biochemical Pathways and Networks between Starting Compounds and Drug Product

If a candidate metabolite was found with SimComp, then a metabolic reaction network was generated using BNICE and biased by SimIndex\textsuperscript{18} to save computational effort. If this search did not succeed in resolving to the desired product, a retrosynthetic simulation on the product was run. The forward search network from the SimComp substrate and the retrosynthetic
network from the desired drug product were then matched to search for overlapped chemical intermediates and merged into a single network. Pathways from the candidate metabolite to the final drug product were searched for within this merged network. The pathway search used a Byers-Waterman algorithm to find optimal and near-optimal pathway solutions with pathway lengths at or near the generational step where the product was generated, and was performed after the network generation had completed.¹⁸,²⁷

In the event multiple pathways were found, we used several criteria to prioritize the pathways most amenable to experimental implementation. Pathways with fewer enzymatic steps were prioritized to mitigate uncertainty associated with promiscuous predictions, and inefficiencies that may be associated with long pathways (in terms of increased demand for enzyme synthesis and reduced product yield due to intermediates lost to diffusion). The shortest successful pathways of those that were found are presented in this paper, though some longer examples of note are also included in the supplement. Group contribution thermodynamic calculations were also performed on all reactions to predict the $\Delta G_r$ for each step to evaluate feasibility.²³

For the particular case of tenofovir, successful pathway development required the use of an expanded operator set. During the course of this work, a more broadly-defined operator set, developed elsewhere,²⁸ was successfully used to predict the synthesis of tenofovir. Additionally, this set was also used to develop pathways for two prodrug forms of this molecule: Tenofovir Alafenamide (TAF) and Tenofovir Disoproxil (TDF).

2.5 Identification of Enzyme Candidates for Reactions

Using the predicted substrates and EC class from BNICE, enzymes likely to carry out the desired chemistry on the substrate were identified using SimZyme¹⁸ and the BRENDA²⁹
database. FP2 and FP4-based fingerprints were both used to calculate Tanimoto coefficients\textsuperscript{30,31} between the predicted and known substrates from BNICE and BRENDA, respectively. The SimZyme output thus gives a ranked list of all substrates within all BRENDA entries that share the 3-digit EC values as dictated by BNICE. For each substrate in the SimZyme list, the specific enzyme (4-digit EC) entry is indicated, which can then be used to find the particular isozyme candidates that may be active against the specific substrate. Two chemical transformations did not use this strategy, both within the tert-leucine biosynthesis: (a) the keto-enol shift of the substrate for step 16 was assumed to occur spontaneously, and (b) the decarboxylation class identified for step 17 typically requires an aromatic ring, which was not the case for the putative substrate. However, evidence exists for chemical and biological means for the successful decarboxylation of α-ketoacids, though it may involve an aldehyde intermediate, which would be easily oxidized to the desired carboxylic acid.\textsuperscript{32,33}

Tables 1-3 contain the identified 4-digit EC classes for the three presented pathways in this work, as well as the paired BNICE and BRENDA substrates. Supplementary B contains this same information, the thermodynamic values, and the parent organisms of the EC numbers for the BRENDA-matched substrates.

2.6 Yield Calculations

Yield analysis was conducted in a similar manner to previous publications.\textsuperscript{19} Each overall BNICE-generated design pathway was condensed into a single pathway, including the pathway's starting compounds and the drug molecule of interest as the product. The total number of cofactor-pairs (e.g. ATP/ADP, NADH/NAD) consumed and produced across each overall pathway were also included, as well as any other products released by the pathways.
To calculate the max theoretical yield of each drug molecule, a flux balance analysis calculation was completed for each pathway using the COBRA toolbox for MATLAB\textsuperscript{34}. For each drug, its respective condensed pathway is inserted into the iAF1260\textsuperscript{25} genome-scale metabolic model of \textit{E. coli}. Although all starting substrates have been documented to occur in metabolism, not all are within the iAF1260 model. For those cases, either import reactions were added to allow for the “input” of the precursor, or additional reactions were added to connect the starting molecule to metabolites present in iAF1260. These cases are noted in Table 4, along with the reported maximum theoretical yield values. For each yield calculation, either the uptake flux of either glucose or the import substrate is set to 1 mmol/gDCW/hr. For all pathways, the objective function for the model was set to maximize flux through the condensed drug pathway.

