Knowledge of the Neighborhood of the Reactive Site up to Three Atoms Can Predict Biochemistry and Protein Sequences

Noushin Hadadi[1,+], Homa MohamadiPeyhani[1,+], Ljubisa Miskovic[1], Marianne Seijo[1,#], Vassily Hatzimanikatis[1,*]

[+] contributed equally

[1] Laboratory of Computational Systems Biology (LCSB), EPFL, CH-1015 Lausanne, Switzerland

[#] current address: Institute for Environmental Sciences and Group of Environmental Physico Chemistry, Department F.-A. Forel for Environmental and Aquatic Sciences, University of Geneva, CH-1211, Geneva, Switzerland.

* Corresponding author:
Vassily Hatzimanikatis,
Laboratory of Computational Systems Biotechnology (LCSB),
École Polytechnique Fédérale de Lausanne (EPFL),
CH-1015 Lausanne, Switzerland
Email: vassily.hatzimanikatis@epfl.ch
Phone: + 41 (0)21 693 98 70 Fax: +41 (0)21 693 98 75

Keywords: Gap filling, Reaction similarity, Tanimoto score, Novel reactions, Pathway discovery tools, Orphan reactions, Sequence similarity, Reactive site recognition

Abbreviations: GENRE, GENome scale metabolic REconstruction, **GEM**, GEnome-scale Model, **KEGG,** Kyoto Encyclopedia of Genes and Genomes, **RF,** Reaction fingerprint, **DT,** Discrimination Threshold, **TPR,** True Positive Rate, **FPR,** False Positive Rate, **TP,** True Positives, **TN,** True Negatives, **FP,** False Positives, **FN,** False Negatives, **ROC,** Receiver Operating Characteristics, **AUC,** Area Under a ROC Curve, **GPR,** gene-protein-reaction.

AUTHORS SUMMARY

The recent advances in pathway generation tools have resulted in a wealth of *de novo* hypothetical enzymatic reactions, which lack knowledge of the protein-encoding genes associated with their functionality. Moreover, nearly half of known metabolic enzymes are orphan, i.e., they also lack an associated gene or protein sequence. Proposing genes for catalytic functions of *de novo* and orphan reactions is critical for their utility in various applications ranging from biotechnology to medicine. In this work, we propose a novel computational method that will bridge the knowledge gap and provide candidate genes for both *de novo* and orphan reactions. We demonstrate that information about a small chemical structure around the reactive sites of substrates is sufficient to correctly assign genes to the functionality of enzymatic reactions.

ABSTRACT

Thousands of biochemical reactions with characterized biochemical activities are still orphan. Novel reactions predicted by pathway generation tools also lack associated protein sequences and genes. Mapping orphan and novel reactions back to the known biochemistry and proposing genes for their catalytic functions is a daunting problem. We propose a new method, BridgIT, to identify candidate genes and protein sequences for orphan and novel enzymatic reactions. BridgIT introduces, for the first time, the information of the *enzyme binding pocket* into reaction similarity comparisons. It ascertains the similarity of two reactions by comparing the reactive sites of their substrates and their surrounding structures, along with the structures of the generated products. BridgIT compares orphan and novel reactions to enzymatic reactions with known protein sequences, and then, it proposes protein sequences and genes of the most similar non-orphan reactions as candidates for catalyzing the novel or orphan reactions. We performed BridgIT analysis of orphan reactions from KEGG 2011 (Kyoto Encyclopedia of Genes and Genomes, published in 2011) that became non-orphan in KEGG 2016, and BridgIT correctly predicted enzymes with identical third- and fourth-level EC numbers for 91% and 56% of these reactions, respectively. BridgIT results revealed that it is sufficient to know information about six atoms together with their connecting bonds around the reactive sites of the substrates to match a protein sequence to the catalytic activity of enzymatic reactions with maximal accuracy. Moreover, the same information about only three atoms around the reactive site allowed us to correctly match 87% of the analyzed enzymatic reactions. Finally, we used BridgIT to provide candidate protein sequences for 137000 novel enzymatic reactions from the recently introduced ATLAS of Biochemistry. A web-tool of BridgIT can be consulted at http://lcsb-databases.epfl.ch/BridgIT/.

INTRODUCTION

The utility of genome scale reconstructions of metabolic networks in correlating the genome with physiology hinges on the completeness and accuracy of the annotation of sequenced genomes. Even the genome scale reconstructions of well-characterized organisms such as *Escherichia coli* include orphan reactions, i.e., enzymatic reactions without protein sequences or genes associated with their functionality (3). A recent review on orphan reactions reports that almost half of enzymatic reactions cataloged in KEGG (Kyoto Encyclopedia of Genes and Genomes) (1) lack an associated protein sequence (2).

Similar problems arise in areas such as bioremediation, synthetic biology, and drug discovery, where a need to explore the potential of biological organisms beyond their natural capabilities has prompted the development of tools that are capable of generating *de novo* hypothetical enzymatic reactions and pathways (1–11). *De novo* reactions are behind many success stories in biotechnology, and they can also be used in the gap-filling of metabolic networks (2,8,9,11–14). These enzymatic reactions have well-explained biochemistry that can conceivably occur in metabolism. However, no protein-encoding genes associated with the functionality of these reactions are known, which limits their applicability for the gap-filling of genome scale models, metabolic engineering and synthetic biology applications (15).

Computational methods for identifying candidate genes of orphan reactions have traditionally been based on protein sequence similarity (16–19). Two predominant classes of these methods are based on gene/genome analysis (19–22) and metabolic information (23,24). Several bioinformatics methods combine different aspects of these two classes such as gene clustering, gene co-expression, phylogenetic profiles, protein interaction data and gene proximity for assigning genes and protein sequences to orphan reactions (25–28). All of these methods use the concept of *sequence similarity* to determine the biochemical functions of orphan reactions. However, a large portion of known enzymatic activities is still missing an associated gene due to annotation errors, the incompleteness of gene sequences (29), and the fact that homology-based methods cannot annotate orphan protein sequences with no or little sequence similarity to known enzymes (16,30).

It was argued that sequence similarity methods can provide inaccurate results as small changes in key residues might greatly alter enzyme functionality (31). In addition, these methods are not suitable for the annotation of *de novo* reactions since the current pathway prediction tools do not provide information about their sequences but about their catalytic biotransformation instead. These shortcomings motivated the development of alternative computational methods, based on the *structural similarity of reactions*, for identifying candidate protein sequences for orphan enzymatic

reactions (28,31–36). The idea behind these approaches is to assess the similarity between two enzymatic reactions via the similarity of their reaction fingerprints, i.e., the mathematical descriptors of their structural and topological properties (37). In such methods, the reaction fingerprint of an orphan reaction is compared with a set of non-orphan reference reaction fingerprints, and the genes of most similar reference reactions are then assigned as promising candidate genes for the orphan reaction. Reaction fingerprints can be generated with different similarity metrics such as the bond change, reaction center or structural similarity (36).

A class of these methods considers all of the compounds participating in reactions for comparison (36). The application of this group of methods is restricted to specific enzyme commission (EC) classes (35,38) of enzymatic reactions as there are issues in mapping reactions that involve large cofactors in the reaction mechanism (28,31–36). Another class of these methods uses the chemical structures of reactant pairs for comparison (34). While this class of methods can be applied to all classes of enzymatic reactions, it neglects the crucial role of cofactors in the reaction mechanism. Neither of these two classes was employed for assigning protein sequences to *de novo* reactions (34).

