Towards creating an extended metabolic model (EMM) for *E. coli* using enzyme promiscuity prediction and metabolomics data

Sara A. Amin*, Department of Computer Science, Tufts University, Medford, MA, sara.amin@tufts.edu

Elizabeth Chavez*, Department of Biology, University of North Carolina, Chapel Hill, NC celiz@live.unc.edu

Nikhil U. Nair†, Department of Chemical and Biological Engineering, Tufts University, Medford, MA nikhil.nair@tufts.edu, and

Soha Hassoun†, Departments of Computer Science and Department of Chemical & Biological Engineering, Tufts University, Medford, MA, soha.hassoun@tufts.edu

*Equal contributions

†Co-corresponding authors

Abstract

Background

Metabolic models are indispensable in guiding cellular engineering and in advancing our understanding of systems biology. As not all enzymatic activities are fully known and/or annotated, metabolic models remain incomplete, resulting in suboptimal computational analysis and leading to unexpected experimental results. We posit that one major source of unaccounted metabolism is promiscuous enzymatic activity. It is now well-accepted that most, if not all, enzymes are promiscuous – i.e., they transform substrates other than their primary substrate. However, there have been no systematic analyses of genome-scale metabolic models to predict putative reactions and/or metabolites that arise from enzyme promiscuity.

Results

Our workflow utilizes PROXIMAL – a tool that uses reactant-product transformation patterns from the KEGG database – to predict putative structural modifications due to promiscuous enzymes. Using iML1515 as a model system, we first utilized a computational workflow, referred to as Extended Metabolite Model Annotation (EMMA), to predict promiscuous reactions catalyzed, and metabolites produced, by natively encoded enzymes in *E. coli*. We predict hundreds of new metabolites that can be used to augment iML1515. We then validated our method by comparing predicted metabolites with the *Escherichia coli* Metabolome Database (ECMDB).

Conclusions

We utilized EMMA to augment the iML1515 metabolic model to more fully reflect cellular metabolic activity. This workflow uses enzyme promiscuity as basis to predict hundreds of reactions and metabolites that may exist in *E. coli* but have not been documented in iML1515 or

other databases. Among these, we found that 17 metabolites have previously been documented in *E. coli* metabolomics studies. Further, 6 of these metabolites are not documented for any other *E. coli* metabolic model (e.g. KEGG, EcoCyc). The corresponding reactions should be added to iML1515 to create an Extended Metabolic Model (EMM). Other predicted metabolites and reactions can guide future experimental metabolomics studies. Further, our workflow can easily be applied to other organisms for which comprehensive genome-scale metabolic models are desirable.

Keywords

Metabolic engineering, enzyme promiscuity, extended metabolic model, systems biology, enzyme activity prediction

Background

The engineering of metabolic networks has enabled the production of high-volume commodity chemicals such as biopolymers and fuels, therapeutics, and specialty products [1-3]. Producing such compounds requires transforming microorganisms into efficient cellular factories [4-7]. Biological engineering has been aided via computational tools for constructing synthesis pathways strain optimization, elementary flux mode analysis, discovery of hierarchical networked modules that elucidate function and cellular organization, and many others (e.g., [8-12]). These design tools rely on organism-specific metabolic models that represent cellular reactions and their substrates and products. Model reconstruction tools [13, 14] use homology search to assign function to Open

Reading Frames obtained through sequencing and annotation. Once the function is identified, the corresponding biochemical transformation is assigned to the gene. Additional biological information such as gene-protein-reaction associations is utilized to refine the models. Exponential growth in sequencing has resulted in an "astronomical", or better yet, "genomical", number of sequenced organisms [15]. There are now databases (e.g., KEGG [16], BioCyc [17], and BiGG [18]) that catalogue organism-specific metabolic models. Despite progress in sequencing and model reconstruction, the complete characterizing of cellular activity remains elusive, and metabolic models remain incomplete. One major source of uncatalogued cellular activity is attributed to orphan genes. Because of limitations of homology-based prediction of protein function, there are millions of protein sequences that are not assigned reliable functions [19]. Integrated strategies that utilize structural biology, computational biology, and molecular enzymology continue to address assigning function to orphan genes [20].

We focus in this paper on another major source of uncatalogued cellular activity − promiscuous enzymatic activity, which has recently been referred to as 'underground metabolism' [21, 22]. While enzymes have widely been held as highly-specific catalysts that only transform their annotated substrate to product, recent studies show that enzymatic promiscuity – enzymes catalyzing reactions other than their main reactions – is not an exception but can be a secondary task for enzymes [23-27]. More than two-fifths (44%) of KEGG enzymes are associated with more than one reaction [28]. Promiscuous activities however are not easily detectable *in vivo* since, i) metabolites produced due to enzyme promiscuity may be unknown, ii) product concentration due to promiscuous activity may be low, iii) there is no high-throughput way to relate formed products to specific enzymes, and iv) it is difficult to identify potentially unknown metabolites in complex

biological samples. Outside of *in vitro* biochemical characterization studies to predict promiscuous activities, there are few resources that record details about promiscuous enzymes such as MINEs Database [29], and ATLAS [30]. Despite the current wide-spread acceptance of enzyme promiscuity, and its prominent utilization to engineer catalyzing enzymes in metabolic engineering practice [31-34], promiscuous enzymatic activity is not currently fully documented in metabolic models. Advances in computing and the ability to collect large sets of metabolomics data through untargeted metabolomics provide an exciting opportunity to develop methods to identify promiscuous reactions, their catalyzing enzymes, and their products that are specific to the sample under study. The identified reactions can then be used to complete existing metabolic models.

We describe in this paper a computational workflow that aims to extend preexisting models with reactions catalyzed by promiscuous native enzymes and validate the outcomes using published metabolomics datasets. We refer to the augmented models as Extended Metabolic Models (EMMs), and to the workflow to create them as EMMA (EMM Annotation). Each metabolic model is assumed to have a set of reactions and their compounds and KEGG reaction IDs. Each reaction is assumed to be reversible unless indicated otherwise. EMMA utilizes PROXIMAL [35], a method for creating biotransformation operators from KEGG reactions IDs using RDM (Reaction Center, Difference Region, and Matched Region) patterns [36], and then applying the operators to given molecules. While initially developed to investigate products of Cytochrome P450 (CYP) enzymes, highly promiscuous enzymes utilized for detoxification, the PROXIMAL method is generic. To create an EMM for a known metabolic model, PROXIMAL generates biotransformation operators for each reaction in the model and then applies the operators to known metabolites within the model. The outcome of our workflow is a list of putative metabolites due

to promiscuous enzymatic activity and their catalyzing enzymes and reactions. In this work, we apply EMMA to iML1515, a genome-scale model of *Escherichia coli* MG1655 [37]. EMMA predicts hundreds of putative reactions and their products due to promiscuous activities in *E. coli*. The putative products are then compared to measured metabolites as reported in *Escherichia coli* Metabolome Database, ECMDB [38, 39]. We identify 17 metabolites that are in ECMDB but not in iML15151. Out of the 17 generated metabolites, 11 are already documented in other *E. coli* databases (e.g. EcoCyc [40], and KEGG). The remaining 6 reactions and their metabolites have not been previously recorded as part of *E. coli* metabolism. We therefore recommend extending *E. coli* model iML15151 with at least 17 new reactions that are validated by existing metabolomics data.

Results

Application of PROXIMAL to iML1515 yielded a lookup table with 1,875 biotransformation operator entries. When applied to 106 high concentration metabolites [41] in iML1515, these operators predicted the formation of 1,423 known metabolites of which 1,368 were new to this model. Our workflow recommended 17 balanced reactions that can be used to augment the iML1515 model.

These identified reactions were divided into four categories, C1–C4. The rationale for the various categories is explained using a decision tree (**Fig. 1**), a machine learning model that classifies data into groupings that share similar attributes [42]. With the exception of leaf nodes, each node in the tree tests the presence or absence of a particular attribute. Left branches represent the presence of

the attribute, while the right branch represents the attribute's absence. Each leaf node represents a classification category and is associated with a subset of the 17 reactions. At the root node of the decision tree, we tested if a PROXIMAL predicted metabolite is in the iML1515 model. If it is, and if the enzyme catalyzing the reaction within iML1515 model producing this metabolite is different than the enzyme PROXIMAL used to predict the relevant biotransformation, then it is classified in Category 1 (C1). Reactions belonging to C1 are parallel transformation to the ones in the model. They represent novel biotransformation routes between existing metabolites since they are generated using a different gene/enzyme than what is reported in iML1515. If previous conditions do not apply to the predicted product, then it is discarded as the reaction is already in iML1515.

If a predicted metabolite is not one of the known metabolites in iML1515, the decision tree determines whether the predicted metabolite and reaction set is associated with *E. coli* in other databases (KEGG and EcoCyc). If the biotransformation is present in KEGG or EcoCyc, then the predicted metabolite is classified into Category 2 (C2), reflecting a curation issue where some reactions were not included in the iML1515 model. If the predicted metabolite is not in iML1515 and not associated with *E. coli* in KEGG nor listed in EcoCyc, then the decision tree determines if the same chemical transformation (same substrate and same product) is documented to occur in other organisms. Predicted biotransformations documented in KEGG for organisms other than *E. coli* are classified in Category 3 (C3). While biotransformations not found in KEGG are classified as Category 4 (C4).

Each Category consists of a set of reactions. C1 consists of five reactions that are predicted to be catalyzed by enzymes that are different than those in iML1515 (**Fig. 2**). The redox transformation between L-alanine and 2-aminoacrylic acid (**Fig. 2A**), is predicted to be catalyzed by EC 1.3.1.98 (UDP-*N*-acetylmuramate dehydrogenase). 2-Aminoacrylic acid, also known as dehydroalanine, is also formed/consumed in *E. coli* due to EC 4.3.1.17 (serine deaminase). Another predicted reaction is the redox transformation between 2-oxoglutarate and 2-hydroxyglutarate by EC 1.1.1.79 (glyoxylate reductase) (**Fig. 2B**). 2-Hydroxyglutarate is involved in reactions associated with EC 1.1.1.95 (phosphoglycerate dehydrogenase) in *E. coli*. The phosphoribosyltransferase reaction between cytosine and cytidine-5'-monophosphate (CMP) is predicted to occur in *E. coli* due to EC 2.4.2.10 (orotate phosphoribosyltransferase) (**Fig. 2C**). CMP, a nucleotide, is already known to be involved in a number of *E. coli* reactions – ECs 2.4.99.12, 2.4.99.13, 2.4.99.14, 2.4.99.15 (all of which are the same KDO transferase), 2.7.1.48 (uridine kinase), 2.7.4.25 (dCMP kinase), 2.7.8.5 (glycerol-3-phosphate phosphatidyltransferase), 2.7.8.8 (phosphatidylserine synthase), 3.1.3.5 (5'-nucleotidase), 3.2.2.10 (pyrimidine-5'-nucleotide nucleosidase), 3.6.1.9 (nucleotide diphosphatase), 3.6.1.26 (CDP diacylglycerol hydrolase), 3.6.1.65 (CTP diphosphatase), 4.6.1.12, (MECDP synthase) and 6.3.2.5 (phosphopantothenate-cysteine ligase) [16]. Another predicted reaction is the transformation between bicarbonate and carboxyphosphate catalyzed by EC 3.6.1.7 (acylphosphatase) (**Fig. 2D**). Carboxyphosphate is also formed/consumed in *E. coli* due to EC 6.3.5.5 (carbamoyl-phosphate synthase). The last prediction is the coenzyme A transferase reaction between acetoacetyl-CoA and acetoacetate due to EC 2.8.3.10 (citrate CoA-transferase) (**Fig. 2E**). Acetoacetate is also known to be formed/consumed in *E. coli* due to ECs 2.8.3.8 (acetate CoA-transferase) and 2.8.3.9 (butyrate-acetoacetate CoA-transferase).