3. Results

Of the list of 13 tested molecules, successful pathways were developed for 8 compounds entirely composed of predictive biochemical reactions. These pathways all stem from metabolites known to exist biochemically and have been assigned individual enzyme classes. Each predicted reaction was linked to enzyme-substrate pairs with demonstrated activity on BRENDA, as well as chemical substrate similarity. Several of these pathways are presented below, while the rest can be found in Supplementary F and G.

3.1 Pyrazinamide

For pyrazinamide, a member of the tetrad combination treatment for \textit{M. tuberculosis}, pyrazine, while not a known natural metabolite, was selected as the starting substrate. Initial pathway analysis identified methyl pyrazine as an intermediate. Because methyl pyrazine is known to be produced by \textit{Corynebacterium glutamicum},\textsuperscript{35} in practice methyl pyrazine may be preferable to unsubstituted pyrazine as a starting compound.
BNICE simulations gave a range of pathways, with the shortest pathways we were able to find shown in Figure 2. Although the pathways for pyrazinamide showed significant branching in the larger network (Supplementary Figure E), all but one of these pathways converged to that presented in Figure 2. This starts with the methylation of the pyrazine ring in step 1, followed by a number of oxidation steps to yield a carboxylic acid that is aminated to give the final drug product. Two short pathway branches are also shown in Figure 2. A single divergent pathway was seen in the expanded network in Supplementary E, where steps 4 and 5 were replaced with a six-step pathway bypassing pyrazinoic acid. Matched substrates, mostly consisting of varying substituted aromatic ring structures, can be seen in Table 1.

3.2 L-Tert Leucine

Searching for pathways for the synthesis of L-tert leucine was motivated by its use as a key precursor in the chemical synthesis of atazanavir, an HIV protease inhibitor for which we could not find a biochemical synthesis pathway. L-tert leucine is a non-biological amino acid with an isobutyl sidechain. This molecule has also demonstrated utility in the synthesis of cancer chemotherapeutics, additional antivirals, and for asymmetric organocatalysis. Using a SimComp search for tert-leucine, 2-dehydropantoate was selected as a starting substrate due to it having a quaternary carbon center. By combining forward and retrosynthetic simulations, paths were found in as few as eight steps.

Though the pathway network in Figure 3 appears complex, the diversity of these reactions is fairly low. Pathways from 2-dehydropantoate to tert-leucine consist of either seven or eight biochemical reactions (not including step 16 which is a spontaneous keto-enol tautomerism). In the first half of the network, steps 1-11, all of the reactions are one of three chemical transformations – an oxidoreduction (e.g., step 1), an aminotransferase (e.g., step 2), or
an aldehyde C-C acetyl transferase (e.g., step 4). These three chemistries occur largely independent of order (though an aldehyde is a prerequisite for the acetylation).

For the remaining reaction steps, the previous parts of the pathways converged to a stricter set of ordered steps. With the lengthened carbon chain, the previously primary, now secondary, alcohol on the tert-butyl group is “shifted” farther along the chain, through an additional hydroxylation in steps 12-15, though even here the order of hydroxylation and dehydration is flexible. This pseudo-isomerization is necessary because the adjacent quaternary carbon prohibits direct dehydration to a double bond. The new structure is then decarboxylated twice to yield the final drug product. All predicted substrates were successfully matched to BRENDA substrates (Table 2), although the match for step 17 was in a class of enzymes that had only demonstrated activity for aromatic substrates. Though this does not preclude the possibility for an enzyme to perform this reaction, further work is necessary to determine the necessity of an aromatic backbone for this reaction.

In addition to the dehydropantoate pathway shown in Figure 3, an analysis of the retrosynthetic network alone suggested an alternative synthesis scheme. The retrosynthetic network predicted an aldehyde C-C ligase reaction between glyoxylate, a common metabolite, and isobutane, a non-metabolite. Inspired by this reaction, an additional simulation was attempted with glyoxylate and prenol, an isoprenoid derivative that contains a C-C double bond to a tertiary carbon that could potentially be ligated to glyoxylate. Though it was hypothesized that this bond would activate C-C bond formation with glyoxylate, the prenol-based network converged to a range of pathways to synthesize isobutane from prenol, followed by the same isobutane pathway. Both are shown in Supplementary D.