In this study, we introduce a novel computational method, BridgIT, to assign protein sequences to both *de novo and orphan* reactions. BridgIT belongs to the methods that use the reaction fingerprints to compare enzymatic reactions. Compared to currently existing methods, whose reaction fingerprints contain information about the reactants and products of reactions, BridgIT introduces an additional level of specificity by capturing the critical information of the enzyme binding pocket into reaction fingerprints. More precisely, BridgIT is substrate-reactive-site centric, and its reaction fingerprints reflect the specificities of biochemical reaction mechanisms that arise from the type of enzymes catalyzing those reactions. In BridgIT, we use the Tanimoto score to quantify the similarity of reaction fingerprints. Values of the Tanimoto scores near 0 indicate reactions with no or a negligible similarity, whereas values near 1 indicate reactions with a high similarity.

BridgIT allows us to do the following: (i) compare a given novel or orphan reaction to a set of reactions that have associated sequences, subsequently referred to as the *reference reactions*; (ii) rank the identified *similar* reactions based on the computed Tanimoto scores; and (iii) propose the sequences of the highest ranked reference reactions as possible candidates for encoding the enzyme of the given *de novo* or orphan reaction.

We demonstrate through several studies the effectiveness of utilizing the BridgIT fingerprints for mapping novel and orphan reactions to the known biochemistry. We show that BridgIT is capable of correctly predicting enzymes with an identical third–level EC (39) number for 91% of orphan reactions from KEGG 2011 that became non-orphan in KEGG 2016. We also study how the size of the BridgIT fingerprint impacts the BridgIT predictions, and we find that it is sufficient to use the fingerprints that

describe six atoms along with their connecting bonds around the reactive sites to correctly predict protein sequences. Finally, we study 137000 novel reactions from the ATLAS of biochemistry, a database of all theoretically possible biochemical reactions (40), and we provide candidate enzymes that can potentially catalyze the biotransformation of these reactions to the research community.

RESULTS AND DISCUSSION

Reference database

The BridgIT reference reaction database consists of well-characterized reactions with associated genes and protein sequences, and it was built on the basis of the KEGG 2016 reaction database (Methods). The KEGG reaction database is the most comprehensive database of enzymatic reactions, and it provides information about the biochemical reactions together with their corresponding enzymes and genes. However, half of KEGG reactions lack associated genes and protein sequences, and they are hence considered to be orphan reactions. The reference database was built with KEGG reactions that (i) are non-orphan and elementally balanced and (ii) can be reconstructed by the existing BNICE.ch generalized reaction rules. As a result, the reference reaction database contains information on approximately 5049 out of 9556 KEGG reactions (Supplementary material S1).

From reaction chemistry to detailed enzyme mechanisms. Approximately 15% of KEGG reactions (1532 reactions) are assigned to more than one enzyme (EC number), i.e., multiple enzymes can catalyze their biotransformation through different enzymatic mechanisms. For example, KEGG reaction R00217 is assigned to three different enzymes, 1.1.1.40, 1.1.1.38 (malate dehydrogenase) and 4.1.1.3 (oxaloacetate carboxy-lyase), and it can undergo two different enzymatic mechanisms (Figure 1). For the 4.1.1.3 enzyme, the reaction mechanism is well understood, as this enzyme belongs to the carboxy-lyases, where a carbon-carbon bond is broken and a molecule of CO_2 is released. In contrast, for the 1.1.1.40 and 1.1.1.38 enzymes, there is ambiguity about their detailed enzyme mechanisms. As discussed in Swiss-Prot (41), these two enzymes are NAD-dependent dehydrogenases, but they also have the ability to decarboxylate oxaloacetate. These enzymes are found in bacteria and insects (1.1.1.38) as well as in fungi, animals and plants (1.1.1.40).

We applied the BridgIT algorithm to R00217, and we obtained two distinct reaction fingerprints that corresponded to the two different enzyme mechanisms. More precisely, the BNICE.ch generalized reaction rules 1.1.1.- and 4.1.1.- reacted upon two different reactive sites of oxaloacetate to break the carbon-carbon bond and release CO_2 and pyruvate (Figure 1). The 1.1.1.- rules recognized a larger, i.e., more specific, reactive site compared to the one recognized by 4.1.1.- (Figure 1).

Therefore, a reaction from KEGG can be translated into more than one fingerprint in the BridgIT reference database. In addition, BNICE.ch describes approximately 42% of the reactions assigned to a single enzyme in KEGG with multiple reaction rules. Consequently, multiple reaction fingerprints can describe the biotransformation of each of such reactions.

This way, by virtue of preserving the information about enzyme binding pockets, the reconstructed BridgIT reference reaction database expands from 5049 *reactions* to 17657 *reaction fingerprints* corresponding to 17657 *detailed reaction mechanisms*.

Comparison of BridgIT and BLAST predictions

As a means to relate *reaction structural similarity* with *reaction sequence similarity*, we simultaneously applied BridgIT on a subset of reactions from the reference reaction database together with BLAST (42) on their corresponding protein sequences. Based on the assumption that similar protein sequences have similar functions (43), we compared the similarity results of BridgIT with those from BLAST, and we statistically assessed BridgIT performance using ROC curve analysis (Figure 2).

We chose *E. coli* BW29521 (EBW) as our benchmark organism for this analysis. We extracted all of the non-orphan reactions of EBW from the BridgIT reference database together with their associated protein sequences (Supplementary material S1). There were 531 non-orphan reactions in EBW associated with 413 protein sequences. In total, there were 731 reaction-gene associations (Supplementary material S1), as there were reactions with more than one associated gene, and there were also genes associated with more than one reaction (Figure 2, vignette 1). We then used BridgIT to assess the structural similarity of 531 EBW reactions to BridgIT reference reactions using the Tanimoto score (Figure 2, vignette 2), and we also applied BLAST to quantify the similarity of the 413 EBW protein sequences to the KEGG protein sequence database using e-values (Figure 2, vignette 3). We provided a list of BridgIT reaction-reaction comparisons together with BLAST sequence-sequence comparisons (Supplementary material S2).

Comparing reaction and sequence similarity scores. We considered two sequences to be similar if BLAST reports an e-value less than 10^{-10} for their alignment. For a chosen discrimination threshold, DT, of the global Tanimoto score, we considered the BridgIT prediction of similarity between an EBW reaction and a BridgIT reference reaction with a Tanimoto score of T_G as:

(i) True Positive, TP, if T_G > DT and their associated sequence(s) were similar (e-value $<10^{-10}$);

(ii) True Negative, TN, if not similar for both BridgIT (T_G<DT) and BLAST+ (e-value $>10^{-10}$);

(iii) False Positive, FP, if similar for BridgIT (T_G>DT) but not similar for BLAST+ (e-value $>10^{-10}$);

(iv) False Negative, FN, if not similar for BridgIT (T_G<DT) but similar for BLAST+ (e-value $<10^{-10}$).

We then counted the number of TPs, TNs, FPs and FNs for all 531 reactions, and we summed up these quantities to obtain the total number of TPs, TNs, FPs and FN per chosen DT. We repeated this procedure for a set of DT values varying across the interval between 0 and 1 (Figure 2, vignette 4). Finally, we used the total number of TPs, TNs, FPs and FNs to compute the true positive and false positive rates for the ROC curve analysis (Figure 2, vignette 5).

The ROC curve indicated that the reaction comparison based on *reaction structural similarity* (BridgIT) is comparable to the one based on *reaction sequence similarity* (BLAST). Indeed, the obtained AUC score for the BridgIT classifier was 0.91 (Figure 3, panel A). We next studied if the type of compared reactions affected the accuracy of BridgIT predictions. We categorized reactions according to their first level EC class, and we performed the ROC analysis for each class separately (Figure 3, panel A). The analysis revealed that BridgIT performed well with all enzyme classes. For all the classes, we obtained high AUC scores, ranging from 0.88 (EC 1) to 0.96 (EC 5).

