Microbial Cell Factories

RESEARCH

Open Access

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

Sara A. Amin[1†], Elizabeth Chavez[2†], Vladimir Porokhin[1], Nikhil U. Nair[3*] and Soha Hassoun[1,3*]

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 *Escherichia 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 may have not been documented in iML1515 or other databases. We provide detailed analysis of 23 predicted reactions and 16 associated metabolites. Interestingly, nine of these metabolites, which are in ECMDB, have not previously been documented in any other *E. coli* databases. Four of the predicted reactions provide putative transformations parallel to those already in iML1515. We suggest adding predicted metabolites and reactions to iML1515 to create an extended metabolic model (EMM) for *E. coli*.

Keywords: Metabolic engineering, Enzyme promiscuity, Extended metabolic model, Systems biology, Enzyme activity prediction

*Correspondence: nikhil.nair@tufts.edu; soha.hassoun@tufts.edu
†Sara A. Amin and Elizabeth Chavez contributed equally to this work.
[3] Department of Chemical and Biological Engineering, Tufts University, Medford, MA, USA
Full list of author information is available at the end of the article

Background

The engineering of metabolic networks has enabled the production of high-volume commodity chemicals such as biopolymers and fuels, therapeutics, and specialty products [1–5]. Producing such compounds requires transforming microorganisms into efficient cellular factories [6–9]. 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. [10–14]). These design tools rely on organism-specific metabolic models that represent cellular reactions and their substrates and products. Model reconstruction tools [15, 16] 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 [17]. There are now databases (e.g. KEGG [18], BioCyc [19], and BiGG [20]) 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 [21]. Integrated strategies that utilize structural biology, computational biology, and molecular enzymology continue to address assigning function to orphan genes [22].

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' [23–25]. 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 [26–31]. More than two-fifths (44%) of KEGG enzymes are associated with more than one reaction [32]. 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 [33], and ATLAS [34]. Despite the current wide-spread acceptance of enzyme promiscuity, and its prominent utilization to engineer catalyzing enzymes in metabolic engineering practice [35–38], 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, and thus transformation, is assumed to be reversible unless indicated otherwise. EMMA utilizes PROXIMAL [39], a method for creating biotransformation operators from KEGG reactions IDs using RDM (Reaction Center, Difference Region, and Matched Region) patterns [40], 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 [41]. 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 [42, 43]. We identify 23 new reactions and 16 new metabolites that we recommend adding to the *E. coli* model iML1515. Four of these reactions have not been catalogued prior for *E. coli* or other organisms, suggesting novel undocumented promiscuous transformations, while five other reactions are catalogued for species other than *E. coli.* Further, there were ten reactions that were cataloged in other *E. coli* databases (e.g.

EcoCyc [44], and KEGG), but not in iML1515. These 19 reactions led to the addition of the 16 metabolites that are new to iML1515. Additionally, there were four new reactions that present putative transformation routes that are in parallel to existing reactions in *E. coli.* No new metabolites are added due to these four reactions.

Results

The application of PROXIMAL to iML1515 yielded a lookup table with 1875 biotransformation operator entries. The operators were applied on two sets of metabolites. One set consisted of 106 iML1515 metabolites with predicted or measured concentration values above 1 μM [45]. We focused on these metabolites as the assumption is that high concentration metabolites are more likely to undergo transformation by promiscuous enzymatic activity and form detectible derivatives. When applied to this set, the operators predicted the formation of 1423 known (with PubChem IDs) metabolites of which 57 were identified to exist in *E. coli* per ECMDB. After manual curation (per Step 1 in the "Methods" section), our workflow recommended 16 new metabolites and 23 reactions that can be used to augment the iML1515 model. The second set of metabolites consisted of the non-high concentration metabolites in iML1515. Our workflow predicted the formation of 3694 known (with PubChem IDs) metabolites. Out of the predicted metabolites of the second set 210 derivatives are found in ECMDB. We provide a listing of all derivatives in Additional file 1. For the remainder of the "Results" section, we focus on detailed analysis of derivative products due to high-concentration metabolites. Results of flux balance analysis and flux variability analysis for the added EMMA reactions are reported in Additional file 2.

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 [46]. 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 23 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