C2 consists of six reactions known to be in *E. coli* but missing from the iML1515 model. The first predicted reaction is the aminoacyltransferase reaction between L-glutamate and γ-glutamyl-β-cyanoalanine due to EC 2.3.2.2 (γ-glutamyltransferase) (**Fig. 3A**). The second is a predicted ligase reaction between L-glutamic acid and THF to form/consume THF-L-glutamic acid by EC 6.3.2.17 (tetrahydrofolate synthase) (**Fig. 3B**). The third is an acyltransferase transformation between propanoyl-CoA and 2-methylacetoacetyl-CoA catalyzed by EC 2.3.1.9 (acetoacetyl-CoA thiolase) (**Fig. 3C**). Fourth, PROXIMAL predicted the phosphotransferase reaction between of D-ribulose-5-phosphate and D-ribulose-1,5-bisphosphate by EC 2.7.1.19 (phosphoribulokinase) (**Fig. 3D**). The fifth predicted reaction known to be in *E. coli* is the redox transformation of D-gluconic acid to 2-keto-D-gluconic acid by EC 1.1.1.215 (gluconate 2-dehydrogenase) (**Fig. 3E**). Lastly, the workflow predicted glycosyltransferase transformation of 5-amino-4-imidazolecarboxamide to/from 1-(5'-phosphoribosyl)-5-amino-4-imidazolecarboxamide by EC 2.4.2.7 (AMP pyrophosphorylase) (**Fig. 3F**).

C3 consists of three predicted reactions that are not documented in *E. coli* but are known in other organisms. The first of these, the transformation between pyruvate and 4-carboxy-4-hydroxy-2-oxoadipate (**Fig. 4A**) catalyzed by EC 4.1.3.17 (HMG aldolase), is present in many organisms, including bacteria, as part of the benzoate degradation pathway (KEGG R00350). The transformation is predicted to occur in *E. coli* due to EC 4.1.3.34 (citryl-CoA lyase). Both EC 4.1.3.17 and EC 4.1.3.34 are lyases enzymes that form carbon-carbon bonds. 4-Carboxy-4-hydroxy-2-oxoadipate is known to be formed/consumed by EC 4.2.1.80 (2-keto-4-pentenoate hydratase) in *E. coli* (KEGG R04781). Another predicted reaction is the (de)aminating redox transformation between L-histidine and imidazol-5-yl-pyruvate, catalyzed by EC 1.4.1.4

(glutamate dehydrogenase) (**Fig. 4B**). Imidazol-5-yl-pyruvate is not known to be produced in any other way in *E. coli*, according to ECMDB and KEGG databases. The transformation of L-histidine to/from imidazol-5-yl-pyruvate is known to occur in the bacterium *Delftia acidovorans* by EC 2.6.1.38 (histidine transaminase) [43]. Lastly, C3 includes the predicted aryltransferase reaction between geranyl diphosphate and geranyl hydroxybenzoate by EC 2.5.1.39 (4-hydroxybenzoate transferase) (**Fig. 4C**). While the general reaction of all-*trans*-polyprenyl diphosphate to 4-hydroxy-3-polyprenylbenzoate is known to occur in *E. coli*, the specific transformation between geranyl diphosphate to geranyl hydroxybenzoate is known to occur in plants as part of shikonin biosynthesis, by EC 2.5.1.93 (4-hydroxybenzoate geranyltransferase) [44].

C4 consists of three predicted reactions that are not currently catalogued in KEGG for any organism (**Fig. 5**). The first reaction (**Fig. 5A**) is the oxidoreductive interconversion between aminomalonate and L-serine by EC 1.1.1.23 (histidinol dehydrogenase). There is one reaction (KEGG R02970) catalyzed by EC 2.6.1.47 (L-alanine:oxomalonate aminotransferase) that produces aminomalonate; but it is not a redox reaction and is associated with rat and silkworm, not *E. coli* [45]. The second, is a hydrolytic decarboxylation reaction between *N*-acetylputrescine and *N*-acetylornithine (**Fig. 5B**) predicted to be catalyzed by EC 4.1.1.36 (PPC decarboxylase). The product, *N*-acetylputrescine, is involved in a number of enzymatic reactions – ECs 1.4.3.4 (monoamine oxidase), 2.3.1.57 (spermidine acetyltransferase), and 3.5.1.62 (acetylputrescine deacetylase) – in many organisms that include both eukaryotes and bacteria [16]. The last reaction in this category is the hydrolytic decarboxylation reaction between 3-ureidopropionate and *N*-carbamoyl-L-aspartate also catalyzed by EC 4.1.1.36 (PPC decarboxylase). 3-Ureidopropionate is

present in eukaryotes and bacteria (but not *E. coli*) and is involved in reactions catalyzed by ECs 3.5.1.6 (β-ureidopropionase) and 3.5.2.2 (dihydropyrimidinase).

Discussion

Current practices for reconstructing genome-scale metabolic models, which are derived using sequencing and functional annotation, can be improved by utilizing metabolomics data. However, directly utilizing metabolomics measurements to augment existing models is challenging. Not every metabolite is measurable due to limited resolution and fidelity of mass spectrometry instruments. Further, assigning chemical identities to measured metabolites remains a challenge. Even if new metabolites are identified, their formation cannot be easily assigned to enzymes without significant experimental effort involving either genetic or biochemical screens. Additionally, metabolomics data alone cannot differentiate reactions catalyzed by different enzymes yet between the same substrates-product pairs without additional experimental efforts. Computational tools and workflows, as presented in this paper, can significantly guide such studies and aid in metabolic model construction and augmentation based on metabolomics data.

The workflow that we developed here is designed to identify metabolites that can form due to promiscuous enzymatic activity. Further, the workflow provides balanced reactions to document such enzymatic activities. We utilized PROXIMAL [35], which first identifies patterns of structural transformations associated with enzymes in the biological sample and then applies these transformations to known sample metabolites to predict putative metabolic products. Using PROXIMAL in this way allows attributing putative metabolic products to specific enzymatic

activity and deriving balanced biochemical reactions that capture the promiscuous activity. Using PROXIMAL offers an additional advantage – the derived promiscuous transformations are specific to the sample under study, and are not limited to hand-curated biotransformation operators as in prior works [29, 30]. PROXIMAL therefore allows exploration of a variety of biotransformations that are commensurate with the biochemical diversity of the biological sample. The EMMA workflow, which utilized PROXIMAL, was previously developed to engineer a candidate set from a metabolic model for metabolite identification [49]. EMMA did not aim to augment existing metabolic models or derive balanced reactions as utilized in this study.

Future experimental and computational efforts can further advance this work. Experimentally, the list of putative products generated by PROXIMAL but not documented in any metabolomics databases can be used as a resource to identify as yet unidentified metabolites. Experimental validation of reactions in the various categories, especially C3 and C4, provide a means for expanding existing databases such as KEGG and EcoCyc. Computationally, PROXIMAL can be upgraded to consider enzymes that act on more than one Reaction Center (R) within a metabolite (e.g. transketolase). This would produce multiple operators per reaction and generate a more comprehensive list of putative reactions and products. We applied PROXIMAL transformation patterns to only 106 high concentration metabolites with the goal of increasing the probability verifiable predictions. Derivatives of metabolites with lower concentration, however, can also be considered. Additionally, we applied only one iteration of the workflow – i.e., we did not consider whether products of promiscuous reactions can themselves act as new substrates for promiscuous reactions. This is due to the large number of putative products. We are currently developing machine learning techniques to improve the prediction accuracy of PROXIMAL.

Conclusion

This study investigates creating Extended Metabolic Models (EMMs) through the augmentation of existing metabolic models with reactions due to promiscuous enzymatic activity. Our workflow, EMMA, first utilizes PROXIMAL to predict putative metabolic products, and then compares these products against metabolomics data. EMMA was applied to iML1515, the genome-scale model of *E. coli* MG1655. PROXIMAL generated 1,875 biochemical operators based on reactions in iML1515 and predicted 1,368 derivatives of 106 high-concentration metabolites. To validate these products, EMMA compared the set of putative derivatives with the set of metabolites documented in ECMDB as part of *E. coli* metabolism. For the overlapping set, we generated corresponding atom-balanced reactions by adding suitable cofactors and/or co-substrates to the substrate-derivative pair suggested by PROXIMAL. The balanced reactions were compared with data recorded in EcoCyc and KEGG. Our workflow generated a list of 17 new reactions that should be utilized to extend the iML1515 model, including parallel reactions between existing metabolites, novel routes to existing metabolites, and new paths to new metabolites. Importantly, this study is foundational in providing a systemic way of coupling computational predictions with metabolomics data to explore the complete metabolic repertoire of organisms. Applying this workflow to other biological samples and their metabolomics data promise to enhance our understanding of natural, synthetic, and xenobiotic metabolism.

Methods

The EMMA workflow was customized to augment the *E. coli* iML1515 model based on the availability of the metabolic measurements in ECMDB, and the availability of cataloged reactions and metabolites for *E. coli* in other databases (EcoCyc and KEGG) (**Fig. 6**). The iML1515 model consists of 1,877 metabolites, 2,712 reactions and 1,516 genes. Our workflow consists of the following three steps.

Step 1 – Predict promiscuous products using PROXIMAL

EMMA used PROXIMAL to predict putative products that can be added to the model. PROXIMAL utilizes RDM patterns [36] specific to the model's reactions to create lookup tables that map reaction centers to structural transformation patterns. An RDM pattern specifies local regions of structural similarities/differences for reactant-product pairs based on a given biochemical reaction. An RDM pattern consists of three parts: i) A Reaction Center (R) atom exists in both the substrate and reactant molecule and is the center of the molecular transformation. ii) Difference Region (D) atoms are adjacent to the R atom and are distinct between substrate and product. iii) Matched Region (M) atoms are adjacent to the R atom but remain unmodified by the transformation. All atoms are labelled using KEGG atom types [50]. Only transformations requiring the presence of one Reaction Centers (R) for the biotransformation to occur are utilized. PROXIMAL constructs a lookup table of all possible biotransformations that can occur due to promiscuous activity of enzymes based on the RDM patterns of reactions catalyzed by enzymes associated with genes in the iML1515 gene list. The "key" in the lookup table consisted of the R and M atom(s) in the reactant, while the "value" is the R and D atom(s) in the product. RDM

patterns were initially available through the (RPAIR) database, but they are now catalogued in KEGG's RClass database. The biotransformation operators in the lookup table were then applied to model metabolites. To increase the probability of predicting verifiable reactions, the biotransformations were applied to only 106 metabolites in iML1515 with predicted or measured concentration values above 1 µM [41]. The assumption here is that high concentration metabolites are more likely to undergo transformation by promiscuous enzymatic activity and form detectible derivatives. The outcome of this step is a list of predicted products due to putative enzymatic activity.

Step 2 – Compare predicted products with metabolomics dataset

Metabolites predicted by PROXIMAL were compared with measured metabolic data in ECMDB. ECMDB contains 3,760 metabolites detected in *E. coli* strain K-12 and related information such as reactions, enzymes, pathways, and other properties. This information was either collected from resources and databases such as EcoCyc, KEGG, EchoBase [51], UniProt [52, 53], YMDB [54], and CCDB [55], or from literature, or validated experimentally by the creators of ECMDB. Partial information about metabolites such as KEGG compound IDs, metabolites cell location, and chemical formulas is provided in ECMDB.

For each putative product, a mol file was generated and then converted to a SMILES string using Pybel [56], a python wrapper for the chemical toolbox Open Babel [57]. Based on the SMILES string, we initially retrieved the corresponding PubChem ID and InchiKey from PubChem using Pybel. To ensure consistency, we confirmed that retrieved PubChem IDs and InchiKeys of

PROXIMAL predicted metabolites matched the corresponding entries in ECMDB. During this process, we noted some discrepancies. In some cases, the information retrieved from PubChem, such as InchiKeys did not match those in ECMDB. In cases of a mismatch, we sought additional information to confirm metabolite identities of ECMDB products. We utilized the values of the CAS ID, BioCyc ID, Chebi ID and KEGG ID fields to retrieve PubChem IDs using Pybel. The retrieved PubChem IDs are used to determine the ID through a majority vote. For example, if the PubChem ID associated with InchiKey, KEGG ID and CAS ID matched, but did not match the PubChem ID provided in ECMDB, then we considered the one retrieved by Pybel as the correct PubChem ID. Out of 3,760 metabolites in ECMDB, we identified 3,397 metabolites with consistent information with data retrieved from PubChem. Once PubChem IDs were identified for ECMDB metabolites, we compared our predicted metabolites against ECMDB metabolites using PubChem IDs.

Step 3 – Curation of stoichiometric reactions

If a metabolite predicted by PROXIMAL was in ECMDB, then steps 1 and 2 resulted in the identification of a *verifiable* predicted promiscuous transformation of an *E. coli* metabolite. Otherwise, our analysis in step 1 yielded a putative transformation. For each verifiable predicted transformation by PROXIMAL, we developed a new reaction by examining the reaction(s) template associated with the enzymatic transformation and adding suitable cofactors to the reactant and product of the biotransformation identified. The set of balanced reactions developed, where the added cofactors to a reaction caused the number of atoms of reactants and products to match on both sides of the reaction, are then compared to reactions recorded in EcoCyc and KEGG. If the reaction could not be balanced, it was discarded from further analysis. Here, 34 products were

matched to measured metabolites reported in ECMDB. We identified 17 products and their balanced reactions after curation to remove stoichiometrically unbalanced reactions.