We next analyzed the accuracy of BridgIT classification as a function of the discrimination threshold of the Tanimoto score, DT (Figure 3, panel B). The accuracy ranged from 43% for DT=0.01 to 85% for DT=0.29. For values of DT > 0.29, the accuracy was monotonically decreasing toward a value of 62% for DT=1. The classifier was overly conservative for values of DT > 0.29, and it was rejecting true positives (Figure 3, panel B). More specifically, for DT=0.29, the TP percentage was 38%, whereas for DT=1, it was reduced to 3%. In contrast, the TN percentage increased very slightly for the values of DT> 0.29, where for DT=0.29, it was 46%, and for DT=1, it was 57% (Figure 3, panel B). Based on this analysis, we have chosen a DT of 0.29 as an optimal threshold value for further studies.

Details about the statistical procedure are provided in Supplementary material S3.

BridgIT analysis of KEGG reactions

We applied BridgIT to 5270 KEGG reactions that could be reconstructed by the BNICE.ch generalized reaction rules (Supplementary material S1). Among them, 5049 reactions were non-orphan, and they existed in the BridgIT reference reaction database, whereas the remaining 221 reactions were orphan. BridgIT correctly mapped each of the 5049 non-orphan reactions to itself in the reference reaction database. Additionally, BridgIT identified the reference reactions with Tanimoto scores higher than the optimal threshold value of 0.29 for 92% of the orphan reactions. The remaining 8% of orphan reactions had a low similarity with the reference reactions.

BridgIT reaction fingerprints offer improved predictions. We repeated the analysis from the previous section using the standard reaction difference fingerprint (Methods), which is utilized in methods such as RxnSim (32) and RxnFinder (33), to assess the benefits of introducing the information about the

reactive site of substrates into the reaction fingerprints. Comparison of the two sets of predictions on 5049 non-orphan reactions showed that the predictions obtained with BridgIT modified fingerprints were significantly better than the standard ones. The BridgIT fingerprints mapped all non-orphan reactions correctly, whereas the standard fingerprints mismatched approximately 29%, i.e., 1464 of non-orphan reactions.

The mismatch arose from cancelations of the fragments from the substrate and product sets inside fingerprint description layers (Methods). More specifically, the fragments from the substrate and product sets were canceled out during algebraic summation in all eight description layers for 246 non-orphan reactions, i.e., their standard fingerprints were empty. The information about reactive sites introduced in the BridgIT reaction fingerprints prevents such cancellations. For example, the standard reaction fingerprint of KEGG reaction R00722 was empty (Figure 4, panel A). In contrast, we identified R00722 and R00330 as the most similar reactions to R00722 with the BridgIT fingerprint. Indeed, according to the KEGG database, the enzyme 2.7.4.6 catalyzes both of these reactions.

Furthermore, the first description layer of the standard fingerprint was empty for an additional 1129 reactions, which indicated that these fingerprints did not represent the bond changes during the course of the reaction. The remaining 89 mismatched non-orphan reactions had partial cancelations in the fingerprint description layers (Supplementary material S1). For example, we incorrectly identified R03132 as the most similar to R00691 with the standard fingerprint, whereas we identified R00691 and R01373 as the most similar to R00691 with the BridgIT fingerprint (Figure 4, panel B). KEGG reports that both R00691 and R01373 can be catalyzed by either EC 4.2.1.51 or EC 4.2.1.91.

BridgIT analysis of known reactions with common enzymes. The 5049 reactions in the reference database were catalyzed by 2983 enzymes, i.e., there were *promiscuous* enzymes that catalyzed more than one reaction. Out of 2983 enzymes, 844 were promiscuous, and they catalyzed 2432 reactions (Supplementary material S1). BridgIT analysis of these 2432 reactions indicated that more than 80% of them were correctly identified within the groups catalyzed by the same enzyme. An example of such a group is given in Table 1.

We investigated the remaining 20 percent of reactions in depth, and we observed that the Tanimoto scores of the first two description layers (Methods) indicated a very low similarity between the reactions catalyzed by the same enzyme. This result suggested that such enzymes were either multi-functional, i.e., they had more than one reactive site (Figure 5), or incorrectly classified in the EC classification system.

BridgIT predicts correct enzymes for KEGG 2011 orphan reactions. We compared the number of orphan reactions in the two versions of the KEGG reaction database, KEGG 2011 and KEGG 2016. We found that 64 orphan reactions from KEGG 2011 were later associated with enzymes in KEGG 2016, i.e., they became non-orphan reactions (Supplementary material S1). We used these 64 reactions as a benchmark to assess BridgIT performance. For 34 out of 64 (53%) reactions, the BridgIT algorithm proposed the same enzymes that KEGG 2016 assigned to these reactions (Figure 6, Supplementary material S1).

We also compared BridgIT results with the KEGG 2016 assignments up to the third EC level, and strikingly, BridgIT and KEGG 2016 assigned enzymes matched to the third EC level for 58 (91%) reactions (Figure 6, Supplementary material S1). A high matching score in this comparison is likely because BridgIT uses BNICE.ch generalized reaction rules, which describe the biotransformations of reactions with specificity up to the third EC level.

Sensitivity analysis of the BridgIT fingerprint size. The defining characteristic of the BridgIT reaction fingerprint is that it is centered around the reactive site of the reaction substrate. The number of description layers in the BridgIT fingerprint, i.e., its size, defines how large of a chemical structure around the reactive site we consider when evaluating the similarity (Methods). To investigate to what extent the fingerprint size affects the similarity results, we performed sensitivity analysis.

To ensure an unbiased assessment of the effects of the fingerprint size on the similarity results, we started by removing 416 out of the 5049 non-orphan reactions whose substrates could be described with less than seven description layers (Figure 7, panel A). We then formed the reaction fingerprints that contained only the description layer 0 (fingerprint size 0), and we evaluated how many out of remaining 4626 reactions BridgIT could correctly match. We next formed the reaction fingerprints with only the description layers 0 and 1 (fingerprint size 1), and we performed the evaluation again. We repeated this procedure until the final step, where we formed the reaction fingerprints with seven description layers (fingerprint size 7). As expected, the more description layers that were incorporated into the BridgIT fingerprint, the more accurately BridgIT matched the analyzed reactions (Figure 7, panel B). BridgIT correctly matched 96.5% of analyzed reactions for a fingerprint size 5, and it matched all of analyzed reactions for a fingerprint size 7. Considering that the description layers 0 and 1 describe the single atoms and the connected pairs of atoms of the reactive site (Methods), layers 2 to 7 also describe the chemical structure around the reactive site that contains up to eight atoms and seven bonds. Therefore, the information about six atoms along with their connecting bonds *around* the reactive sites is sufficient for BridgIT to correctly match all non-orphan KEGG reactions. Furthermore, we correctly matched 87.7% of the analyzed enzymatic reactions using the same information for only

three atoms around the reactive site (fingerprint size 4), which surpasses the 71% of matched reactions when using the standard reaction fingerprints (fingerprint size 7). The enzymes that catalyzed the 12.3% of reactions that we could not match with the BridgIT fingerprints of size 4 acted mostly upon reactive sites that involve ring structures.

BridgIT analysis of novel ATLAS reactions

The ATLAS of biochemistry provides a comprehensive catalog of theoretically possible biotransformations between KEGG compounds, and it can be mined for novel biosynthetic routes for a wide range of applications in metabolic engineering, synthetic biology, identification of drug targets and bioremediation (40).