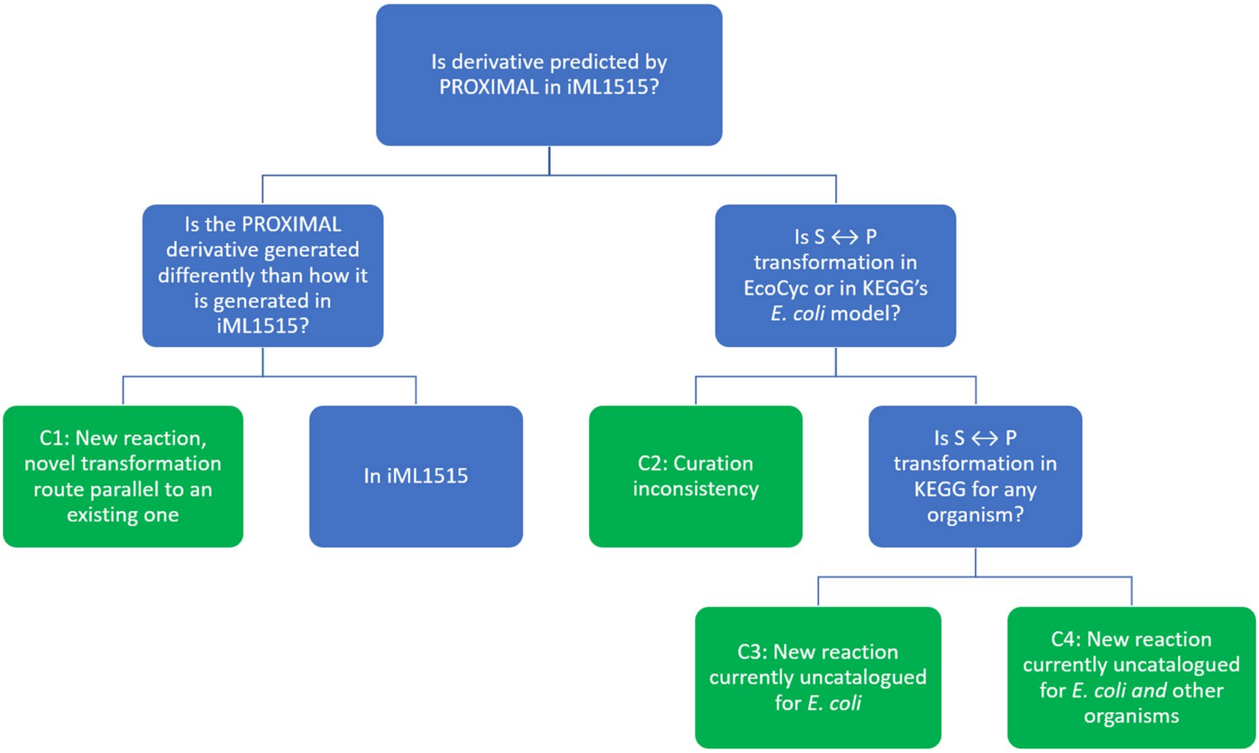
Is derivative predicted by PROXIMAL in iML1515?
Is the PROXIMAL derivative generated differently than how it is generated in iML1515?
Is S ↔ P transformation in EcoCyc or in KEGG's *E. coli* model?
C1: New reaction, novel transformation route parallel to an existing one
In iML1515
C2: Curation inconsistency
Is S ↔ P transformation in KEGG for any organism?
C3: New reaction currently uncatalogued for *E. coli*
C4: New reaction currently uncatalogued for *E. coli and* other organisms

Fig. 1 Decision tree for classifying reactions identified based on enzyme promiscuity. When analyzing the iML1515 *E. coli* model, reaction categories C1, C2, C3, and C4 had 4, 10, 5, and 4 predicted reactions, respectively

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 are 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 four reactions that are predicted to be catalyzed by enzymes that are different than those in iML1515. The details of the predicted reactions are shown in Fig. 2, and Table 1 details a comparison between those predicted reactions and their parallel reactions in iML1515. 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. 2a) and that between 2-oxoglutarate and 2-hydroxyglutarate by EC 1.1.1.79 (glyoxylate reductase) (Fig. 2b). We also predict the transformation between bicarbonate and carboxyphosphate catalyzed by EC 3.6.1.7 (acylphosphatase) (Fig. 2c). While carboxyphosphate is not in iML1515, the transformation is considered parallel to a reaction catalyzed by EC 6.3.5.5 that is documented to occur for *E. coli* in KEGG (see Fig. 3j). 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. 2d).