3.3 Tenofovir
The final simulated pathway is for the synthesis of tenofovir, a recently developed HIV reverse transcriptase inhibitor. This molecule is distinct as it contains two chemical moieties uncommon in native metabolism: an organophosphate and an ether bond. The formation of organophosphate bonds was catalyzed by phosphoenolpyruvate mutases, producing phosphonopyruvate. However, the initial set of BNICE operators did not encode for the formation of an ether bond, which prevented pathway resolution.

Broader specificity operators, developed based on the MetaCyc database, were then used to successfully find pathways, including a step for the formation of the ether bond (Figure 4). Because the broader specificity resulted in many more predicted reactions, multistep simulations became computationally intractable. To counter this, analysis was conducted one generation at a time with manual curation between each step. Because manual curation precludes the use of large pools of substrates for subsequent generational steps, there were fewer redundancies in this pathway than was the case with the prior molecules. Potential enzyme-substrate pairs were found for each step as shown in Table 3. Steps 1-5 involve phosphorylating the alcohol moieties on 1,2 propanediol, a molecule with demonstrated biosynthetic routes. One phosphate is used with a transferase to bring in the aforementioned phosphonopyruvate (Step 6), creating the ether bond that was not possible with the original operator set. The other is used to exchange in the adenine nucleotide base (Step 7).

In addition to the production routes for tenofovir, the broadened operator set was also used to develop pathways for the biosynthesis of two tenofovir prodrugs: tenofovir alafenamide fumarate (TAF) and tenofovir disoproxil fumarate (TDF). Both pathways were assembled starting from the tenofovir drug product structure itself, as the final prodrugs do not alter the core of the molecule. For TAF, because the change involves a symmetric addition of two carbonate-
based sidechains, the pathway is relatively short. For TDF, the synthesis is more complex to allow for two distinct attached chains. These paths are presented in Supplementary F.

3.4 Yield and Thermodynamics Analysis

Yield calculations varied between the different pathways and, in some cases, variations in yield between pathways for the same molecules were also present. For all possible pathways from the BNICE simulations, a merged overall reaction was used in a flux balance analysis within the iAF1260 model. For most pathways, the starting substrates were already in the model, so yields were calculated based on glucose uptake. In a few cases the model had to be modified to account for metabolites not within the E. coli model but known to exist elsewhere in metabolism. Variation across pathways was a result of analyzing pathways of multiple lengths, and cases where multiple enzyme classes, with different cofactor usage, were predicted to perform the same chemical reaction. Ethambutol showed the lowest yield of 0.15 mmol/mmol substrate. The other pathways all had higher values, with maximum value ranging from 0.47-0.85. Interestingly, pyrazinamide showed the highest theoretical maximum yield of 0.85. Overall reactions, and how the yields varied between the individual pathways for each drug product, are available in Supplementary H.

The thermodynamic predictions for all reactions were considered to be relevant and feasible for the directions shown in the figure of this paper. The highest $\Delta G_r$ was 5.22 kcal/mol, which was the result for the reduction of carboxyls to alcohols, EC class 1.1.1. Not only is it common for enzymes in this family to demonstrate reversibility, but this value is also below similar thresholds used by others. We therefore do not anticipate thermodynamics to be an obstacle for these predicted reactions, which is not surprising since all reaction rules are based on
chemistries known to occur biochemically, and the operators were originally designed with reaction directionality taken into account.

4. Discussion

4.1 Flexible and Rigid Regimes Present Themselves in Multiple Pathways

Several of the pathway networks partitioned into two regimes, described here as flexible and rigid. In the flexible regime, pathways typically utilize diverse operators/reactions, have reactions that show some order-independence of reactions, and typically diverge from or converge to a singular compound. This can be seen particularly well in the network for tert-leucine in Figure 5. In part A, the pathway used 5 different chemical changes, illustrated by the 5 different colored arrows, that tolerated varying substrates throughout the path. The rigid regime, shown in B, involved a much stricter series of reactions, suggesting there were fewer biochemical options available to achieve these transformations and specific ordering was required. The rigid regimes are stricter than the flexible ones in the molecular substructures that are required for the reactions to proceed.