We utilized BridgIT to identify candidate enzymes of more than 137000 *de novo* ATLAS reactions. If the identified candidate enzymes can catalyze their ATLAS reactions, we can use them directly in the systems biology designs. Otherwise, we can use their amino acid sequences as initial guesses in protein engineering. We found that 7% of ATLAS novel reactions were matched to known KEGG reactions with a Tanimoto score of 1, while 88% were similar to KEGG reactions with a Tanimoto score higher than 0.29. The remaining 5% of the ATLAS novel reactions were not similar to the well-characterized known enzymatic reactions.

We illustrated the procedure of identifying candidate enzymes of *de novo* reactions through an example. We applied the generalized reaction rules to the substrates of the novel ATLAS reaction rat 127359 (Figure 8, panel 1), and three rules, 3.1.2.-, 3.3.1.- and 6.3.1.-, were able to describe this reaction (Figure 8, panel 2). We then constructed reaction fingerprints around three identified reactive sites, and we compared them with the reference reaction fingerprints based on the Tanimoto score (Figure 8, panel 3). BridgIT suggested reaction R07294 as the best candidate, which had a high similarity with rat 127359 regarding the structure of the substrate and the reaction mechanism, and it also had an identical third-level EC number (Figure 8, panel 4).

Finding well-characterized reactions, that are similar to the novel ones is crucial for evolutionary protein engineering as well as computational protein design and consequently for the experimental implementation of the *de novo* reactions.

Results of BridgIT analysis on the ATLAS reactions are available on the website http://lcsb-databases.epfl.ch/atlas/.

METHODS

The BridgIT method allows us to link orphan reactions and *de novo* reactions, predicted by pathway design tools such as BNICE.ch (12), retropath2 (11), DESHARKY (6), and SimPheny (8), with well-

characterized enzymatic reactions and their associated genes. BridgIT is inspired by the "lock and key" principle, which is used in protein docking methods (44). The enzyme binding pocket is considered to be the "lock" and the ligand is a "key". If a molecule has the same reactive sites and similar surrounding structure as the native substrate of a given enzyme, it is then rational to expect that the enzyme will catalyze the same biotransformation on this molecule. Following this reasoning, BridgIT uses the structural similarity of the reactive sites of participating substrates together with their surrounding structure as a metric for the similarity of enzymatic reactions.

We used the curated generalized reaction rules of BNICE.ch to extract information about the reactive sites of participating substrates and integrated it into BridgIT reaction fingerprints. BNICE.ch, its applications, and the concept of generalized reaction rules are discussed elsewhere (40,45,46).

BridgIT workflow consists of four main steps: 1) reactive site identification, 2) reaction fingerprint construction, 3) reaction similarity evaluation and 4) scoring, ranking and gene assignment (Figure 9).

Reactive site identification. An enzymatic reaction triggers when its substrate(s) fits perfectly in the binding site of the enzyme. Since the structure and geometry of the binding sites of enzymes are complex and most of the time not fully characterized, we propose focusing on the similarity of the reactive sites of their substrates. Following this, we used the generalized reaction rules of BNICE.ch to identify the reactive sites of substrates. The expert-curated reaction rules have third-level EC identifiers, e.g., EC 1.1.1, and they encompass the following biochemical knowledge of enzymatic reactions: (i) the information about atoms of the substrate's reactive site; (ii) their connectivity (atom-bond-atom); and (iii) the exact information of bond breakage and formation in the course of the reaction. As of July 2017, BNICE.ch contains 361*2 bidirectional generalized reaction rules that can reconstruct 6528 KEGG reactions (40).

Given a novel or orphan reaction, we identify the reactive sites of its substrate(s) in three steps. In the first step, we identify BNICE.ch generalized reaction rules that can be applied to groups of atoms from the analyzed substrates. We then store the information about the identified rules and the corresponding groups of atoms. Subsequently, we refer to these groups of atoms as the candidate substrate reactive sites. In the second step, among the identified rules, we keep only the ones that can recognize the connectivity between the atoms of the candidate substrate reactive sites. In the third step, we then test if the biotransformation of a substrate(s) to a product(s) can be explained by the rules retained after the second step. The candidate reactive sites corresponding to the rules that have passed the three-step test are validated and used further for the construction of reaction fingerprints.

We exemplify this procedure on the *de novo* reaction rat 132064, which catalyzes the conversion of 3,4-dyhydroxymandelonitrile, substrate A, to protocatechualdehyde and cyanide (Figure 9). In the first

step, we identified 164 rules out of 361*2 rules that could be applied to groups of atoms of substrate A (Figure 9, panel 1a). Out of the 164 rules, 103 matched the connectivity (Figure 9, panel 1b). Finally, we applied 103 reaction rules to substrate A for bond breaking and formation comparisons, and one rule could explain the transformation of substrate A to the products (Figure 9, panel 1c).

Reaction fingerprint construction. Linear representations of the structures of molecules, molecular fingerprints, have been used in many methods and for different applications, especially for structural comparison of compounds (47,48). In one of the most commonly used molecular fingerprints, the Daylight fingerprint (47), a molecule is decomposed into eight layers starting from layer zero that accounts only for atoms. Layer 1 expands one bond away from all of the atoms and accounts for atom-bond-atom connections. This procedure is continued until layer 7 that includes seven connected bonds from each atom. There are two types of Daylight reaction fingerprints: (i) structural reaction fingerprint, which is a simple combination of reactant and product fingerprints, and (ii) reaction difference fingerprint, which is the algebraic summation of reactant and product fingerprints multiplied by their stoichiometry coefficients in the reaction. In this study, we propose a modified version of the reaction difference fingerprint. The procedure of formulating BridgIT reaction fingerprints is demonstrated through an example reaction (Figure 9, panels 2.a and 2.b).

Starting from the atoms of the identified substrate reactive site, we formed eight description layers of a molecule, where each layer consisted of fragments with different lengths. Fragments were composed of atoms connected through unbranched sequences of bonds. Depending on the number of bonds included in the fragments, we formed different description layers of a molecule as follows:

Layer 0: Described the type of each atom of the reactive site together with its count. For example, the substrate of the example reaction at layer 0 was described with 1 oxygen, 1 nitrogen and 2 carbon atoms (Figure 9, panel 2.a).

Layer 1: Described the type and count of each bond between pairs of atoms in the reactive site. In the example, the substrate at layer 1 was described with three fragments of length 1: 1 C-O, 1 C-C and 1 C≡N bond (Figure 9, panel 2.a). Fragments are shown by their SMILES molecular representation (49).

Layer 2: Described the type and count of fragments with 3 connected atoms. While layers 0 and 1 described the atoms of reactive sites, starting from layer 2, we also described atoms that were outside of the reactive site. In the illustrated example, we had 3 different fragments of this type (Figure 9, panel 2.a).

We used the same procedure to describe the molecules up to layer 7, as descriptions up to this layer are known to capture the structure of most of the metabolites in biochemical reactions (37).

Not all description layers are needed to describe less complex molecules. For example, product C (cyanide) was fully described with layer 0 and layer 1 (Figure 9, panel 2.a). For very large molecules, we can use the description layers that contain fragments with more than 8 connected atoms.

For each layer, we next formed the following: (i) the substrate set by merging all of the fragments, their type and their count in the substrate molecules of the reaction and (ii) the product set by merging all of the fragments (type and count) in the product molecules of the reaction. In both sets, we multiplied the count of each fragment by the stoichiometric coefficients of the corresponding compound in the reaction. Finally, we created the reaction fingerprints by summing the fragments of the substrate and product sets for each layer (Figure 9, panel 2.b).

Introducing the specificity of reactive sites into the reaction fingerprint allows BridgIT to capitalize on the information about enzyme bonding pockets. In order to keep this valuable information throughout the generation of reaction fingerprints, BridgIT does not consider the atoms of the reactive site(s) when performing the algebraic summation of the substrate and product set fragments. Consequently, the BridgIT algorithm enables retaining, tracking and emphasizing the information of the reactive site(s) in all of the layers of the reaction fingerprint, which distinguishes it from the existing methods.