C2 consists of 10 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

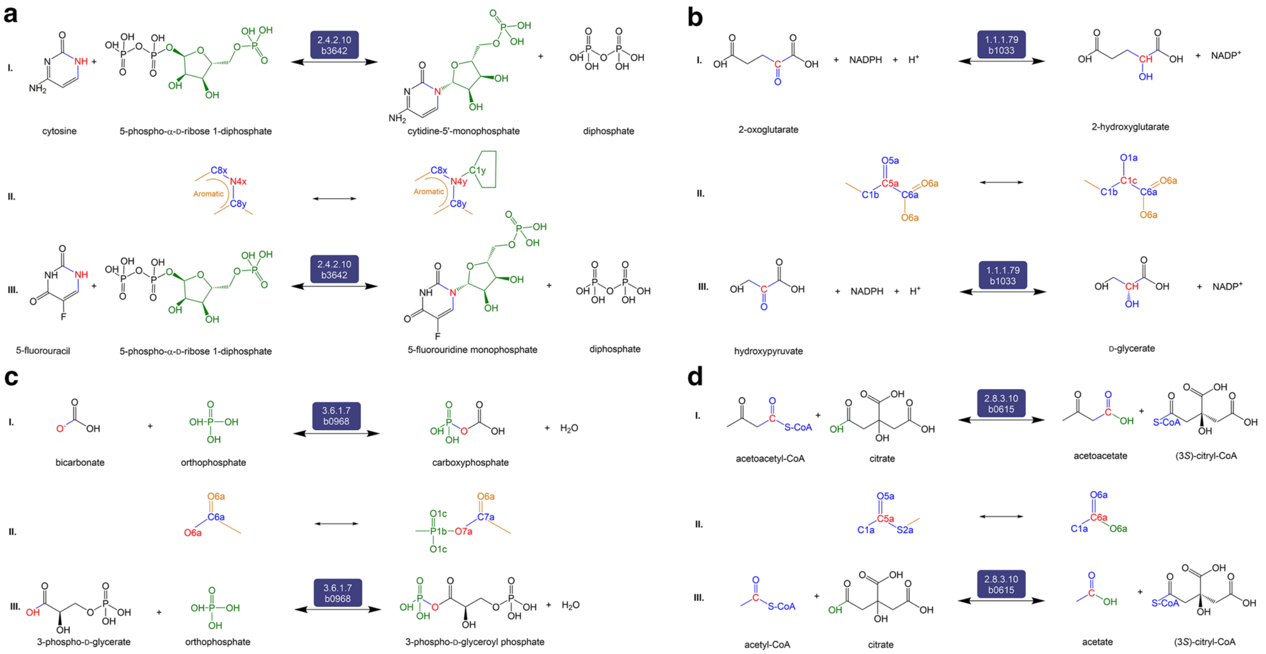

Fig. 2 The set of four reactions, panels **a**–**d**, belonging to Category 1 (C1). Reactions in C1 are predicted to be catalyzed by enzymes different than those in iML1515. Each of the four panels is divided into three sections (I) the balanced reaction developed by our workflow indicating the reactants, 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

Table 1 List of C1 reactions predicted by EMMA and their parallel reactions in *E. coli* iML1515

	EC number (gene)	Reaction
Predicted	2.4.2.10 (b3642)	Cytosine + 5-phospho-α-D-ribose-1-diphosphate ⇋ CMP + diphosphate
iML1515	3.2.2.10 (b2795)	Cytosine + D-ribose-5-phosphate ⇋ CMP + H_2O
Predicted	1.1.1.79 (b1033)	2-Oxoglutarate + NADPH + H^+ ⇋ 2-hydroxyglutarate + $NADP^+$
iML1515	1.1.1.95 (b2913)	2-Oxoglutarate + NADH + H^+ ⇋ 2-hydroxyglutarate + NAD^+
Predicted	3.6.1.7 (b0968)	Bicarbonate + orthophosphate ⇋ carboxyphosphate + H_2O
KEGG	6.3.5.5 (b0032 or b0033)	Bicarbonate + ATP ⇋ carboxyphosphate + ADP
Predicted	2.8.3.10 (b0615)	Acetoacetyl-CoA + citrate ⇋ acetoacetate + (3*S*)-citryl-CoA
iML1515	2.8.3.8 (b2221 and b2222 or b1694) or 2.8.3.9 (b2221 and b2222)	Acetoacetyl-CoA + acetate ⇋ acetoacetate + acetyl-CoA

Each of the predicted/iML1515 reaction pair occurs between the same substrate and product but utilize different co-substrate or cofactors

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). The workflow also 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). The seventh predicted reaction is the transformation between pyruvate and 4-hydroxy-2-oxoglutarate by EC 4.1.3.24 (Fig. 3g). The eighth reaction is catalyzed by EC 2.4.2.10 to transform guanine to/from GMP (Fig. 3h). Also, PROXIMAL predicted the transformation between glycerate and tartrate by EC 4.1.1.73 (Fig. 3i). Lastly, bicarbonate is transformed to/from carboxyphosphate by EC 3.6.1.7 (Fig. 3j).

C3 consists of five 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) [47]. C3 also 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) [48]. The fourth predicted reaction is the redox transformation between D-alanine and 2-aminoacrylic acid (Fig. 4d). This reaction is predicted to be catalyzed by EC 1.3.1.98 (UDP-*N*-acetylmuramate dehydrogenase). While 2-aminoacrylic acid is not known to be produced in *E. coli* in any other way, the transformation between D-alanine and 2-aminoacrylic acid occurs in other organisms such as *Staphylococcus aureus* [49]. Lastly, our workflow predicts the transformation between phenylpyruvate and phenyllactate by EC 1.1.1.100 (Fig. 4e). This transformation is known to occur in plants by EC 1.1.1.237 [50].