The outcomes were divided into four categories. C1 reactions consisted of metabolites predicted by PROXIMAL that are already in iML1515 but catalyzed by different enzymes than the ones already listed in the model. These reactions reflect promiscuous activity that enabled the same biotransformation catalyzed by a different gene in the model. C2 reactions already existed in EcoCyc and/or KEGG but not in iML1515. This reflected a curation problem where some reactions were not included in the iML1515 model. C3 reactions were not in EcoCyc but documented in KEGG for other organisms. C4 reactions did not exist in either databases EcoCyc nor in KEGG. These reactions were thus novel reactions that have not been reported in the literature.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The *E. coli* iML1615 model is available on the BiGG database and can be found at http://bigg.ucsd.edu/models/iML1515. Data from ECMDB can be directly downloaded from the

ECMDB website. The EMMA workflow will be made available at https://github.com/HassounLab/ (once published). In addition, a full list of derivatives predicted by PROXIMAL can be found in Supplementary File 1.

Competing interests

Not applicable

Funding

This work is funded under NSF grant #1421972 and NIH grants 1DP2HD091798 and 1R03CA211839-01.

Author's contribution

SH conceived the EMMA concept. SA developed the EMMA workflow. EC curated the results. NN and SH supervised the work done through the development of the workflow and data curation. Manuscript was written by SA and EC, and revised by NN and SH.

Acknowledgments

Not applicable

Figures

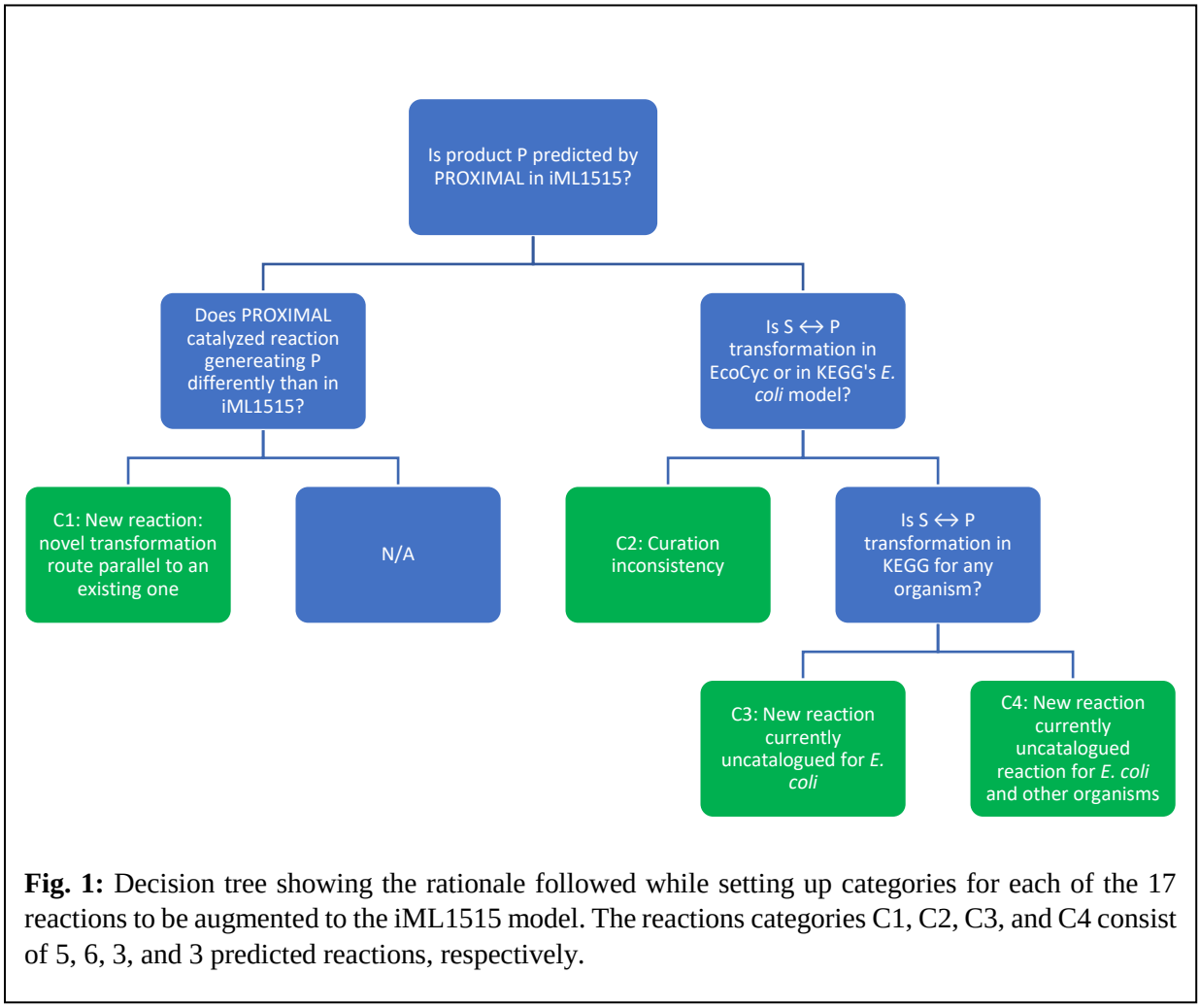


Fig. 1: Decision tree showing the rationale followed while setting up categories for each of the 17 reactions to be augmented to the iML1515 model. The reactions categories C1, C2, C3, and C4 consist of 5, 6, 3, and 3 predicted reactions, respectively.

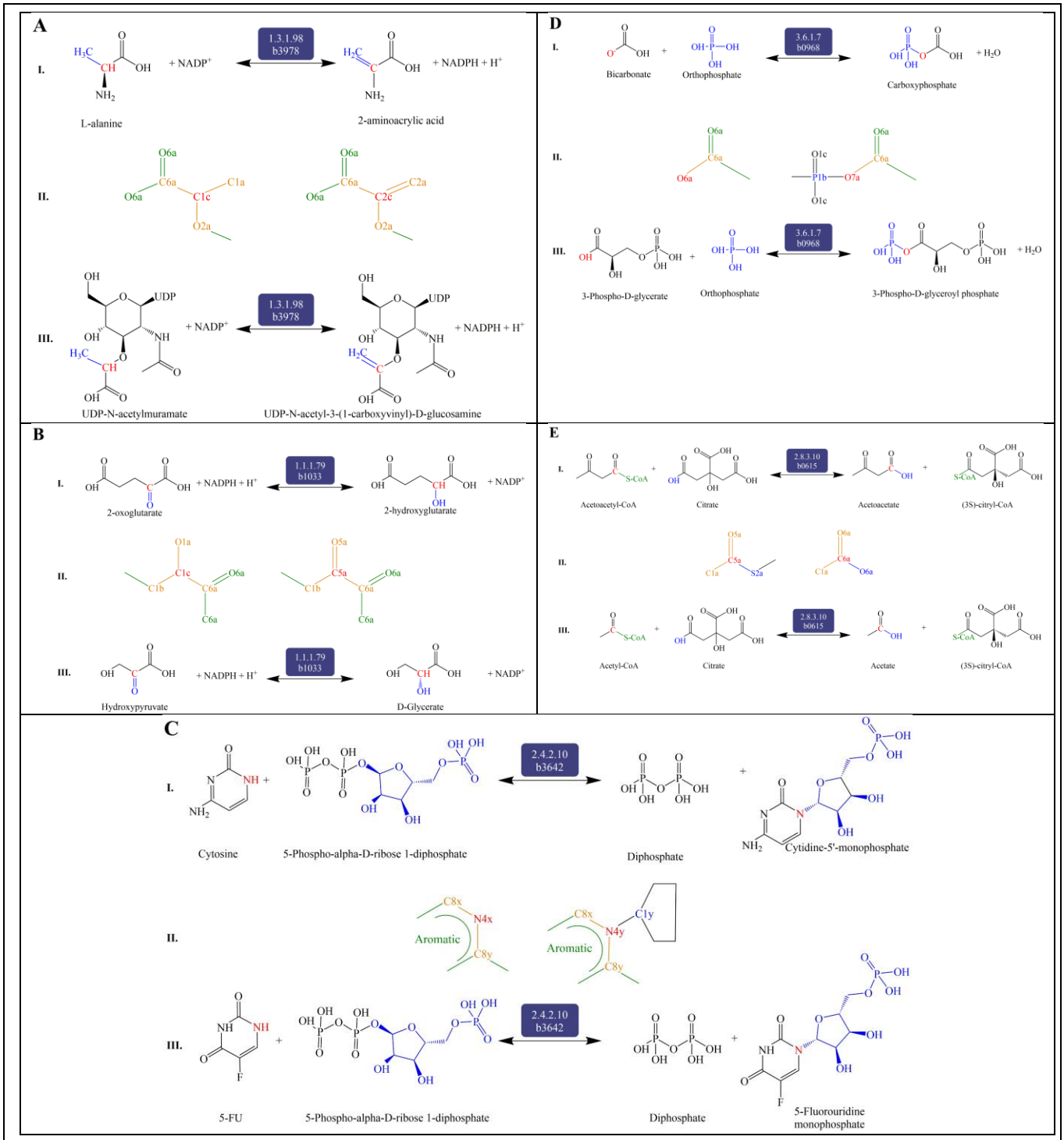


Fig. 2: The set of five reactions belonging to Category 1 (C1). Reactions in C1 are predicted to be catalyzed by enzymes different than those in iML1515. Each of the five panels is divided into three sections I) the balanced reaction developed by our workflow indicating the reactants, products, and the promiscuous enzyme, II) the RDM pattern showing the Reaction Center (R) in red where the biotransformation occurs, and III) the native reaction catalyzed by the potentially promiscuous enzyme, as catalogued in KEGG.

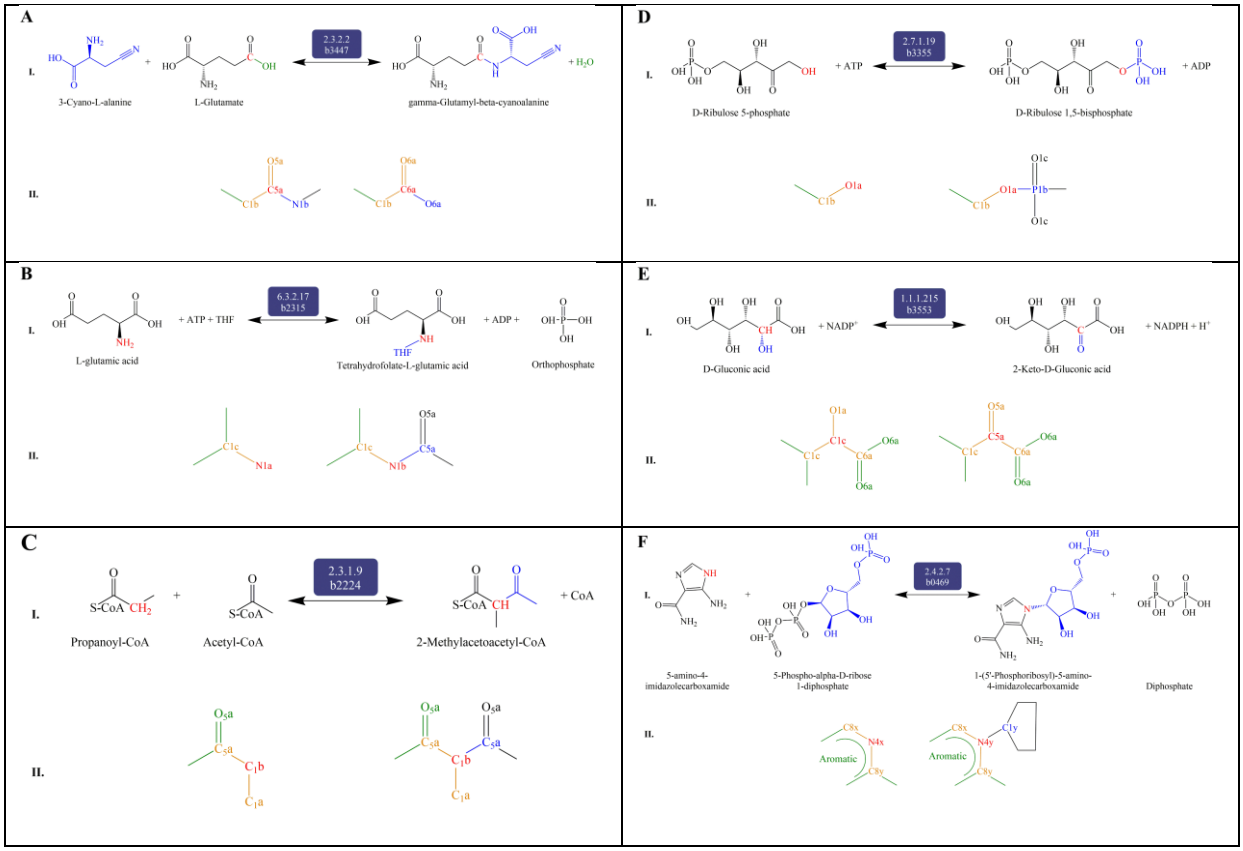


Fig. 3: The set of six reactions belonging to Category 2 (C2). Reactions in C2 are associated with derivatives not present in iML1515 but are associated with *E. coli* in KEGG and/or EcoCyc. Each of the six panels is divided into two sections I) the balanced reaction developed by our workflow, that is also documented in KEGG, indicating the reactants, products, and the promiscuous enzyme n, and II) the RDM pattern showing the Reaction Center (R) in red where the biotransformation occurs.