Reaction similarity evaluation. We quantified the similarity of two reactions with the similarity score between their fingerprints, subsequently referred to as reaction fingerprints 1 and 2. In this study, we used the Tanimoto score, which is an extended version of the Jaccard coefficient and cosine similarity (50). We calculated the Tanimoto score for each descriptive layer, T_{Lk}, together with the global Tanimoto score, T_G. The Tanimoto score for the k-th descriptive layer was defined as:

$$T_{Lk} = \frac{c_k}{a_k + b_k - c_k}$$

where a_k was the count of the fragments in the k-th layer of reaction fingerprint 1; b_k is the count of the fragments in the k-th layer of reaction fingerprint 2; and c_k was the number of common k-th layer fragments of reaction fingerprints 1 and 2. Two fragments are equal if their canonical SMILES and their stoichiometric coefficients are identical. We defined the global Tanimoto similarity score, T_G, as follows:

$$T_G = \frac{\sum_{k=0}^{7} c_k}{\sum_{k=0}^{7} a_k + \sum_{k=0}^{7} b_k - \sum_{k=0}^{7} c_k}$$

For each reaction fingerprint, we computed its Tanimoto similarity score against the reaction fingerprints from the BridgIT reference database, which contained reaction fingerprints of all known well-characterized enzymatic reactions (Figure 9, panel 3).

Sorting, ranking and gene assignment. For a given input reaction, we ranked the reference reactions using the computed T_G scores. We distinguished the reference reactions with the same T_G score based on the T_L score of layers 0 and 1. The algorithm also allows the assignment of ranking weights to layers specified by the user. We then assigned the protein sequences associated with the highest ranked, i.e., the most similar, reference reactions to the input reaction (Figure 9, panel 4).

CONCLUSIONS

We developed the computational tool BridgIT to evaluate and quantify the structural similarity of biochemical reactions by exploiting the biochemical knowledge of BNICE.ch generalized reaction rules. Benefiting from the capability of the generalized reaction rules to identify reactive sites of substrates, BridgIT translates the structural definition of biochemical reactions into a novel type of reaction fingerprint that explicitly describes the atoms of the substrates reactive sites and their surrounding structure. These reaction fingerprints can then be used to compare and score all novel and orphan reactions with well-characterized reference reactions and, consequently, to link them with genes, genomes, and organisms. We demonstrated through several examples improvements that the BridgIT fingerprint brings compared to the fingerprints currently existing in the literature.

Unlike traditional sequence similarity methods, BridgIT can also identify the protein sequence candidates for *de novo* reactions. We applied BridgIT to *de novo* reactions of the ATLAS of Biochemistry database, and we proposed several candidate enzymes for each of them. The candidate enzymes for *de novo* reactions are either capable of catalyzing these reactions or they can serve as initial sequences for enzyme engineering. The obtained BridgIT similarity scores can also be used as a confidence score to assess the feasibility of the implementation of novel ATLAS reactions in metabolic engineering and systems biology studies.

The applications of BridgIT go beyond merely bridging gaps in metabolic reconstructions, as it can be used to identify the potential utility of existing enzymes for bioremediation as well as for various applications in synthetic biology and metabolic engineering. As the field of metabolic engineering is growing and the metabolic engineering applications are increasingly turned toward the production of valuable industrial chemicals such as 1,4-butanediol (51,52), we expect that methods for the design of *de novo* synthetic pathways such as BNICE.ch (12) and methods for identifying candidate enzymes for *de novo* reactions such as BridgIT will grow in importance.

References

1. Gao J, Ellis LBM, Wackett LP. The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. Nucleic Acids Res. 2010 Jan;38(suppl_1):D488–91.

2. Hatzimanikatis V, Li CH, Ionita JA, Henry CS. Exploring the diversity of complex metabolic networks. Bioinformatics. 2005;21:1603–1609.

3. Hatzimanikatis V, Li C, Ionita JA, Broadbelt LJ. Metabolic networks: enzyme function and metabolite structure. Curr Opin Struct Biol. 2004 Jun;14(3):300–6.

4. Soh KC, Hatzimanikatis V. DREAMS of metabolism. Trends Biotechnol. 2010 Oct;28(10):501–8.

5. Carbonell P, Planson A-G, Fichera D, Faulon J-L. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Syst Biol. 2011;5(1):122.

6. Rodrigo G, Carrera J, Prather KJ, Jaramillo A. DESHARKY: automatic design of metabolic pathways for optimal cell growth. Bioinformatics. 2008 Nov 1;24(21):2554–6.

7. Cho A, Yun H, Park J, Lee S, Park S. Prediction of novel synthetic pathways for the production of desired chemicals. BMC Syst Biol. 2010;4(1):35.

8. Yim H, Haselbeck R, Niu W, Pujol-Baxley C. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat Chem Biol. 2011;445–452.

9. Campodonico MA, Andrews BA, Asenjo JA, Palsson BO, Feist AM. Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path. Metab Eng. 2014;25:140–158.

10. Prather KLJ, Martin CH. De novo biosynthetic pathways: rational design of microbial chemical factories. Curr Opin Biotechnol. 2008 Oct;19(5):468–74.

11. Delépine B, Duigou T, Carbonell P, Faulon J-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. 2017 Jun 29 [cited 2017 Aug 18]; Available from: http://biorxiv.org/lookup/doi/10.1101/141721

12. Hadadi N, Hatzimanikatis V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. Curr Opin Chem Biol. 2015;28:99–104.

13. Carbonell P, Parutto P, Herisson J, Pandit SB, Faulon JL. XTMS: pathway design in an eXTended metabolic space. Nucleic Acids Res. 2014;42:389–394.

14. Hadadi N, Soh KC, Seijo M, Zisaki A. A computational framework for integration of lipidomics data into metabolic pathways. Metab Eng. 2014;23:1–8.

15. Rolfsson O, Palsson BØ, Thiele I. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. BMC Syst Biol. 2011;5(1):155.

16. Sorokina M, Stam M, Medigue C, Lespinet O, Vallenet D. Profiling the orphan enzymes. Biol Direct. 2014;9.

17. Karp PD. Call for an enzyme genomics initiative. Genome Biol. 2004;5.

18. Orth JD, Palsson BO. Systematizing the Generation of Missing Metabolic Knowledge. Biotechnol Bioeng. 2010;107:403–412.

19. Osterman A, Overbeek R. Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol. 2003;7:238–251.

20. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. In: Proceedings of the National Academy of Sciences of the United States of America 1999, 96. p. 2896–2901.

21. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. In: Proceedings of the National Academy of Sciences of the United States of America 1999, 96. p. 4285–4288.

22. Chen V. Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biol. 2006;17.

23. Overbeek R, Begley T, Butler RM, Choudhuri JV. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33:5691–5702.

24. Vallenet D, Labarre L, Rouy Z, Barbe V. a microbial genome annotation system supported by synteny results. Nucleic Acids Res. 2006;34:53–65.

25. Kharchenko P, Chen LF, Freund Y, Vitkup D, Church GM. Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics. 2006;7.

26. Yamanishi Y, Mihara H, Osaki M, Muramatsu H. Prediction of missing enzyme genes in a bacterial metabolic network - Reconstruction of the lysine-degradation pathway of Pseudomonas aeruginosa. Febs J. 2007;274:2262–2273.

27. Chen Y, Mao FL, Li G, Xu Y. Genome-wide discovery of missing genes in biological pathways of prokaryotes. BMC Bioinformatics. 2011;12.