C4 consists of four 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* [51]. 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

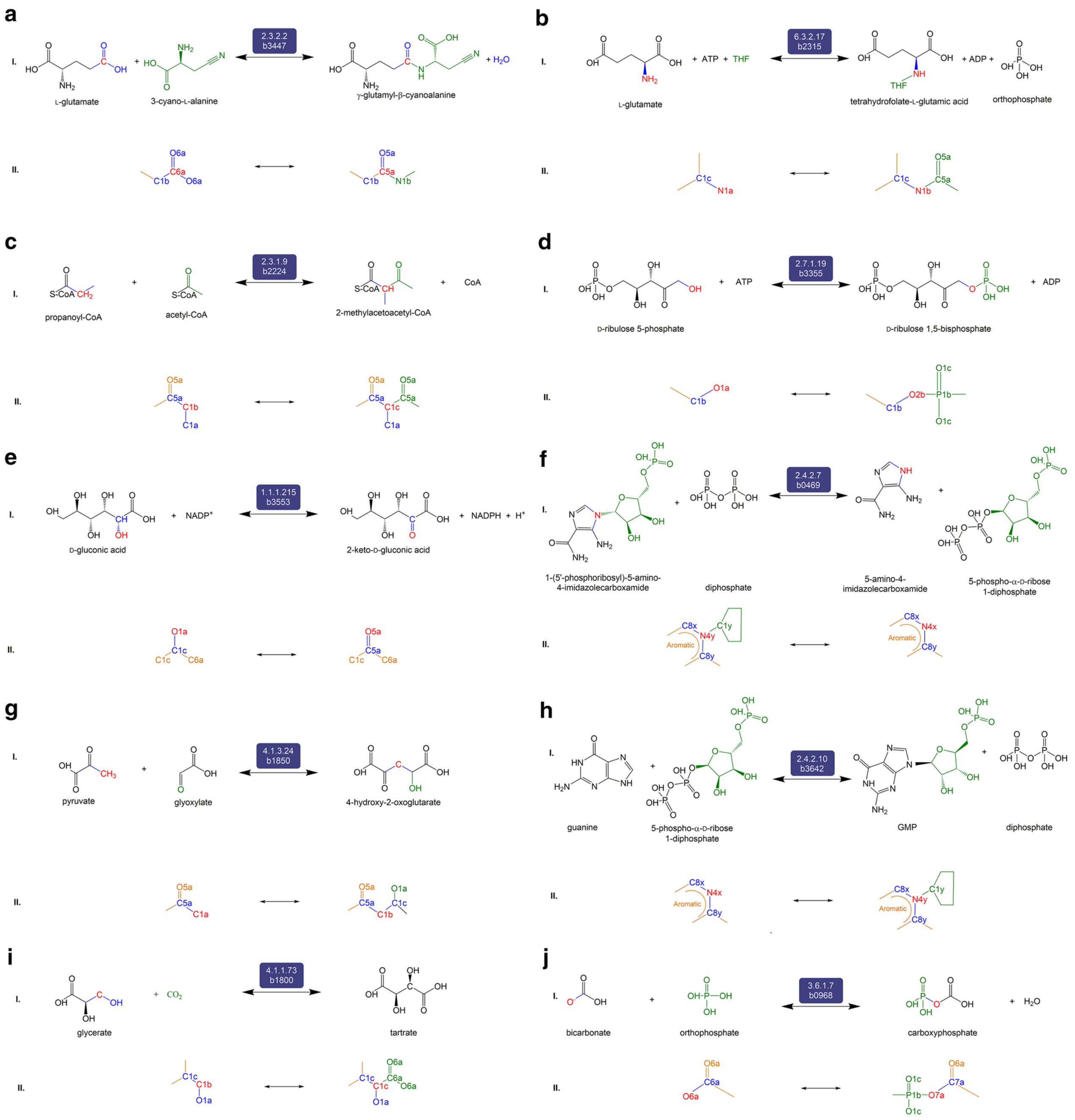

Fig. 3 The set of ten reactions, panels **a**–**j**, belonging to Category 2 (C2). Reactions in C2 are associated with derivatives not present in iML1515 but are associated with *E. coli* in KEGG and/or EcoCyc. Each of the ten panels is divided into two sections (I) the balanced reaction developed by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

[16]. The third 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). The last reaction is the transformation between D-gluconic acid and D-galactarate by EC 1.1.1.23. D-Galactarate is involved in reactions catalyzed by 4.2.1.158 that is present in *Oceanobacillus iheyensis* [52].