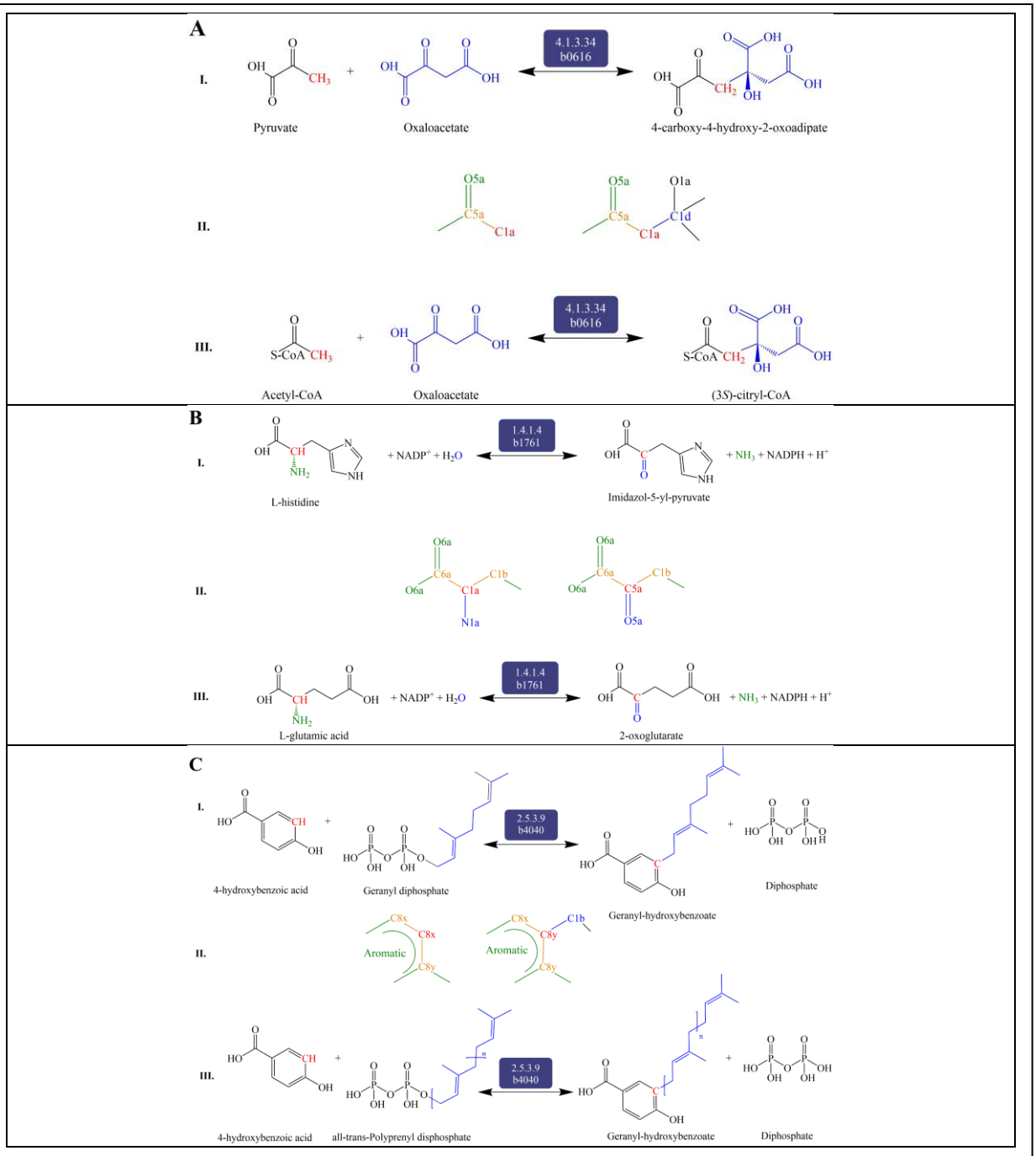


Fig. 4: The set of three reactions belonging to Category 3 (C3). C3 reactions and derivatives are neither present in iML1515 nor associated with *E. coli* in KEGG and EcoCyc. However, according to KEGG, the reactions occur in other organisms. Each of the three panels is divided into three sections I) the balanced reaction developed by our workflow indicating the reactants, products, and the promiscuous enzyme, II) the RDM pattern showing the Reaction Center (R) in red, and III) the native reaction catalyzed by the potentially promiscuous enzyme, as catalogued in KEGG.

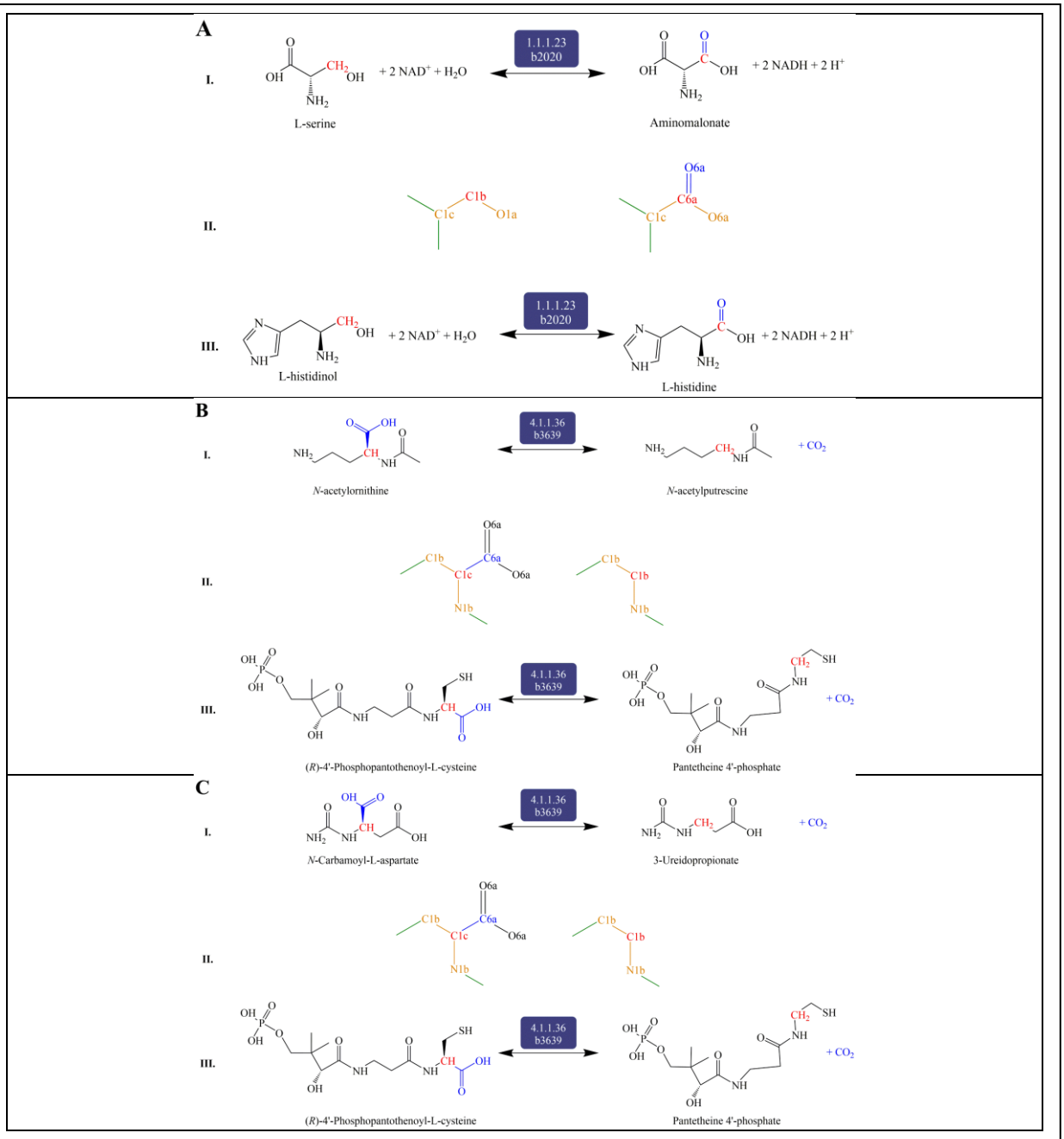


Fig. 5: The set of three reactions belonging to Category 4 (C4). C4 reactions and derivatives are neither present in iML1515 nor associated with any other organism in KEGG or EcoCyc. Each of the three panels is divided into three sections I) the balanced reaction developed by our workflow indicating the reactants, products, and the promiscuous enzyme, II) the RDM pattern showing the Reaction Center (R) in red, and III) the native reaction catalyzed by the potentially promiscuous enzyme, as catalogued in KEGG.