28. Smith AAT, Belda E, Viari A, Medigue C, Vallenet D. The CanOE Strategy: Integrating Genomic and Metabolic Contexts across Multiple Prokaryote Genomes to Find Candidate Genes for Orphan Enzymes. Plos Comput Biol. 2012;8.

29. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. Plos Comput Biol. 2009;5.

30. Green ML, Karp PD. Using genome-context data to identify specific types of functional associations in pathway/genome databases. Bioinformatics. 2007;23:205–211.

31. Matsuta Y, Ito M, Tohsato Y. ECOH: An Enzyme Commission number predictor using mutual information and a support vector machine. Bioinformatics. 2013;29:365–372.

32. Giri V, Sivakumar TV, Cho KM, Kim TY, Bhaduri A. RxnSim: a tool to compare biochemical reactions. Bioinformatics. 2015;31:3712–3714.

33. Hu QN, Deng Z, Hu HA, Cao DS, Liang YZ. RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. Bioinformatics. 2011;27:2465–2467.

34. Moriya Y, Yamada T, Okuda S, Nakagawa Z. Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate-Product Pairs. J Chem Inf Model. 2016;56:510–516.

35. Hu QN, Zhu H, Li XB, Zhang MM. Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. Plos One. 2012;7.

36. Rahman SA, Cuesta SM, Furnham N, Holliday GL, Thornton JM. EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat Methods. 2014 Feb;11(2):171–4.

37. DAYLIGHT, Version 4.62, DAYLIGHT Inc., Mission Viejo, CA.

38. Sacher O, Reitz M, Gasteiger J. Investigations of Enzyme-Catalyzed Reactions Based on Physicochemical Descriptors Applied to Hydrolases. J Chem Inf Model. 2009 Jun 22;49(6):1525–34.

39. International Union of Biochemistry and Molecular Biology, Webb EC, editors. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press; 1992. 862 p.

40. 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 Oct 21;5(10):1155–66.

41. Pundir S, Magrane M, Martin MJ, O'Donovan C, The UniProt Consortium. Searching and Navigating UniProt Databases: Searching and Navigating UniProt Databases. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, editors. Current Protocols in Bioinformatics [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2015 [cited 2017 Aug 18]. p. 1.27.1-1.27.10. Available from: http://doi.wiley.com/10.1002/0471250953.bi0127s50

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990;215:403–410.

43. Duan Z-H, Hughes B, Reichel L, Perez DM, Shi T. The relationship between protein sequences and their gene ontology functions. BMC Bioinformatics. 2006;7(Suppl 4):S11.

44. Rogers DJ, Tanimoto TT. A Computer Program for Classifying Plants. Science. 1960(132):1115–1118.

45. Hadadi N, Hatzimanikatis V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. Curr Opin Chem Biol. 2015 Oct;28:99–104.

46. Hadadi N, Hafner J, Soh KC, Hatzimanikatis V. Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites. Biotechnol J. 2017 Jan;12(1):1600464.

47. Briem H, Lessel UF. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. In: Klebe G, editor. Virtual Screening: An Alternative or Complement to High

Throughput Screening? [Internet]. Dordrecht: Kluwer Academic Publishers; 2002 [cited 2017 Aug 18]. p. 231–44. Available from: http://link.springer.com/10.1007/0-306-46883-2_13

48. 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.

49. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." Journal of chemical information and computer sciences 28.1 (1988): 31-36. J Chem Inf Comput Sci. 28(1):31–6.

50. Leydesdorff L. On the normalization and visualization of author co-citation data: Salton's Cosineversus the Jaccard index. J Am Soc Inf Sci Technol. 2008 Jan 1;59(1):77–85.

51. Burgard A, Burk MJ, Osterhout R, Van Dien S, Yim H. Development of a commercial scale process for production of 1,4-butanediol from sugar. Curr Opin Biotechnol. 2016;42:118–125.

52. Andreozzi S, Chakrabarti A, Soh KC, Burgard A. Identification of metabolic engineering targets for the enhancement of 1,4-butanediol production in recombinant E. coli using large-scale kinetic models. Metab Eng. 2016;35:148–159.

FIGURES

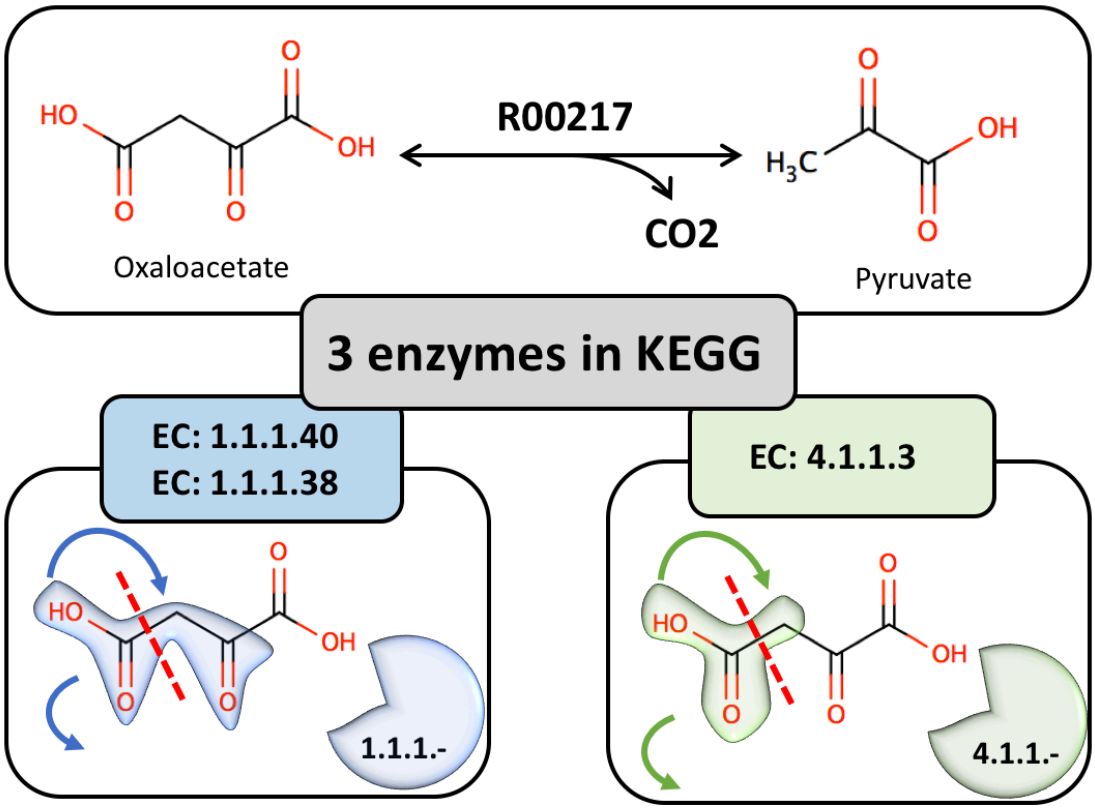

Figure 1. A multi-enzyme reaction such as R00217 can be catalyzed by more than one enzyme. BridgIT identified two distinct fingerprints for this reaction that correspond to two reactive sites of oxaloacetate. The reactive site recognized by the 1.1.1.- rule is more specific (blue substructure) than the one recognized by the 4.1.1.- rule (green substructure).