The two regimes were typically evident in the larger pathways that were found. The alternative route presented in Supplementary D for prenol + glyoxylate synthesis explored a wide range of chemical transformations until converging to the same rigid isobutane pathway. Similarly, in an expanded network for the production of pyrazinamide (Supplementary E), the two regimes can also be seen, but reversed; The first two steps (the same as in Figure 2) create (pyrazine-2-yl)methanol, which is further catalyzed more broadly to arrive at the final drug molecule.
The tendency for these BNICE networks to proceed in flexible vs. rigid regimes can influence future pathway development for similar applications. While the current approach removes intermediates that fall below the averaged similarity of each generational pool\textsuperscript{18}, a different methodology could isolate similarity outliers with scores much higher than the rest and artificially enter the rigid regime, allowing the development of longer pathways in a shorter time. Additionally, the operators can now be generated to encode varying levels of molecular substructure specificity in the recognized active sites.\textsuperscript{41} Networks could artificially create this flexible and rigid trend by selective use of the low specificity operators for flexible regimes, and high specificity for rigid.

Flexible and rigid regimes can also inform experimental studies. The rigid regime inherently has fewer enzymatic options, and thus is a reasonable point to begin proof-of-concept studies. If these reactions fail, then there are limited alternative enzyme-substrate options available, and the pathway should be rejected. Alternatively, a reaction that can not be implemented experimentally can be re-run in a BNICE simulation to find multi-step alternative pathways that avoid the unsuccessful reaction. The flexible regime, however, is better for later experiments, as the wide variety of substrate-enzyme pairs will provide alternatives should one reaction fail. Many combinations of enzymes and substrates can be studied in parallel for the flexible regimes, while the rigid steps will be controlled more strongly by biochemical feasibility, independent of enzymatic capabilities.

\textbf{4.2 Designed Pathways and Natural Pathways Use Enzyme Classes at the Same Frequency}

Although our synthetic targets were all non-natural, the biochemical pathways exhibit trends that were analogous to natural pathways. The operators that were the most prevalent in the
pathways were representative of the most populous EC classes. The top four operators used in the networks all fell within the top ten largest EC classes (1.1.1, 1.14.13, 2.1.1, and 4.2.1), and are shown with arrows colored red, blue, green, and purple, respectively, in Figures 2-3. Beyond that, the rest of the operators present in the pathways were within the top 9% of the largest third-level EC families. This suggests that our coverage of enzymatic data is strong for EC families with the broadest substrate representation. The largest classes, with the most available reactions to generalize, gave us the broadest rules. These larger classes (and rulesets) are useful for biosynthesis as they describe chemistries achievable through catalysis on a wide range of substrates. However, these broader enzyme classes may be dominating the network generation, as they generate so many pathway intermediates as to make network generation of long pathways computationally intractable. The remaining less extensive classes contain many interesting potential biochemistries. Smaller EC classes present a challenge, as BNICE operators attempt to generalize observations. Without as many examples, operators may be overly conservative with regard to the substrates they will accept, even if the enzymes are more promiscuous. Only by broadening the experimental observations of an EC class, or by loosening the generalization criteria beyond the EC system for these smaller enzyme families, can we enhance the utilization of particular biochemistries.

4.3 Challenges in Assembling Large Molecules

Successful pathways rarely contain large changes in molecular size or weight. This is partially a result of the focus on chemically similar starting compounds; however, even in the cases where networks did have a large change in molecular weight, like the N-alkyl transferase in the ethambutol pathway (Supplementary G), enzymatic options for those biosynthetic steps were restrictive. Networks tended to converge on those points, suggesting that biosynthesis, or
our current rule set at the very least, contains many rules that excel at chemical modifications and changes in functional groups but are lacking in anabolic biochemistry (i.e., C-C bond forming). This is reflective of native biochemistry as well, as four out of the top five largest native EC families are involved in either the modification or transfer of small functional groups: alcohol oxidoreductions (1.1.1), methylation (2.1.1), acetyl-CoA mediated acetylation (2.3.1), and hydroxylation (1.14.13). These four classes, representing roughly 20% of the defined reactions in KEGG, either do not alter carbon numbers, or do so by no more than two.

Implementing the designed pathways into experimental systems could provide exciting new opportunities for pharmaceutical synthesis, though a major challenge will be ensuring enzyme promiscuity towards non-native substrates. Here, we use SimZyme to identify enzymes with substrates that are chemically similar to the non-native pathway substrates. Our own internal experimental studies have shown roughly a success rate of 75% for these predictions (data not shown). A range of other strategies can be used to find particularly promiscuous enzymes, such as: utilizing known promiscuous superfamilies, predicting based on amino acid sequence, identifying ancestrally divergent enzymes, or engineering broader substrate specificity. The kinetics of the promiscuous rates are likely much lower than would be practical for industrial systems. Fortunately, these types of reactions are prime targets for rational engineering or directed evolution to increase the desired activity.