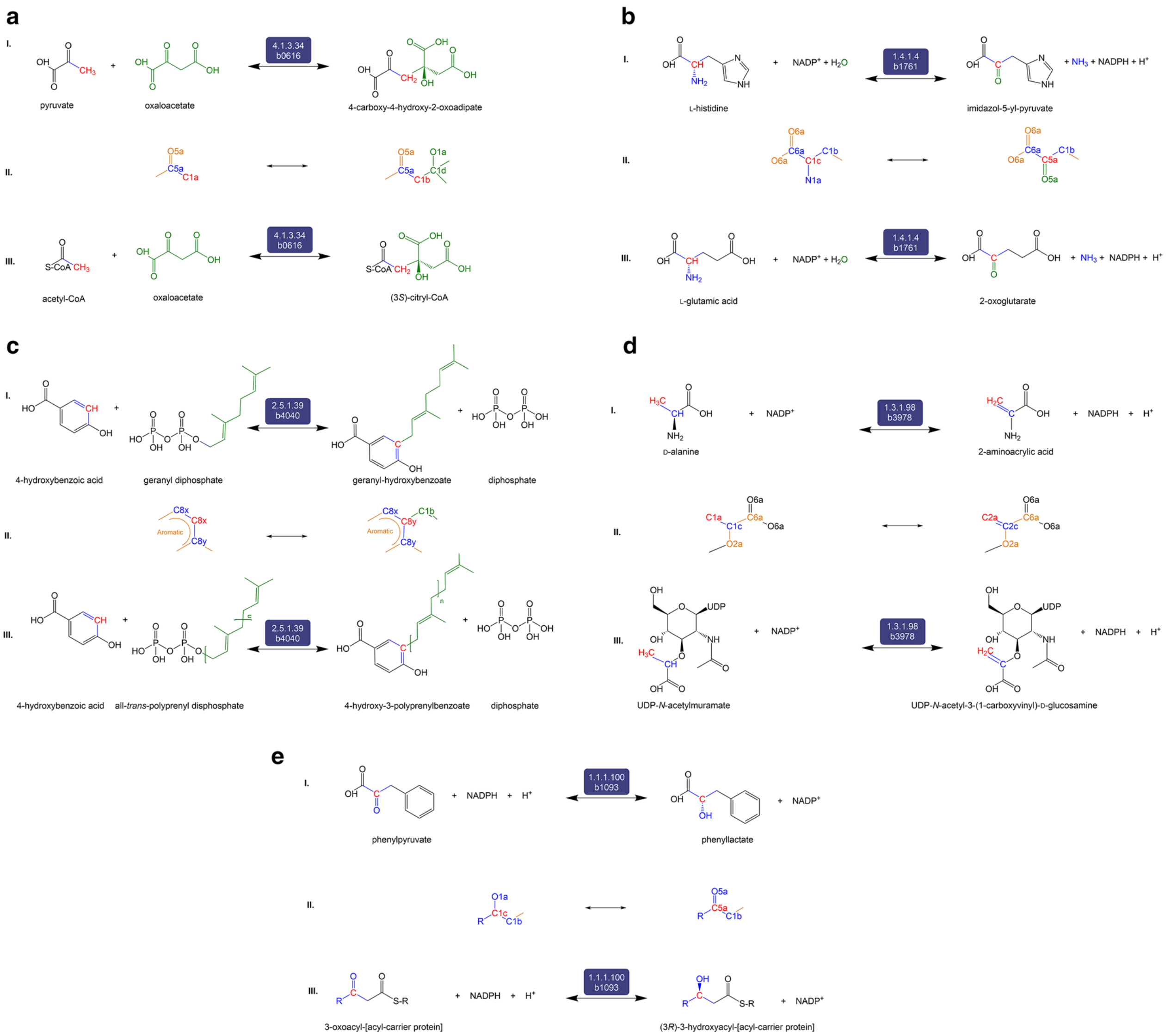

Fig. 4 The set of five reactions, panels **a**–**e**, belonging to Category 3 (C3). C3 reactions and derivatives are neither present in iML1515 nor associated with *E. coli* in KEGG and EcoCyc. However, according to KEGG, the reactions occur in other organisms. Each of the five panels is divided into 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, and (III) the native reaction catalyzed by the potentially promiscuous enzyme, as catalogued in KEGG

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.