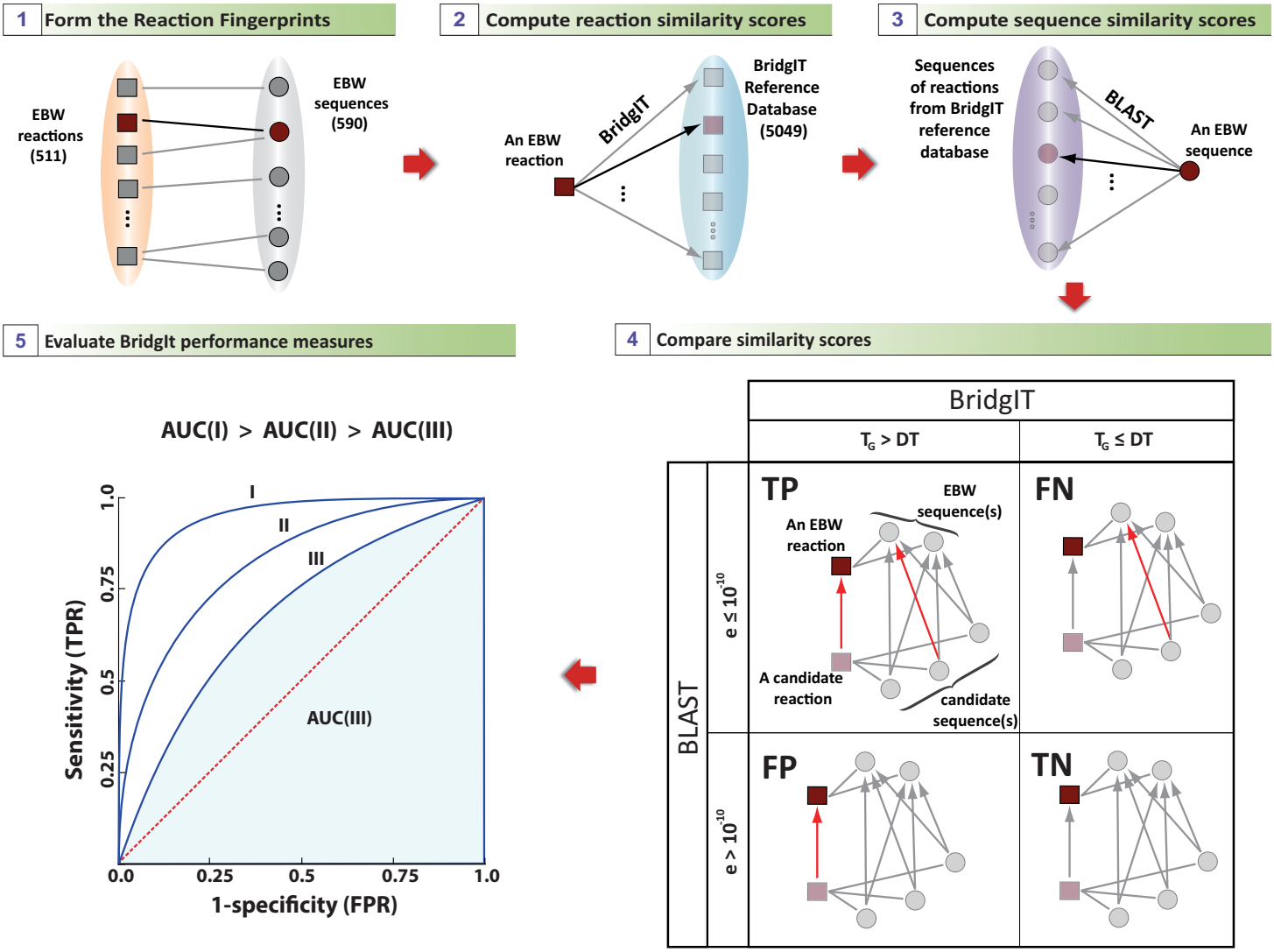

Figure 2. Five steps in the BridgIT cross validation procedure.

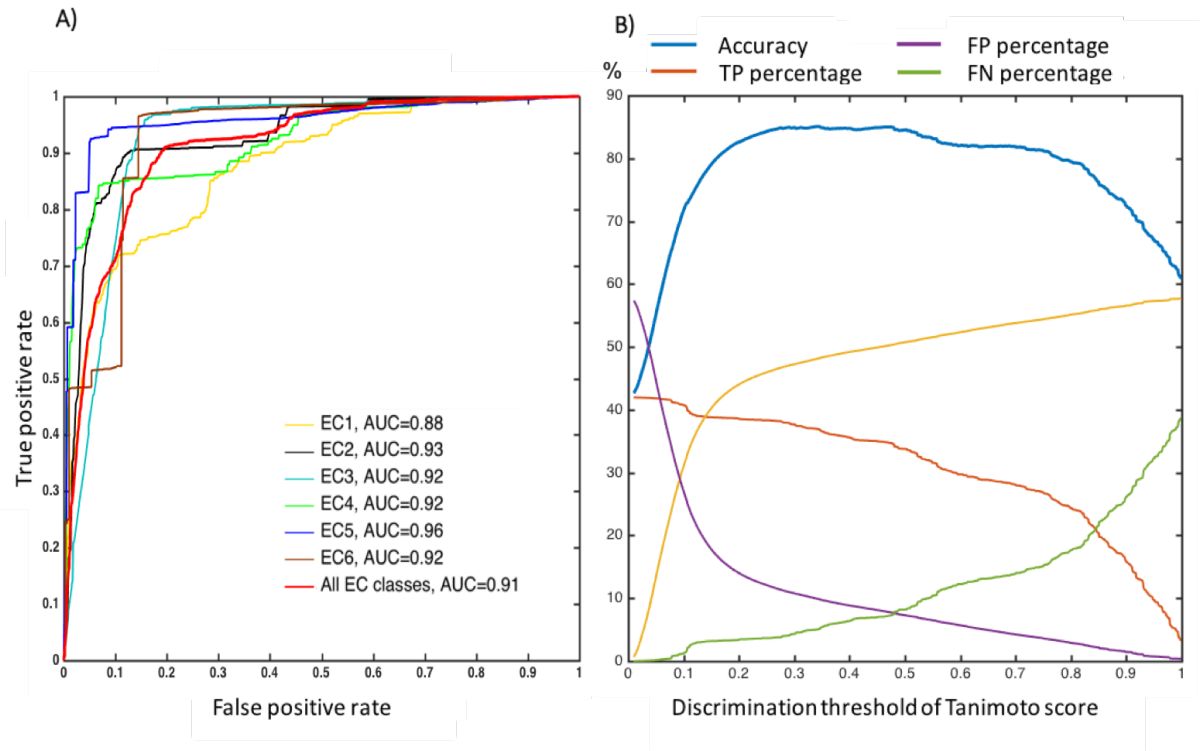

Figure 3. Panel A: ROC curve for the BridgIT classifier among all EC classes and inside each class. Panel B: Accuracy characteristics and the percentages of TP, TN, FP and FN as a function of the discrimination threshold DT. The percentages are computed as X %=100.X/(TP+TN+FN+FP) where X can be TP, TN, FP or FN.