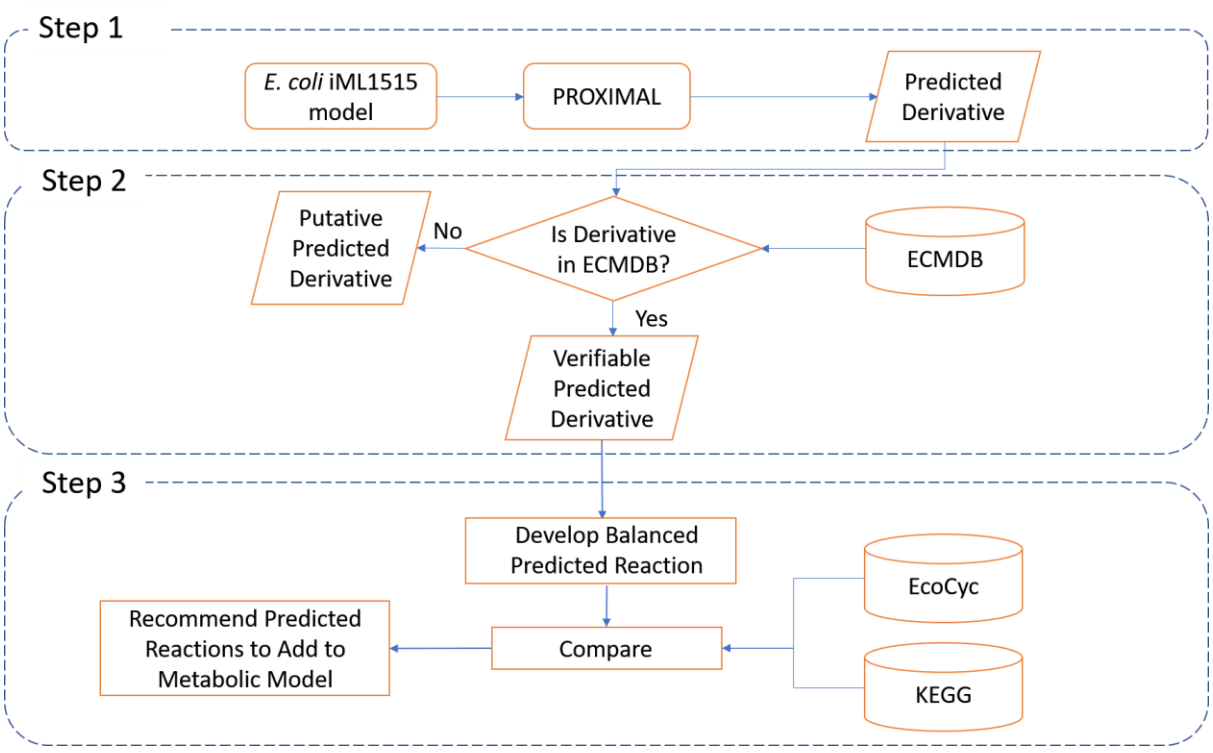


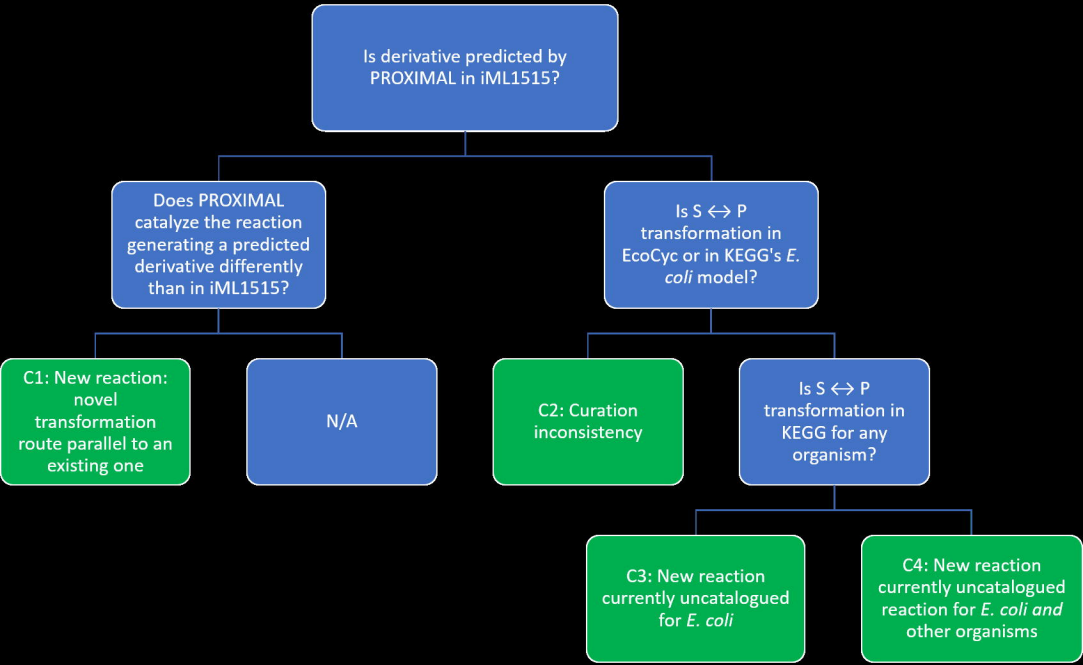
Fig. 6: Main steps of EMM-B workflow customized to extend the *E. coli* iML1515 model with predicted reactions. Step 1: Predict promiscuous reactions and derivatives using PROXIMAL. Step 2: Compare derivatives with measured metabolic dataset(s). Step 3: Curation and stoichiometric balancing of reactions.

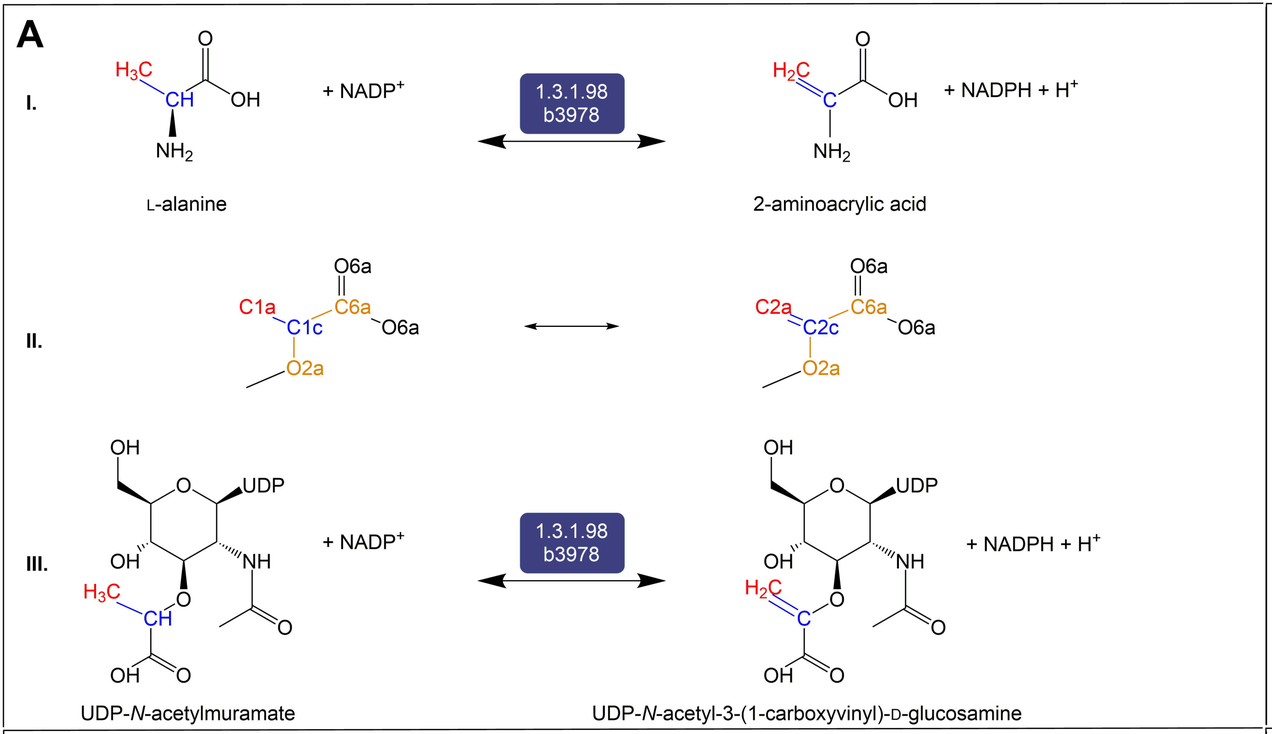
References

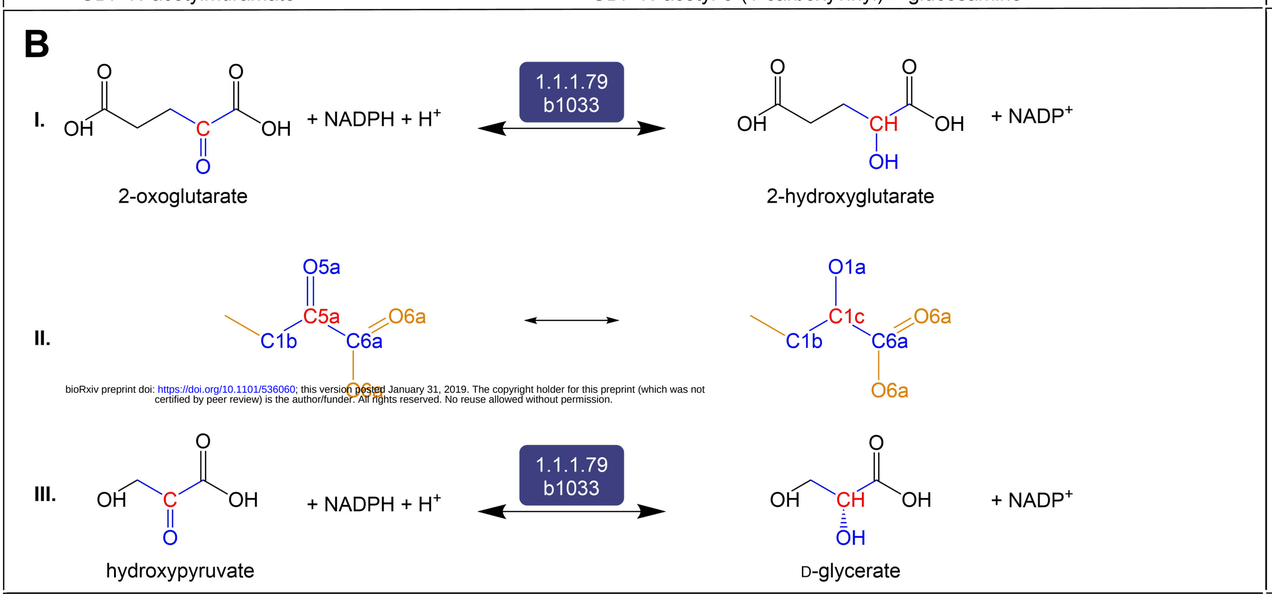
1. Lee SK, Chou H, Ham TS, Lee TS, Keasling JD. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. Current opinion in biotechnology. 2008;19(6):556-63.
2. Trantas EA, Koffas MA, Xu P, Ververidis F. When plants produce not enough or at all: metabolic engineering of flavonoids in microbial hosts. Frontiers in plant science. 2015;6:7.
3. George KW, Alonso-Gutierrez J, Keasling JD, Lee TS. Isoprenoid drugs, biofuels, and chemicals—artemisinin, farnesene, and beyond. Biotechnology of Isoprenoids: Springer; 2015. p. 355-89.
4. Du J, Shao Z, Zhao H. Engineering microbial factories for synthesis of value-added products. Journal of industrial microbiology & biotechnology. 2011;38(8):873-90.
5. Furusawa C, Horinouchi T, Hirasawa T, Shimizu H. Systems metabolic engineering: the creation of microbial cell factories by rational metabolic design and evolution. Future Trends in Biotechnology: Springer; 2012. p. 1-23.
6. Davy AM, Kildegaard HF, Andersen MR. Cell factory engineering. Cell Systems. 2017;4(3):262-75.
7. Lee S, Mattanovich D, Villaverde A. Systems metabolic engineering, industrial biotechnology and microbial cell factories. BioMed Central; 2012.
8. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnology and bioengineering. 2003;84(6):647-57.
9. Ranganathan S, Suthers PF, Maranas CD. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS computational biology. 2010;6(4):e1000744.
10. Yousofshahi M, Lee K, Hassoun S. Probabilistic pathway construction. Metabolic engineering. 2011;13(4):435-44.
11. Wu G, Yan Q, Jones JA, Tang YJ, Fong SS, Koffas MA. Metabolic burden: cornerstones in synthetic biology and metabolic engineering applications. Trends in biotechnology. 2016;34(8):652-64.
12. Gerstl MP, Ruckerbauer DE, Mattanovich D, Jungreuthmayer C, Zanghellini J. Metabolomics integrated elementary flux mode analysis in large metabolic networks. Scientific reports. 2015;5:8930.
13. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY. Recent advances in reconstruction and applications of genome-scale metabolic models. Current opinion in biotechnology. 2012;23(4):617-23.
14. Saha R, Chowdhury A, Maranas CD. Recent advances in the reconstruction of metabolic models and integration of omics data. Current opinion in biotechnology. 2014;29:39-45.
15. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Katta HY, Mojica A, et al. Genomes OnLine database (GOLD) v. 7: updates and new features. Nucleic Acids Research. 2018.
16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.
17. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinformatics. 2017.
18. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic acids research. 2015;44(D1):D515-D22.
19. Sorokina M, Stam M, Médigue C, Lespinet O, Vallenet D. Profiling the orphan enzymes. Biology direct. 2014;9(1):10.
20. Raushel FM. Finding homes for orphan enzymes. Perspectives in Science. 2016;9:3-7.
21. Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, et al. Network-level architecture and the evolutionary potential of underground metabolism. Proceedings of the National Academy of Sciences. 2014;111(32):11762-7.