Computational simulations have had many successful applications for informing novel metabolic engineering strategies. Computational strategies for the engineering of metabolic encompass a wide range of approaches. Some have focused on novel strain design strategies, pathway control, and even applying kinetic simulations. While all of these approaches and others are very useful and important, most pathway design methodologies do not allow for
predictive biochemistry and, thus, will be limited to molecules existent within biology despite many important chemicals for today’s world not being present within metabolism.

However, recent developments have shown that the design of new biosynthetic pathways can dramatically disrupt the limitations of biochemistry. All recent developments, like BNICE, have used predictive reaction rules to design new biosynthetic pathways. GEM-Path, from Campodonico et al, has presented a particularly well-developed framework for the development of novel biosynthetic pathways, not just from the biochemical standpoint but also with respect to strain design and engineering. However, to the author’s knowledge, no other publications have developed pathways for molecules as heterologous to metabolism as those in this study, evidenced by the significantly longer pathways necessary for connecting these molecules to known metabolites. While this does increase the experimental difficulties in implementing such pathways, this work shows that the reaches of putative biochemistry are farther than have been shown thus far.

5. Conclusions

Biochemical pathways for the synthesis of several small molecule drugs are here presented. Starting substrates were chosen from metabolites that were structurally similar to the target compounds. A range of methods were used to generate pathways from the native metabolites to target products, and theoretical maximum yields were calculated for all pathways using flux balance analysis. The predicted pathways often showed an architecture consisting of flexible and rigid regimes, with reactions either demonstrating flexibility or rigidity in the substrates that were required. Further, the set of operators that were represented in our simulations aligned with the broadest types of chemistry known to occur natively. Nearly all
predictions were connected to EC classes containing chemically similar substrates with confirmed activity.

Experimental validation of these types of pathways will, of course, be necessary. Implementing these pathways will require the ability to identify, express, purify, and chemically characterize reaction products from a large number of potential enzymes. High-throughput technologies make this increasingly feasible, but a particularly difficult obstacle will be the necessity to synthesize chemical pathway intermediates to allow for the testing of multiple reactions concurrently. Successes will necessitate the cooperation of not only biochemists, but also organic chemists. Custom synthesis resources will also be useful, though the ability to synthesize some molecules will likely still be intractable.

The BNICE platform is an extensive, robust, and powerful tool for the development of these types of pathways. Further developments in the operators, as well as developments in the ability to accurately predict promiscuous enzyme activities, will enable the implementation of these and similarly developed pathways to change how chemicals are made. Limitations in the design of biochemical pathways that arise from inaccessible chemistry can serve as guideposts for future work in substrate selection. Though it is difficult to definitively state the absolute limits of biochemistry, difficulties with engineering rarer reactions, like promiscuous C-C bond formation, can inform future pathway development and drive the selection of starting substrates, selective operator sets, or the future engineering of new enzymatic reactions.

6. Acknowledgements

The authors would like to thank the Center for Global Health at Northwestern; specifically Dr. Robert Murphy and Dr. Kimberly Scarsi, and the TB Alliance, specifically Dr.
Christopher Cooper and Dr. Takushi Kaneko at the TB Alliance for helpful conversations on selecting priority drugs, and Andrew Stine for the use of his automated operator sets. Funding was provided by the Bill and Melinda Gates Foundation, NSF CBET-0835800, and the Northwestern McCormick School of Engineering.

7. List of Tables

Table 1: Matched substrates for predicted biosynthesis paths for pyrazinamide. When multiple operators performed the same reaction, the SimZyme matched EC number is bold with alternatives in parentheses.

Table 2: Matched substrates for predicted biosynthesis paths for L-tert leucine. When multiple operators performed the same reaction, the SimZyme matched EC number is bold with alternatives in parentheses. Reaction 17 was the single case of an unsuccessfully matched substrate, with the closest match shown with the \( \pi \)-bond of the carbonyl overlaying the \( \pi \)-system of the aromatic ring. \( R_1 \) is a carotenoid-based tail.