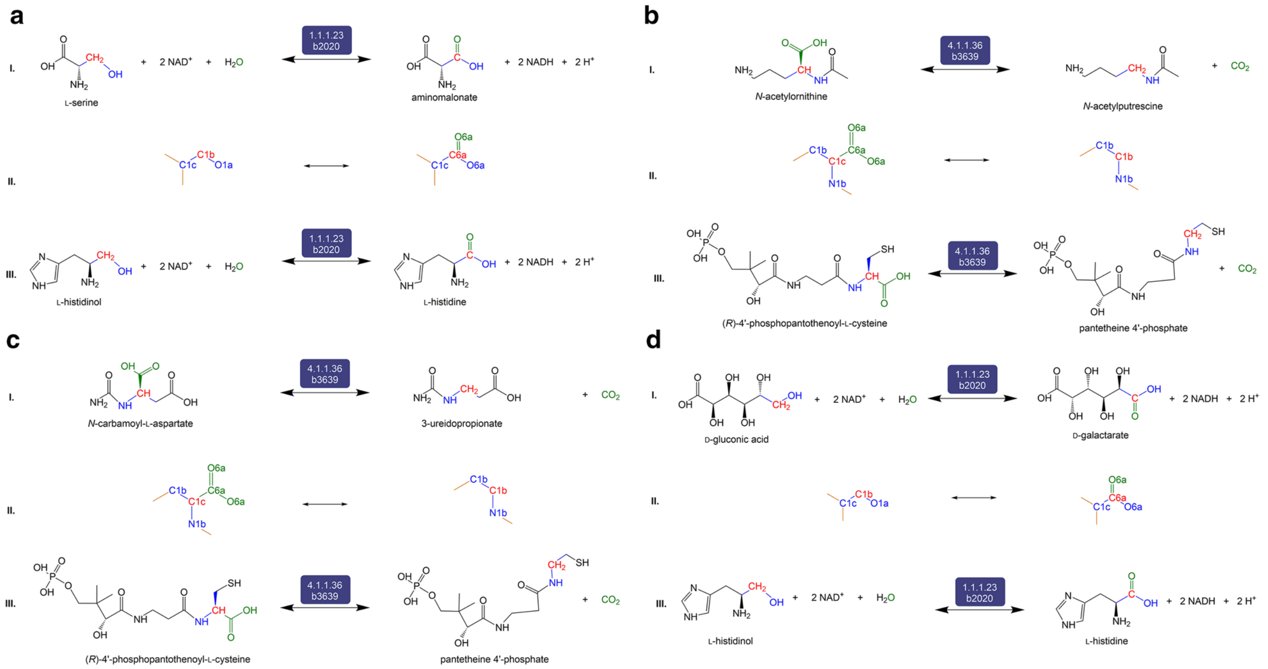

Fig. 5 The set of four reactions, panels **a**–**d**, belonging to Category 4 (C4). C4 reactions and derivatives are neither present in iML1515 nor associated with any other organism in KEGG or EcoCyc. Each of the four panels is divided into three sections (I) the balanced reaction developed by our workflow 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

The workflow we developed here is designed to identify metabolites that can form due to promiscuous enzymatic activity within a specific model organism. Further, the workflow provides balanced reactions to document such enzymatic activities. We utilized PROXIMAL [39], 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 [33, 34]. 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 [53]. 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 C1, C3 and C4 categories would provide further evidence of the suggested reactions, and would 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. When applying PROXIMAL, 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 1875 biochemical operators based on reactions in iML1515 and predicted 1423 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 23 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. The described workflow can be applied to any organism utilizing its metabolic model to predict sample-specific promiscuous enzymatic byproducts. 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 1877 metabolites, 2712 reactions and 1516 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 [40] 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 [54]. PROXIMAL constructs a lookup table of all possible biotransformations that can occur due to promiscuous

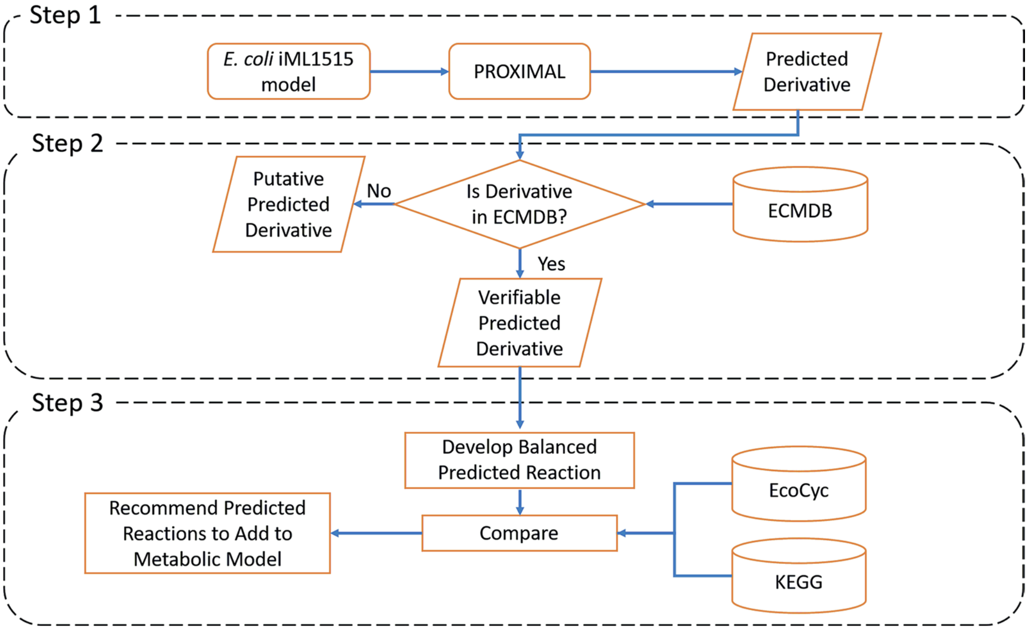
Step 1
E. coli iML1515 model
PROXIMAL
Predicted Derivative
Step 2
Putative Predicted Derivative
No
Is Derivative in ECMDB?
ECMDB
Yes
Verifiable Predicted Derivative
Step 3
Develop Balanced Predicted Reaction
EcoCyc
Recommend Predicted Reactions to Add to Metabolic Model
Compare
KEGG

Fig. 6 Main steps of EMMA workflow customized to extend the *E. coli* iML1515 model with predicted reactions. Step 1: predict promiscuous transformations and derivatives using PROXIMAL. Step 2: compare derivatives with measured metabolic dataset(s). Step 3: curation and stoichiometric balancing of reactions

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. The biotransformation operators in the lookup table were then applied to model metabolites. 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 3760 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 [55], UniProt [56, 57], YMDB [58], and CCDB [59], 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 [60], a python wrapper for the chemical toolbox Open Babel [61]. 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 3760 metabolites in ECMDB, we identified 3397 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. Each predicted transformation was manually examined and compared against the RDM pattern causing the transformation. Transformations were discarded if the they seemed infeasible, if the substrate was a cofactor, or if the RPAIR entry associated with the PROXIMAL operator required the presence of more than one Reaction Center (R). For each valid 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 developed balanced reactions, 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, KEGG, or the literature.