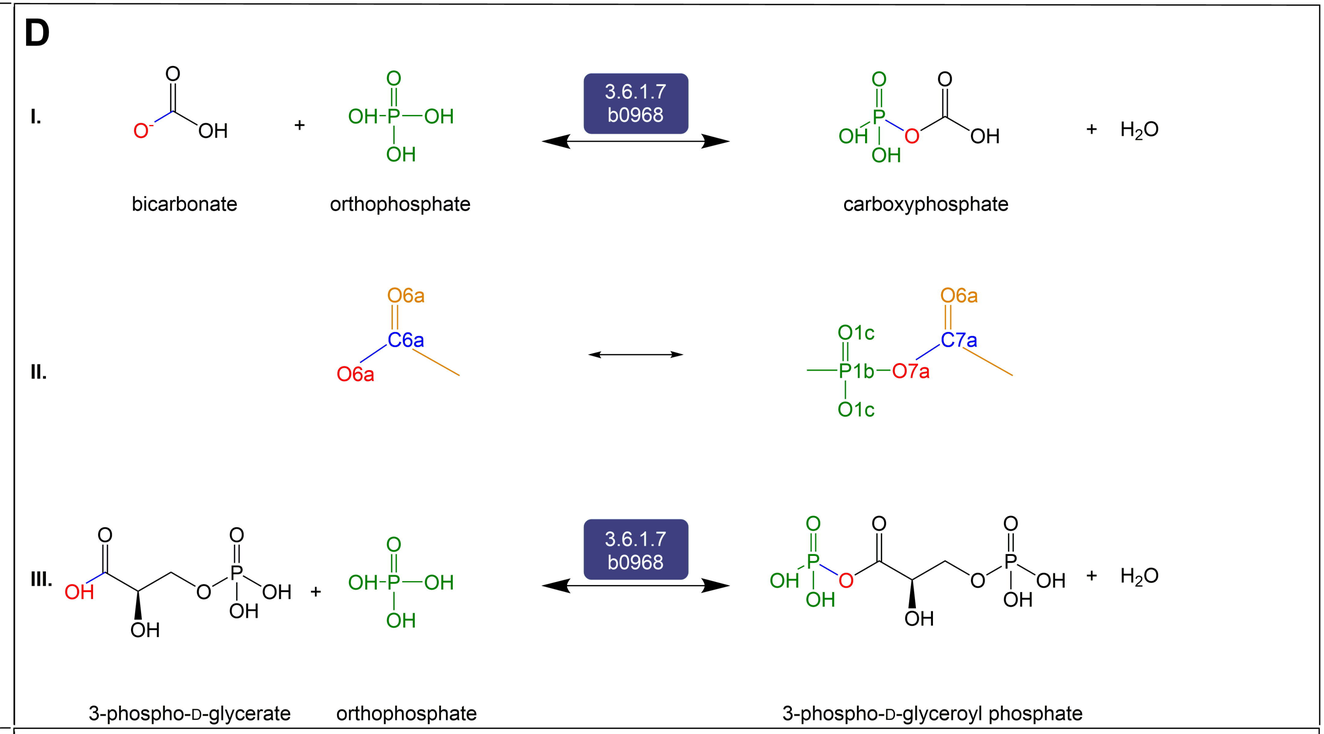
22. Notebaart RA, Kintses B, Feist AM, Papp B. Underground metabolism: network-level perspective and biotechnological potential. Current Opinion in Biotechnology. 2018;49:108-14.
23. Hult K, Berglund P. Enzyme promiscuity: mechanism and applications. Trends in biotechnology. 2007;25(5):231-8.
24. Khersonsky O, Roodveldt C, Tawfik DS. Enzyme promiscuity: evolutionary and mechanistic aspects. Current opinion in chemical biology. 2006;10(5):498-508.
25. Tawfik OK, S D. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annual review of biochemistry. 2010;79:471-505.
26. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nature biotechnology. 2009;27(2):157.
27. D'Ari R, Casadesús J. Underground metabolism. BioEssays. 1998;20(2):181-6.
28. Carbonell P, Faulon J-L. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics. 2010;26(16):2012-9.
29. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. Journal of cheminformatics. 2015;7(1):44.
30. Hadadi N, Hafner J, Shajkofci A, Zisaki A, Hatzimanikatis V. ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. ACS synthetic biology. 2016;5(10):1155-66.
31. Arora B, Mukherjee J, Gupta MN. Enzyme promiscuity: using the dark side of enzyme specificity in white biotechnology. Sustainable Chemical Processes. 2014;2(1):25.
32. Poppe L, Paizs C, Kovács K, Irimie F-D, Vértessy B. Preparation of unnatural amino acids with ammonia-lyases and 2, 3-aminomutases. Unnatural Amino Acids: Springer; 2012. p. 3-19.
33. Atsumi S, Hanai T, Liao JC. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. nature. 2008;451(7174):86.
34. Song CW, Kim JW, Cho IJ, Lee SY. Metabolic engineering of Escherichia coli for the production of 3-hydroxypropionic acid and malonic acid through β-alanine route. ACS synthetic biology. 2016;5(11):1256-63.
35. Yousofshahi M, Manteiga S, Wu C, Lee K, Hassoun S. PROXIMAL: a method for Prediction of Xenobiotic Metabolism. BMC systems biology. 2015;9(1):94.
36. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic acids research. 2010;38(suppl_2):W138-W43.
37. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes Escherichia coli traits. Nature biotechnology. 2017;35(10):904.
38. Guo AC, Jewison T, Wilson M, Liu Y, Knox C, Djoumbou Y, et al. ECMDB: the E. coli Metabolome Database. Nucleic acids research. 2012;41(D1):D625-D30.
39. Sajed T, Marcu A, Ramirez M, Pon A, Guo AC, Knox C, et al. ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. Nucleic acids research. 2015;44(D1):D495-D501.
40. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, et al. EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic acids research. 2005;33(suppl_1):D334-D7.
41. Tepper N, Noor E, Amador-Noguez D, Haraldsdóttir HS, Milo R, Rabinowitz J, et al. Steady-state metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load. PloS one. 2013;8(9):e75370.
42. Quinlan JR. Simplifying decision trees. International journal of man-machine studies. 1987;27(3):221-34.
43. Coote J, Hassall H. The role of imidazol-5-yl-lactate-nicotinamide-adenine dinucleotide phosphate oxidoreductase and histidine-2-oxoglutarate aminotransferase in the degradation of imidazol-5-yl-lactate by Pseudomonas acidovorans. Biochemical Journal. 1969;111(2):237.

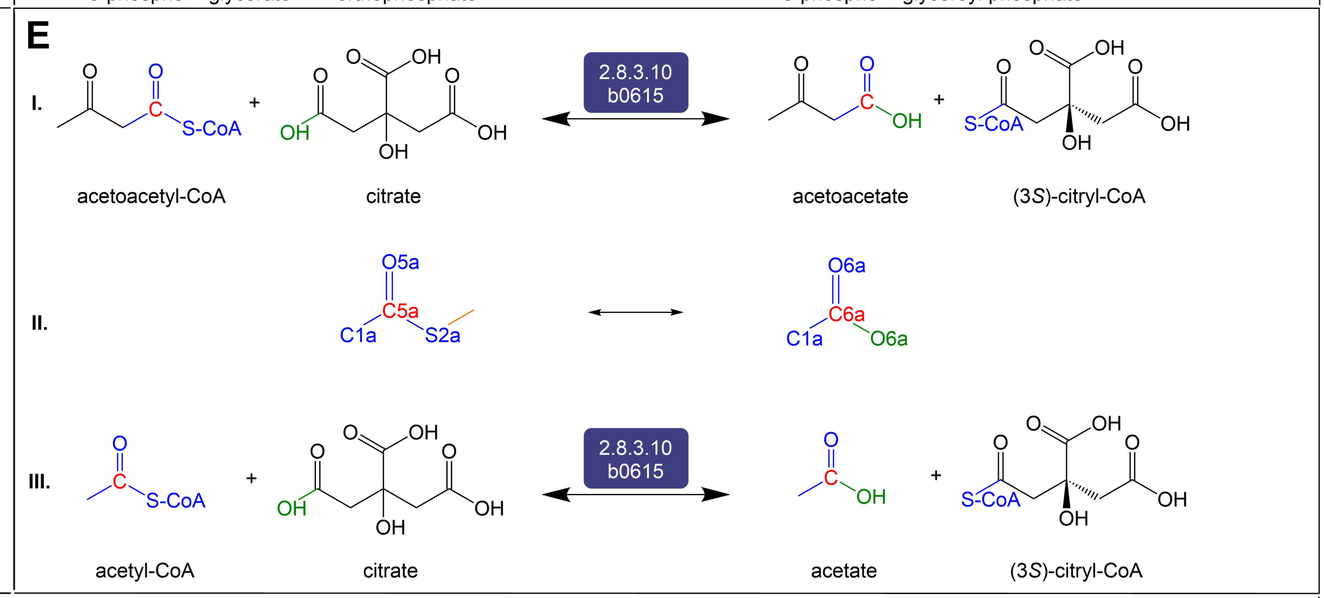
44. Mühlenweg A, Melzer M, Li S-M, Heide L. 4-Hydroxybenzoate 3-geranyltransferase from Lithospermum erythrorhizon: purification of a plant membrane-bound prenyltransferase. Planta. 1998;205(3):407-13.
45. Nagayama H, Muramatsu M, Shimura K. Enzymatic formation of aminomalonic acid from ketomalonic acid. Nature. 1958;181(4606):417.
46. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular systems biology. 2007;3(1):121.
47. Pertusi DA, Moura ME, Jeffryes JG, Prabhu S, Biggs BW, Tyo KE. Predicting novel substrates for enzymes with minimal experimental effort with active learning. Metabolic engineering. 2017;44:171-81.
48. Yun EJ, Oh EJ, Liu J-J, Yu S, Kim DH, Kwak S, et al. Promiscuous activities of heterologous enzymes lead to unintended metabolic rerouting in Saccharomyces cerevisiae engineered to assimilate various sugars from renewable biomass. Biotechnology for biofuels. 2018;11(1):140.
49. Hassanpour N. Computational Methods to Advance Directed Evolution of Enzymes and Metabolomics Data Analysis: Tufts University; 2018.
50. Hattori M, Okuno Y, Goto S, Kanehisa M. 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. 2003;125(39):11853-65.
51. Misra RV, Horler RS, Reindl W, Goryanin II, Thomas GH. Echo BASE: an integrated post-genomic database for Escherichia coli. Nucleic acids research. 2005;33(suppl_1):D329-D33.
52. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the universal protein knowledgebase. Nucleic acids research. 2004;32(suppl_1):D115-D9.
53. Consortium U. UniProt: a hub for protein information. Nucleic acids research. 2014;43(D1):D204-D12.
54. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, Lee J, et al. YMDB: the yeast metabolome database. Nucleic acids research. 2011;40(D1):D815-D20.
55. Sundararaj S, Guo A, Habibi-Nazhad B, Rouani M, Stothard P, Ellison M, et al. The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. Nucleic acids research. 2004;32(suppl_1):D293-D5.
56. O'Boyle NM, Morley C, Hutchison GR. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chemistry Central Journal. 2008;2(1):5.
57. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of cheminformatics. 2011;3(1):33.

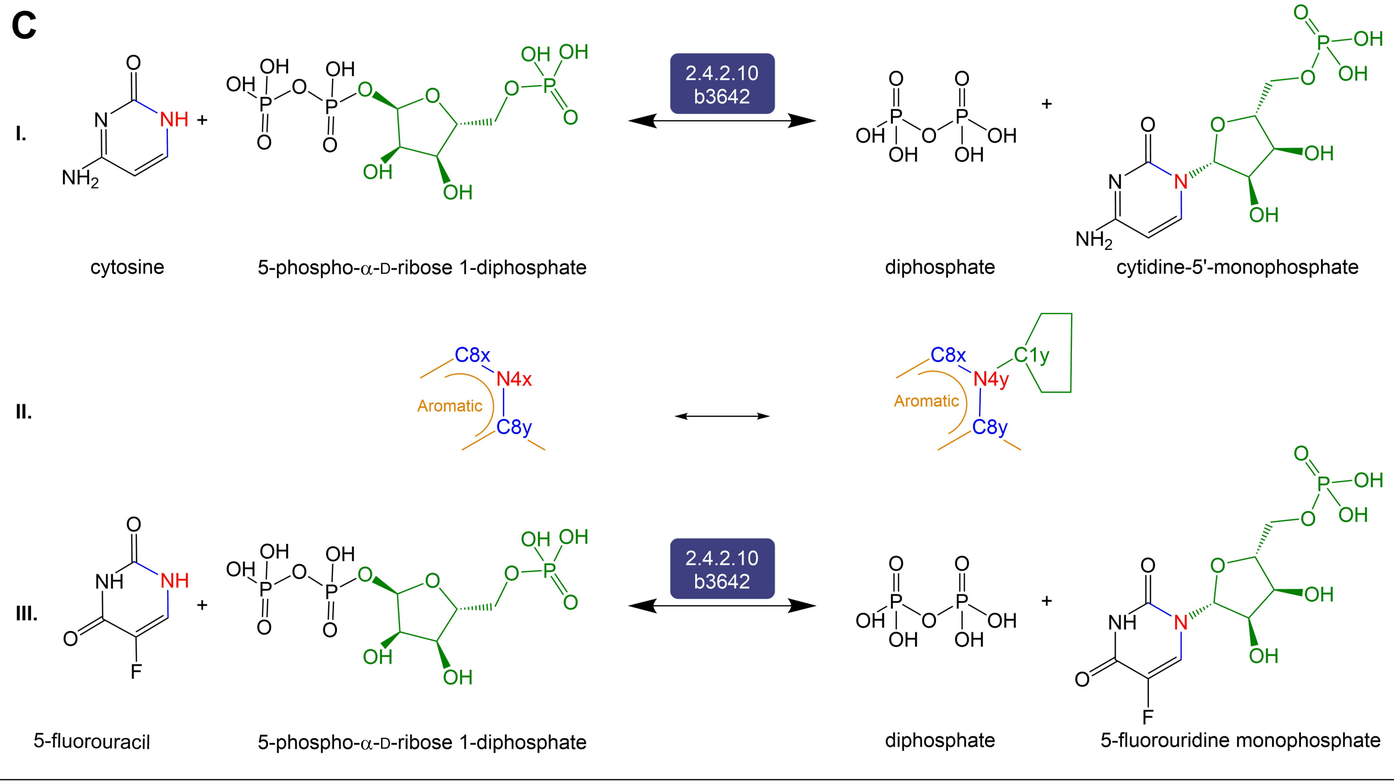

Is derivative predicted by PROXIMAL in iML1515?
Does PROXIMAL catalyze the reaction generating a predicted derivative differently than in iML1515?
Is S ↔ P transformation in EcoCyc or in KEGG's *E. coli* model?
C1: New reaction: novel transformation route parallel to an existing one
N/A
C2: Curation inconsistency
Is S ↔ P transformation in KEGG for any organism?
C3: New reaction currently uncatalogued for *E. coli*
C4: New reaction currently uncatalogued reaction for *E. coli and* other organisms