Table 3: Matched substrates for predicted biosynthesis paths for tenofovir. Reaction 5 contains only three EC digits because the studied enzyme has only been assigned those numbers.\(^{52}\)

8. List of Figures

Figure 1: Diagram of the work flow for pathway development with BNICE and SimZyme. (A) High-impact targets were selected based on information from public health resources, (B) BNICE was then used to find pathways to those targets, and (C) SimZyme then used chemical similarity to known, active substrates to suggest isozyme candidates.
Figure 2: Predicted paths for the biosynthesis of pyrazinamide. Arrows are colored to show the top four EC classes within all pathways generated. Red – 1.1.1, Blue – 1.14.13, Green – 2.1.1

Figure 3: Predicted paths for the biosynthesis of L-tert leucine. Arrows are colored to show the top four EC classes within all pathways generated. Red – 1.1.1, Blue – 1.14.13, Green – 2.1.1, Purple – 4.2.1

Figure 4: Predicted paths for the biosynthesis of tenofovir.

Figure 5: An illustration of the “flexible” (A) and “rigid” (B) regimes within the tert-leucine pathway. Within the flexible regime, colored arrows correspond to the same chemistry occurring on multiple substrates. The orange dashed line represents a multi-step substitute for the oxidation of an alcohol to an aldehyde, through a methoxy intermediate.

9. References


14. Jeffryes, J. G. *et al.* MINEs: Open access databases of computationally predicted enzyme


32. Vlessis, A. A., Bartos, D. & Trunkey, D. Importance of spontaneous α-ketoacid


Figure 2

Pyrazine

1 → 2 → 3 → 4 → 5 → Pyrazinamide

2 → 6

3 → 7

4 → 8 → 9
<table>
<thead>
<tr>
<th>Reaction Number</th>
<th>EC Class and Cofactors</th>
<th>Validated Substrate</th>
<th>BNICE Substrate</th>
<th>Reaction Number</th>
<th>EC Class and Cofactors</th>
<th>Validated Substrate</th>
<th>BNICE Substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.1.1.6</td>
<td>SAM$^+$ → SAH + H$^+$</td>
<td>![Image]</td>
<td>6</td>
<td>2.1.1.4</td>
<td>SAM$^+$ → SAH + H</td>
<td>![Image]</td>
</tr>
<tr>
<td>2</td>
<td>1.14.14.1</td>
<td>NAD(P)H + cyt-c → H$_2$O + NAD(P)$^+$ + cyt-c</td>
<td>![Image]</td>
<td>7</td>
<td>1.14.13.25</td>
<td>NAD(P)H + H$^+$ + O$_2$ → H$_2$O + NAD(P)$^+$</td>
<td>![Image]</td>
</tr>
<tr>
<td>3</td>
<td>1.1.1.307 (1.2.3.1)</td>
<td>NADP$^+$ → NADPH + H$^+$</td>
<td>![Image]</td>
<td>8</td>
<td>1.2.1.10</td>
<td>NAD$^+$ + CoA → NADH + H$^+$</td>
<td>![Image]</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>O$_2$ + H$_2$O → H$_2$O$_2$</td>
<td>![Image]</td>
<td>9</td>
<td>3.1.2.20</td>
<td>H$_2$O → CoA</td>
<td>![Image]</td>
</tr>
<tr>
<td>5</td>
<td>6.3.1.5</td>
<td>ATP + NH$_3$ → AMP + PP$_i$</td>
<td>![Image]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reaction Number</td>
<td>EC Class and Cofactors</td>
<td>Validated Substrate</td>
<td>BNICE Substrate</td>
<td>Reaction Number</td>
<td>EC Class and Cofactors</td>
<td>Validated Substrate</td>
<td>BNICE Substrate</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>1</td>
<td>1.1.1.184 NADP⁺ → NADPH + H⁺</td>
<td><img src="image1.png" alt="figure" /></td>
<td><img src="image2.png" alt="figure" /></td>
<td>10</td>
<td>1.14.13.25 O₂ + NAD(P)H + H⁺ → H₂O + NAD(P)⁺</td>
<td><img src="image3.png" alt="figure" /></td>
<td><img src="image4.png" alt="figure" /></td>
</tr>
<tr>
<td>2</td>
<td>2.6.1.64 (-1.4.1) Glutamine → 2-oxoglutaramine</td>
<td><img src="image5.png" alt="figure" /></td>
<td><img src="image6.png" alt="figure" /></td>
<td>11</td>
<td>2.3.3.13 AcCoA + H₂O → CoA</td>
<td><img src="image7.png" alt="figure" /></td>
<td><img src="image8.png" alt="figure" /></td>
</tr>
<tr>
<td>3</td>
<td>2.1.1.210 SAM⁺ → SAH + H⁺</td>
<td><img src="image9.png" alt="figure" /></td>
<td><img src="image10.png" alt="figure" /></td>
<td>12</td>
<td>1.14.13.25 O₂ + NAD(P)H + H⁺ → H₂O + NAD(P)⁺</td>
<td><img src="image11.png" alt="figure" /></td>
<td><img src="image12.png" alt="figure" /></td>
</tr>
<tr>
<td>4</td>
<td>2.3.3.13 AcCoA + H₂O → CoA</td>
<td><img src="image13.png" alt="figure" /></td>
<td><img src="image14.png" alt="figure" /></td>
<td>13</td>
<td>4.2.1.79</td>
<td><img src="image15.png" alt="figure" /></td>
<td><img src="image16.png" alt="figure" /></td>
</tr>
<tr>
<td>5</td>
<td>2.6.1.64 (1.4.-1) Glutamine → 2-oxoglutaramine</td>
<td><img src="image17.png" alt="figure" /></td>
<td><img src="image18.png" alt="figure" /></td>
<td>14</td>
<td>4.2.1.9</td>
<td><img src="image19.png" alt="figure" /></td>
<td><img src="image20.png" alt="figure" /></td>
</tr>
<tr>
<td>6</td>
<td>1.1.1.184 NADP⁺ → NADPH + H⁺</td>
<td><img src="image21.png" alt="figure" /></td>
<td><img src="image22.png" alt="figure" /></td>
<td>15</td>
<td>1.14.13.25 O₂ + NAD(P)H + H⁺ → H₂O + NAD(P)⁺</td>
<td><img src="image23.png" alt="figure" /></td>
<td><img src="image24.png" alt="figure" /></td>
</tr>
<tr>
<td>7</td>
<td>2.1.1.210 SAM⁺ → SAH + H⁺</td>
<td><img src="image25.png" alt="figure" /></td>
<td><img src="image26.png" alt="figure" /></td>
<td>16</td>
<td>5.3.2</td>
<td><img src="image27.png" alt="figure" /></td>
<td><img src="image28.png" alt="figure" /></td>
</tr>
<tr>
<td>8</td>
<td>2.6.1.64 (1.4.-1) Glutamine → 2-oxoglutaramine</td>
<td><img src="image29.png" alt="figure" /></td>
<td><img src="image30.png" alt="figure" /></td>
<td>17</td>
<td>1.14.13.1 O₂ + NADH + H⁺ → H₂O + NAD⁺</td>
<td><img src="image31.png" alt="figure" /></td>
<td><img src="image32.png" alt="figure" /></td>
</tr>
<tr>
<td>9</td>
<td>2.6.1.64 (1.4.-1) Glutamine → 2-oxoglutaramine</td>
<td><img src="image33.png" alt="figure" /></td>
<td><img src="image34.png" alt="figure" /></td>
<td>18</td>
<td>4.1.1.4</td>
<td><img src="image35.png" alt="figure" /></td>
<td><img src="image36.png" alt="figure" /></td>
</tr>
</tbody>
</table>