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 EcoCyc nor in KEGG. These reactions were thus novel reactions that have not been reported in the literature.

Additional files

Additional file 1. Listing of derivatives that were predicted by PROXIMAL and had a chemical ID in PubChem.

Additional file 2. Results of Flux Balance Analysis and Flux Variability Analysis for the added EMMA reactions.

Acknowledgements
Not applicable.

Authors' contributions

SH conceived the EMMA concept. SA developed the EMMA workflow. EC curated the results. VP verified the analysis. NN and SH supervised the work done through the development of the workflow and data curation. Manuscript was written by SA and EC, reviewed by VP, and revised by NN and SH. All authors read and approved the final manuscript.

Funding
This work is funded under NSF Grant #1421972 and NIH Grants 1DP2HD091798 and 1R03CA211839-01.

Availability of data and materials
The *E. coli* iML1615 model is available on the BiGG database. Data from ECMDB can be directly downloaded from the ECMDB website.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Computer Science, Tufts University, Medford, MA, USA.
[2] Department of Biology, University of North Carolina, Chapel Hill, NC, USA.
[3] Department of Chemical and Biological Engineering, Tufts University, Medford, MA, USA.

Received: 1 February 2019 Accepted: 5 June 2019
Published online: 13 June 2019

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. Curr Opin Biotechnol. 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. Front Plant Sci. 2015;6:7.
3. George KW, Alonso-Gutierrez J, Keasling JD, Lee TS. Isoprenoid drugs, biofuels, and chemicals—artemisinin, farnesene, and beyond. In: Schrader J, Bohlmann J, editors. Biotechnology of Isoprenoids. Berlin: Springer; 2015. p. 355–89.
4. Singh R, White D, Demirel Y, Kelly R, Noll K, Blum P. Uncoupling fermentative synthesis of molecular hydrogen from biomass formation in *Thermotoga maritima*. Appl Environ Microbiol. 2018;84(17):e00998-18.
5. Singh R, Tevatia R, White D, Demirel Y, Blum P. Comparative kinetic modeling of growth and molecular hydrogen overproduction by engineered strains of *Thermotoga maritima*. Int J Hydrog Energy. 2019;44:7125–36.
6. Du J, Shao Z, Zhao H. Engineering microbial factories for synthesis of value-added products. J Ind Microbiol Biotechnol. 2011;38(8):873–90.
7. Furusawa C, Horinouchi T, Hirasawa T, Shimizu H. Systems metabolic engineering: the creation of microbial cell factories by rational metabolic design and evolution. In: Zhong JJ, editor. Future trends in biotechnology. Berlin: Springer; 2012. p. 1–23.
8. Davy AM, Kildegaard HF, Andersen MR. Cell factory engineering. Cell Syst. 2017;4(3):262–75.
9. Lee S, Mattanovich D, Villaverde A. Systems metabolic engineering, industrial biotechnology and microbial cell factories. Microb Cell Fact. 2012;11:156.
10. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng. 2003;84(6):647–57.
11. Ranganathan S, Suthers PF, Maranas CD. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS Comput Biol. 2010;6(4):e1000744.
12. Yousofshahi M, Lee K, Hassoun S. Probabilistic pathway construction. Metab Eng. 2011;13(4):435–44.
13. Wu G, Yan Q, Jones JA, Tang YJ, Fong SS, Koffas MA. Metabolic burden: cornerstones in synthetic biology and metabolic engineering applications. Trends Biotechnol. 2016;34(8):652–64.
14. Gerstl MP, Ruckerbauer DE, Mattanovich D, Jungreuthmayer C, Zanghellini J. Metabolomics integrated elementary flux mode analysis in large metabolic networks. Sci Rep. 2015;5:8930.
15. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY. Recent advances in reconstruction and applications of genome-scale metabolic models. Curr Opin Biotechnol. 2012;23(4):617–23.
16. Saha R, Chowdhury A, Maranas CD. Recent advances in the reconstruction of metabolic models and integration of omics data. Curr Opin Biotechnol. 2014;29:39–45.
17. 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 Res. 2018;47:D649–59.
18. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
19. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform. 2017. https://doi.org/10.1093/bib/bbx085.
20. 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 Res. 2015;44(D1):D515–22.
21. Sorokina M, Stam M, Médigue C, Lespinet O, Vallenet D. Profiling the orphan enzymes. Biol Direct. 2014;9(1):10.
22. Raushel FM. Finding homes for orphan enzymes. Perspect Sci. 2016;9:3–7.
23. Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, et al. Network-level architecture and the evolutionary potential of underground metabolism. Proc Natl Acad Sci. 2014;111(32):11762–7.
24. Notebaart RA, Kintses B, Feist AM, Papp B. Underground metabolism: network-level perspective and biotechnological potential. Curr Opin Biotechnol. 2018;49:108–14.
25. Rosenberg J, Commichau FM. Harnessing underground metabolism for pathway development. Trends Biotechnol. 2019;37(1):29–37. https://doi.org/10.1016/j.tibtech.2018.08.001.
26. Hult K, Berglund P. Enzyme promiscuity: mechanism and applications. Trends Biotechnol. 2007;25(5):231–8.
27. Khersonsky O, Roodveldt C, Tawfik DS. Enzyme promiscuity: evolutionary and mechanistic aspects. Curr Opin Chem Biol. 2006;10(5):498–508.
28. Tawfik OK, Dan S. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem. 2010;79:471–505.
29. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nat Biotechnol. 2009;27(2):157.
30. D'Ari R, Casadesús J. Underground metabolism. BioEssays. 1998;20(2):181–6.
31. Liechti G, Singh R, Rossi PL, Gray MD, Adams NE, Maurelli AT. Chlamydia trachomatis dapF encodes a bifunctional enzyme capable of both D-glutamate racemase and diaminopimelate epimerase activities. MBio. 2018;9(2):e00204–18.
32. Carbonell P, Faulon J-L. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics. 2010;26(16):2012–9.
33. 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. J Cheminformatics. 2015;7(1):44.
34. 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 Synth Biol. 2016;5(10):1155–66.
35. Arora B, Mukherjee J, Gupta MN. Enzyme promiscuity: using the dark side of enzyme specificity in white biotechnology. Sustain Chem Process. 2014;2(1):25.
36. Poppe L, Paizs C, Kovács K, Irimie F-D, Vértessy B. Preparation of unnatural amino acids with ammonia-lyases and 2, 3-aminomutases. In: Pollegioni L, Servi S, editors. Unnatural amino acids. Berlin: Springer; 2012. p. 3–19.
37. Atsumi S, Hanai T, Liao JC. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature. 2008;451(7174):86.
38. 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 Synth Biol. 2016;5(11):1256–63.
39. Yousofshahi M, Manteiga S, Wu C, Lee K, Hassoun S. PROXIMAL: a method for prediction of xenobiotic metabolism. BMC Syst Biol. 2015;9(1):94.
40. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res. 2010;38(suppl_2):W138–43.

41. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes *Escherichia coli* traits. Nat Biotechnol. 2017;35(10):904.
42. Guo AC, Jewison T, Wilson M, Liu Y, Knox C, Djoumbou Y, et al. ECMDB: the *E. coli* metabolome database. Nucleic Acids Res. 2012;41(D1):D625–30.
43. 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 Res. 2015;44(D1):D495–501.
44. 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 Res. 2005;33(suppl_1):D334–7.
45. 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.
46. Quinlan JR. Simplifying decision trees. Int J Man Mach Stud. 1987;27(3):221–34.
47. 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*. Biochem J. 1969;111(2):237.
48. 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.
49. Suda S, Lawton EM, Wistuba D, Cotter PD, Hill C, Ross RP. Homologues and bioengineered derivatives of LtnJ vary in ability to form ᴅ-alanine in the lantibiotic lacticin 3147. J Bacteriol. 2012;194(3):708–14.
50. Häusler E, Petersen M, Alfermann AW. Hydroxyphenylpyruvate reductase from cell suspension cultures of *Coleus blumei* Benth. Zeitschrift für Naturforschung C. 1991;46(5–6):371–6.
51. Nagayama H, Muramatsu M, Shimura K. Enzymatic formation of aminomalonic acid from ketomalonic acid. Nature. 1958;181(4606):417.
52. Rakus JF, Kalyanaraman C, Fedorov AA, Fedorov EV, Mills-Groninger FP, Toro R, et al. Computation-facilitated assignment of the function in the enolase superfamily: a regiochemically distinct galactarate dehydratase from *Oceanobacillus iheyensis*. Biochemistry. 2009;48(48):11546–58.
53. Hassanpour N. Computational methods to advance directed evolution of enzymes and metabolomics data analysis. Tufts University; 2018.
54. 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. J Am Chem Soc. 2003;125(39):11853–65.
55. Misra RV, Horler RS, Reindl W, Goryanin II, Thomas GH. Echo BASE: an integrated post-genomic database for *Escherichia coli*. Nucleic Acids Res. 2005;33(suppl_1):D329–33.
56. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2004;32(suppl_1):D115–9.
57. Consortium U. UniProt: a hub for protein information. Nucleic Acids Res. 2014;43(D1):D204–12.
58. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, Lee J, et al. YMDB: the yeast metabolome database. Nucleic Acids Res. 2011;40(D1):D815–20.
59. 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 Res. 2004;32(suppl_1):D293–5.
60. O'Boyle NM, Morley C, Hutchison GR. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chem Cent J. 2008;2(1):5.
61. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminformatics. 2011;3(1):33.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.