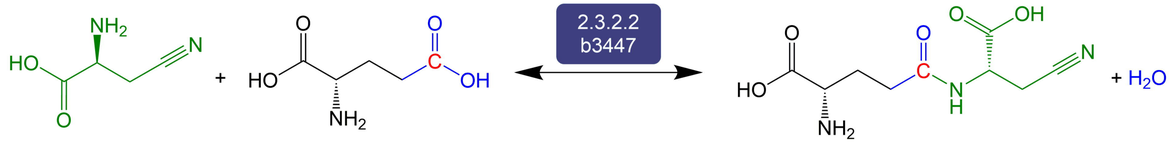






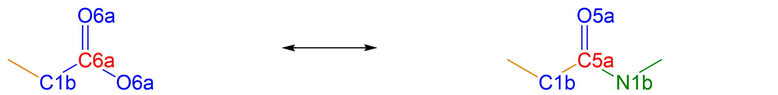
A

I.



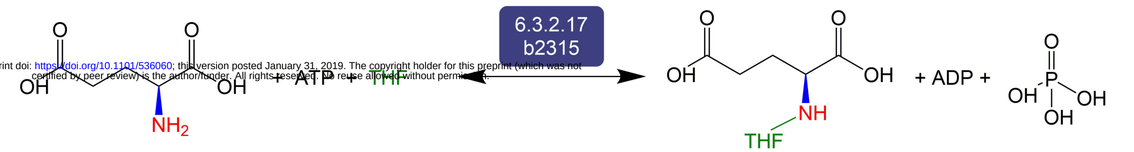
3-cyano-L-alanine + L-glutamate ⟷ γ-glutamyl-β-cyanoalanine

II.



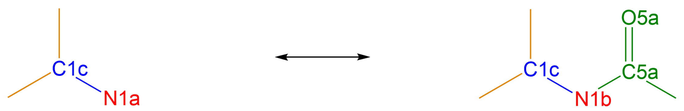
B

I.



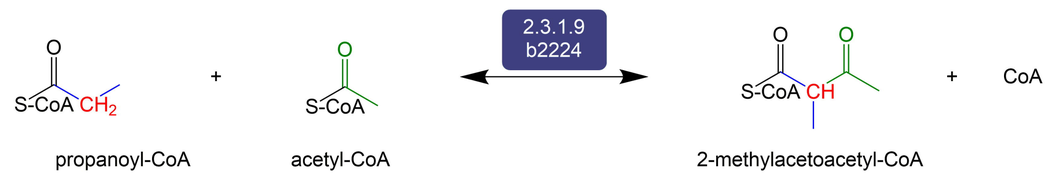
L-glutamic acid + ATP + THF ⟷ tetrahydrofolate-L-glutamic acid + ADP + orthophosphate

II.



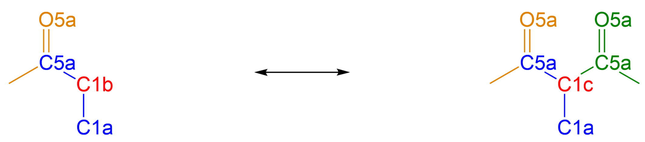
C

I.



propanoyl-CoA + acetyl-CoA ⟷ 2-methylacetoacetyl-CoA + CoA

II.

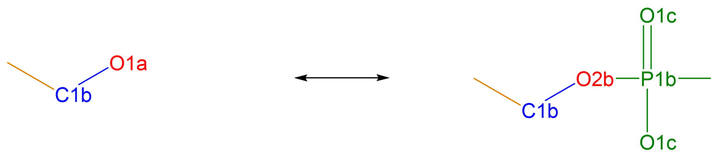


D

I.

D-ribulose 5-phosphate + ATP ⟷ D-ribulose 1,5-bisphosphate + ADP

II.

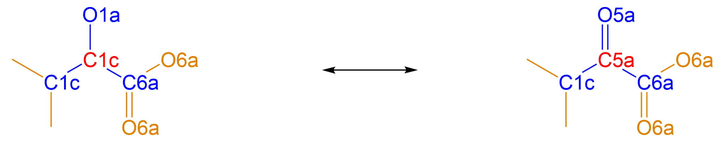


E

I.

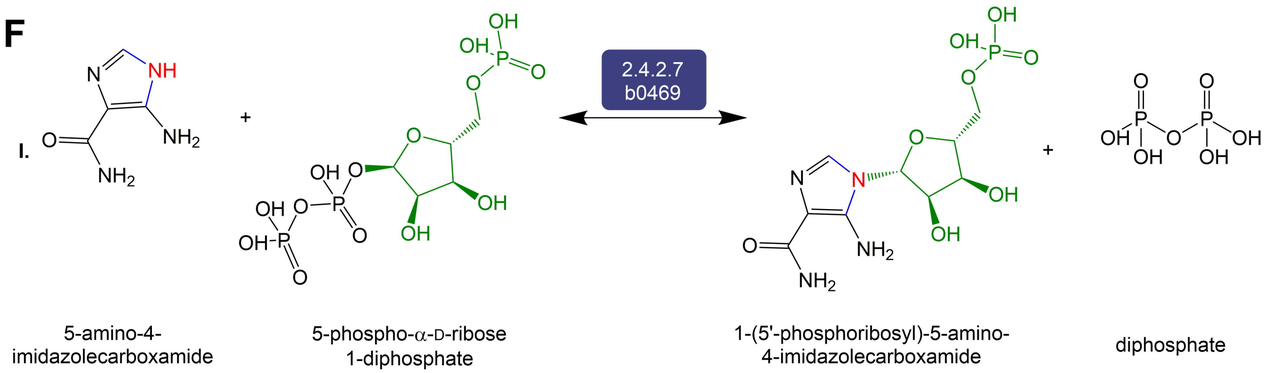
D-gluconic acid + $NADP^+$ ⟷ 2-keto-D-gluconic acid + NADPH + H^+

II.



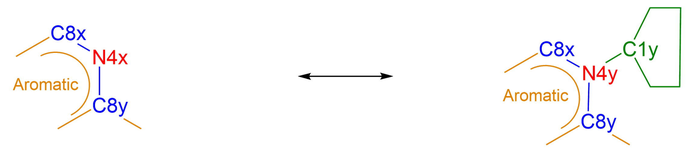
F

I.



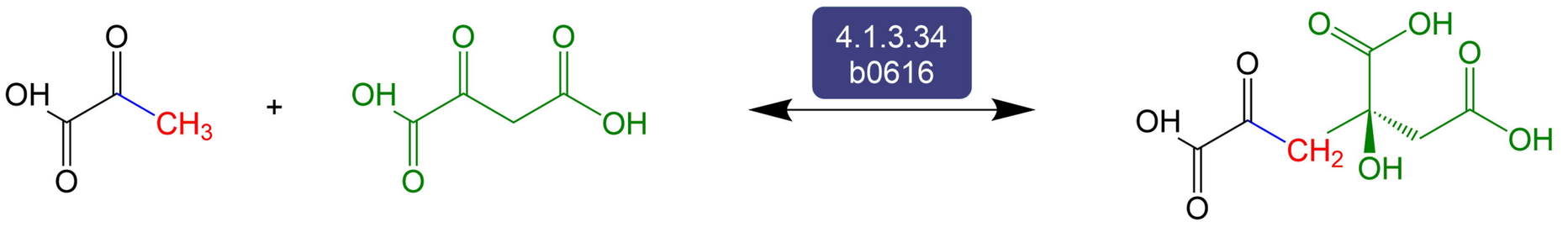
5-amino-4-imidazolecarboxamide + 5-phospho-α-D-ribose 1-diphosphate ⟷ 1-(5'-phosphoribosyl)-5-amino-4-imidazolecarboxamide + diphosphate

II.



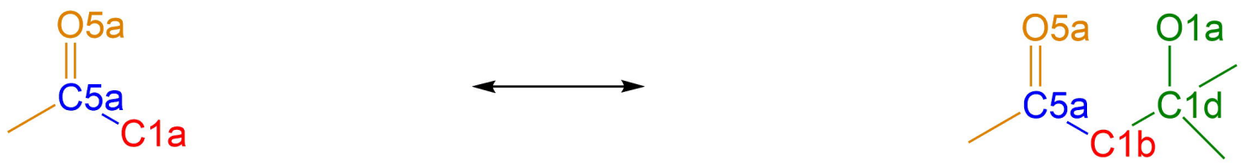
A

I.

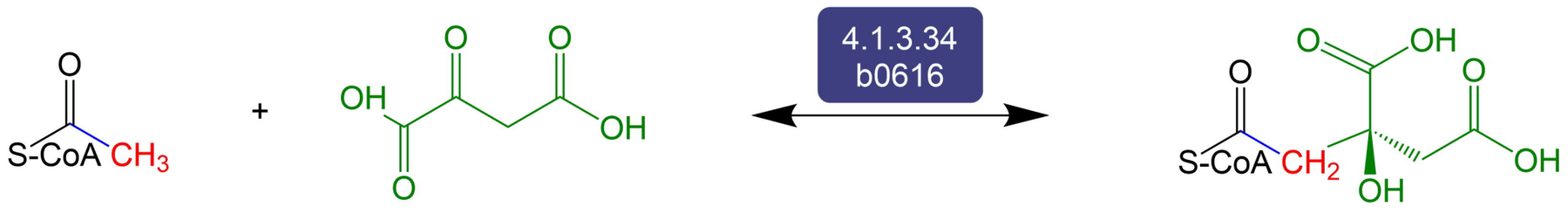


pyruvate + oxaloacetate ⟷ 4-carboxy-4-hydroxy-2-oxoadipate

II.



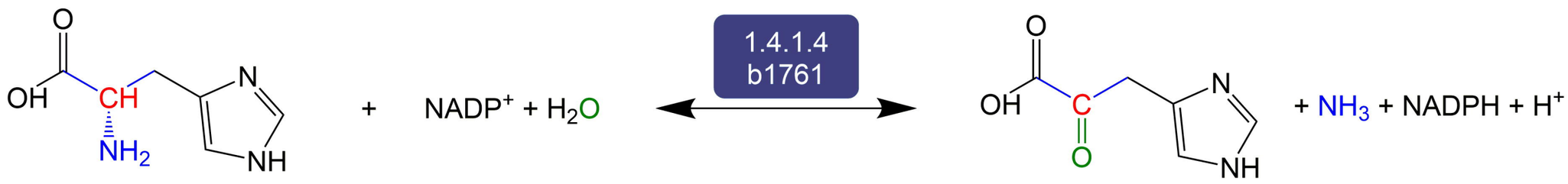
III.



acetyl-CoA + oxaloacetate ⟷ (3*S*)-citryl-CoA

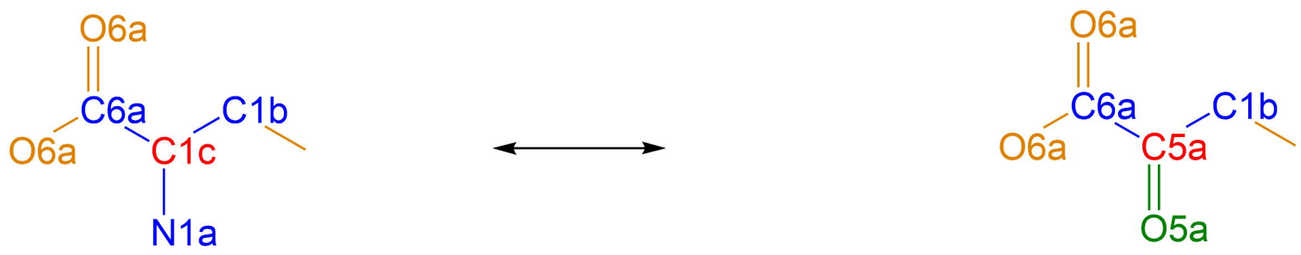
B

I.