Table 2
<table>
<thead>
<tr>
<th>Reaction Number</th>
<th>EC Class and Cofactors</th>
<th>Validated Substrate</th>
<th>BNICE Substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.7.2.1 ATP→ADP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.7.1.165 ATP→ADP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5.4.2.1 AND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.1.3.13 P_i→H_2O</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.7.2- ATP→ADP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.5.1.19 -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.4.2.1 +</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug</td>
<td>Max Yield Range (mmol&lt;sub&gt;prod&lt;/sub&gt;/mmol&lt;sub&gt;C-source&lt;/sub&gt;)</td>
<td>Carbon Source</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.15</td>
<td>2,3-diaminopropanoate^*</td>
<td></td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.37-0.85</td>
<td>Pyrazine†</td>
<td></td>
</tr>
<tr>
<td>Didanosine</td>
<td>0.3-0.52</td>
<td>Glucose</td>
<td></td>
</tr>
<tr>
<td>Stavudine</td>
<td>0.54-0.55</td>
<td>Glucose</td>
<td></td>
</tr>
<tr>
<td>Tert-leucine</td>
<td>0.35-0.63</td>
<td>Glucose</td>
<td></td>
</tr>
<tr>
<td>Tenofovir</td>
<td>0.47</td>
<td>Glucose†</td>
<td></td>
</tr>
</tbody>
</table>