L-histidine + $NADP^+ + H_2O$ ⟷ imidazol-5-yl-pyruvate + NH_3 + NADPH + H^+

II.

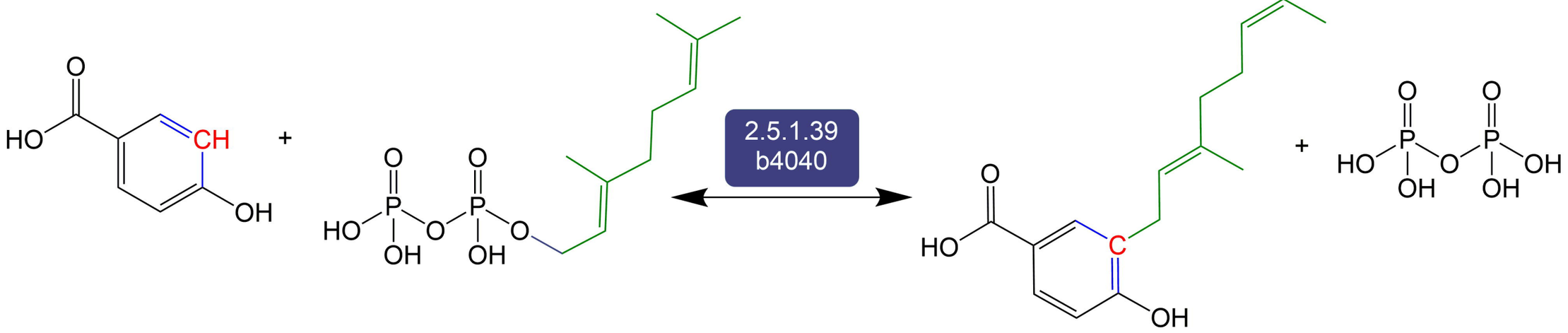


III.

L-glutamic acid + $NADP^+ + H_2O$ ⟷ 2-oxoglutarate + NH_3 + NADPH + H^+

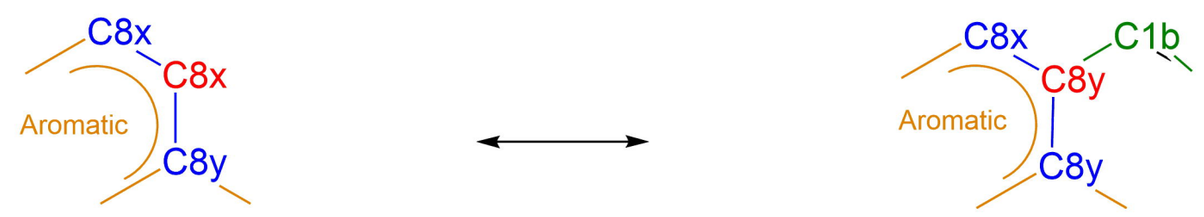
C

I.

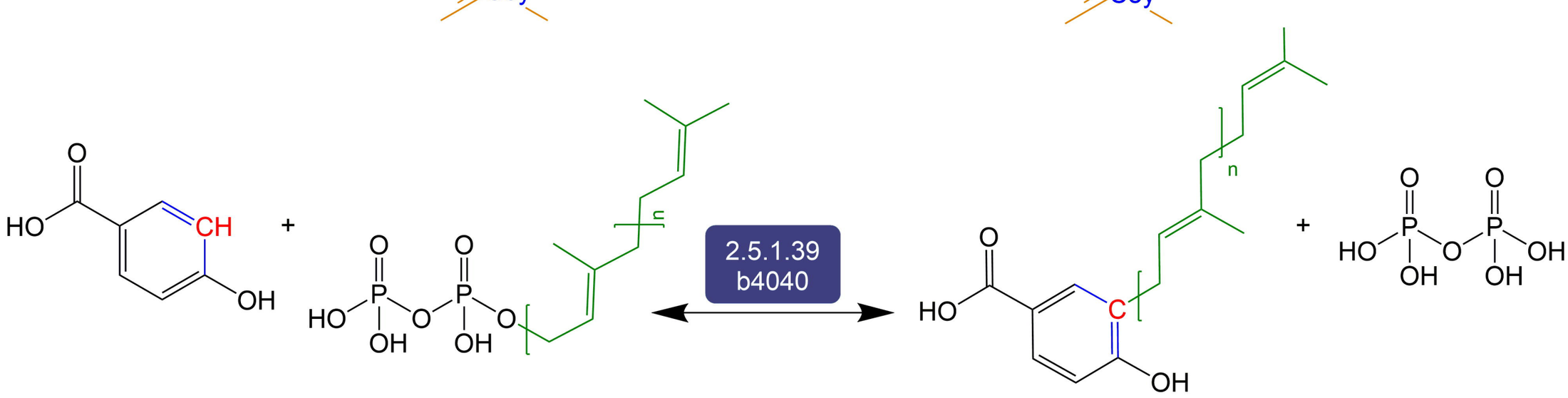


4-hydroxybenzoic acid + geranyl diphosphate ⟷ geranyl-hydroxybenzoate + diphosphate

II.



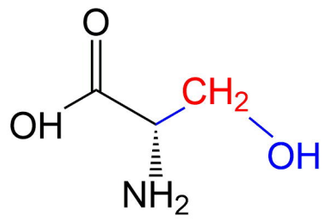
III.



4-hydroxybenzoic acid + all-*trans*-polyprenyl disphosphate ⟷ 4-hydroxy-3-polyprenylbenzoate + diphosphate

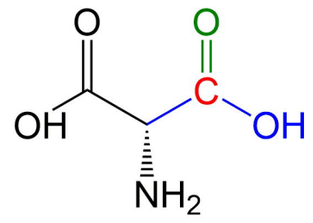
A

I.



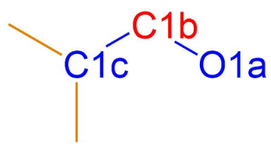
L-serine + 2 NAD^+ + H_2O

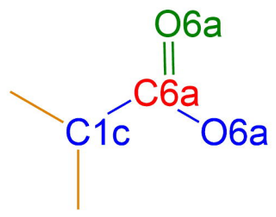
1.1.1.23
b2020



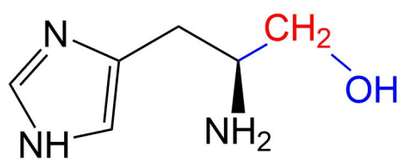
aminomalonate + 2 NADH + 2 H^+

II.



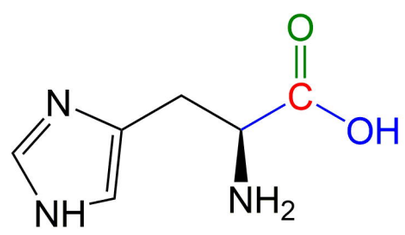


III.



L-histidinol + 2 NAD^+ + H_2O

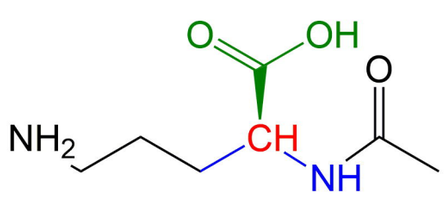
1.1.1.23
b2020



L-histidine + 2 NADH + 2 H^+

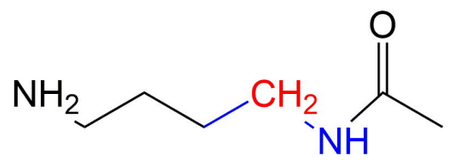
B

I.



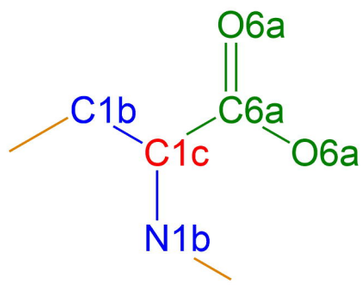
N-acetylornithine

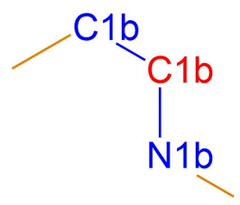
4.1.1.36
b3639



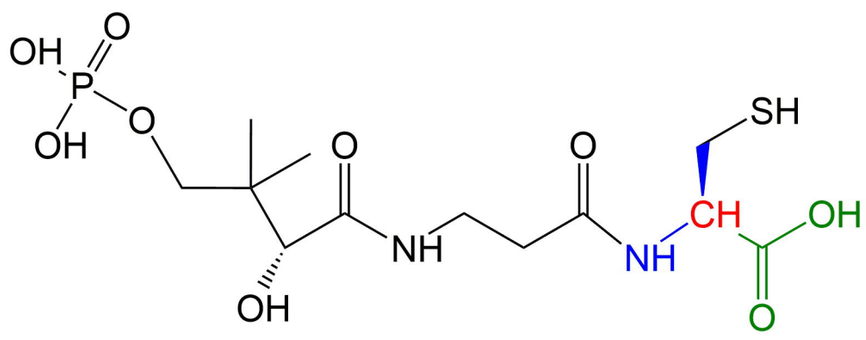
N-acetylputrescine + CO_2

II.



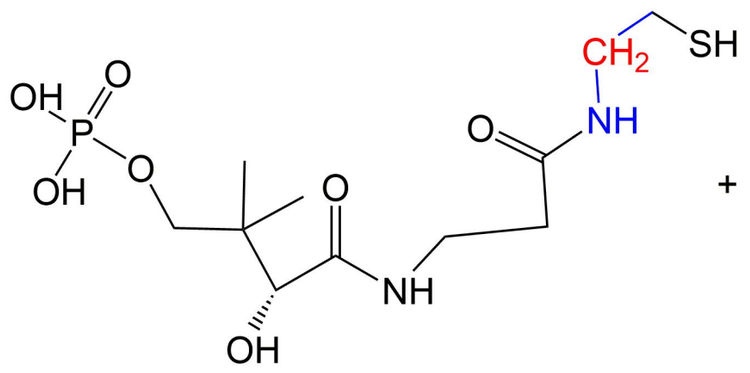


III.



(*R*)-4'-phosphopantothenoyl-L-cysteine

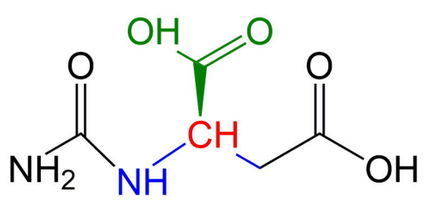
4.1.1.36
b3639



pantetheine 4'-phosphate + CO_2

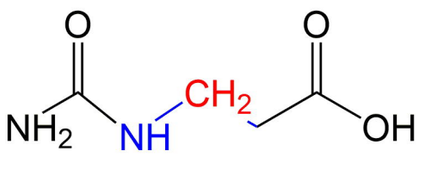
C

I.



N-carbamoyl-L-aspartate

4.1.1.36
b3639



3-ureidopropionate + CO_2

II.

III.

(*R*)-4'-phosphopantothenoyl-L-cysteine

4.1.1.36
b3639

pantetheine 4'-phosphate + CO_2

Step 1

E. coli iML1515 model

PROXIMAL

Predicted Derivative

Step 2

Is Derivative in ECMDB?

No

Putative Predicted Derivative

ECMDB

Yes

Verifiable Predicted Derivative

Step 3

Develop Balanced Predicted Reaction

Compare

Recommend Predicted Reactions to Add to Metabolic Model

EcoCyc

KEGG