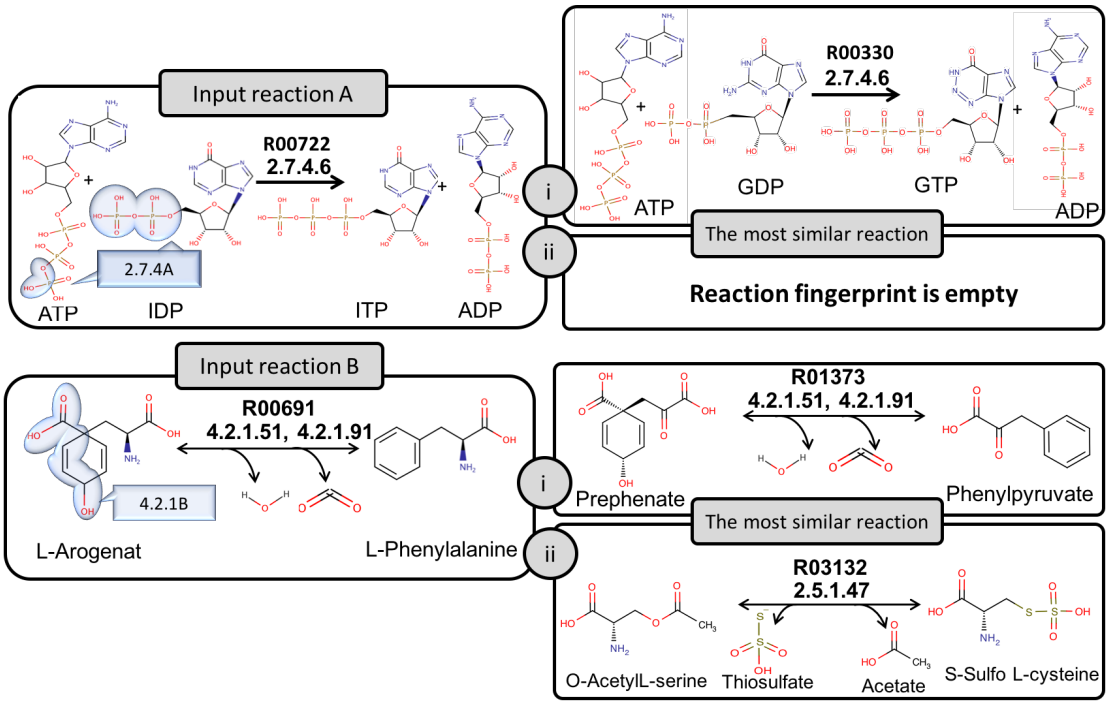

Figure 4. Comparison of the results obtained with the BridgIT and standard fingerprint on two example KEGG reactions. Panel A: the input reaction R00722 (left) and the most similar reactions identified with the BridgIT (right, i) and standard (right, ii) fingerprints. Panel B: the input reaction R00691 (left) and the most similar reactions identified with the BridgIT (right, i) and standard (right, ii) fingerprints.

Table1. A group of 5 reactions catalyzed by 1.1.1.219. High Tanimoto scores indicate that BridgIT correctly predicts the similarity of reactions within this group.

1.1.1.219					
Catalyzed reactions	**R03123**	**R03636**	**R05038**	**R07999**	**R07998**
R03123	1	0.96	0.93	0.93	0.98
R03636	0.96	1	0.96	0.94	0.95
R05038	0.93	0.96	1	0.97	0.91
R07999	0.93	0.94	0.97	1	0.91
R07998	0.98	0.95	0.91	0.91	1

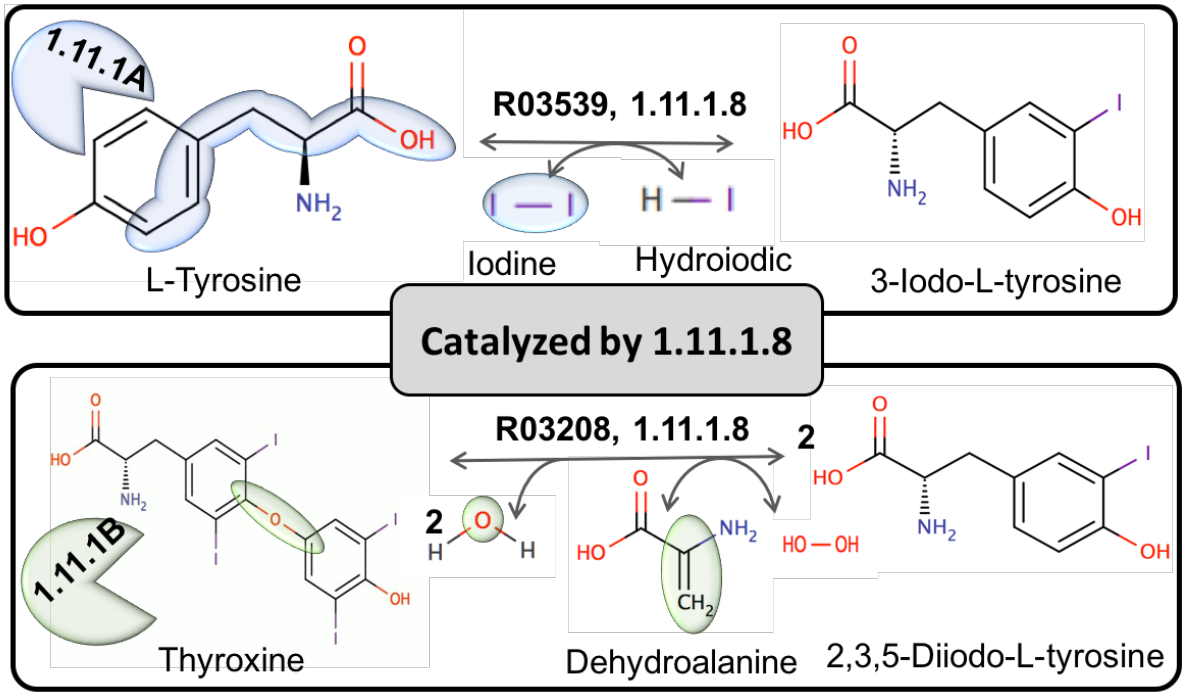

Figure 5. R03539 and R03208 are catalyzed by the same enzyme, 1.11.1.8. However, the reactive sites of substrates are completely different.

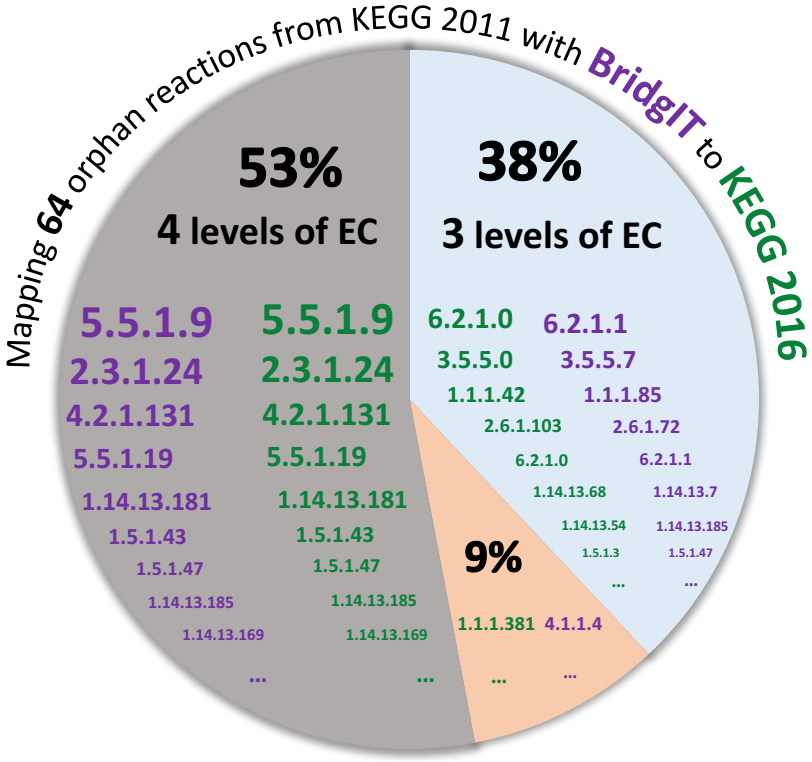

Figure 6. BridgIT successfully predicts enzymes for 58 out of 64 orphan reactions from KEGG 2011 that became non-orphan in KEGG 2016.

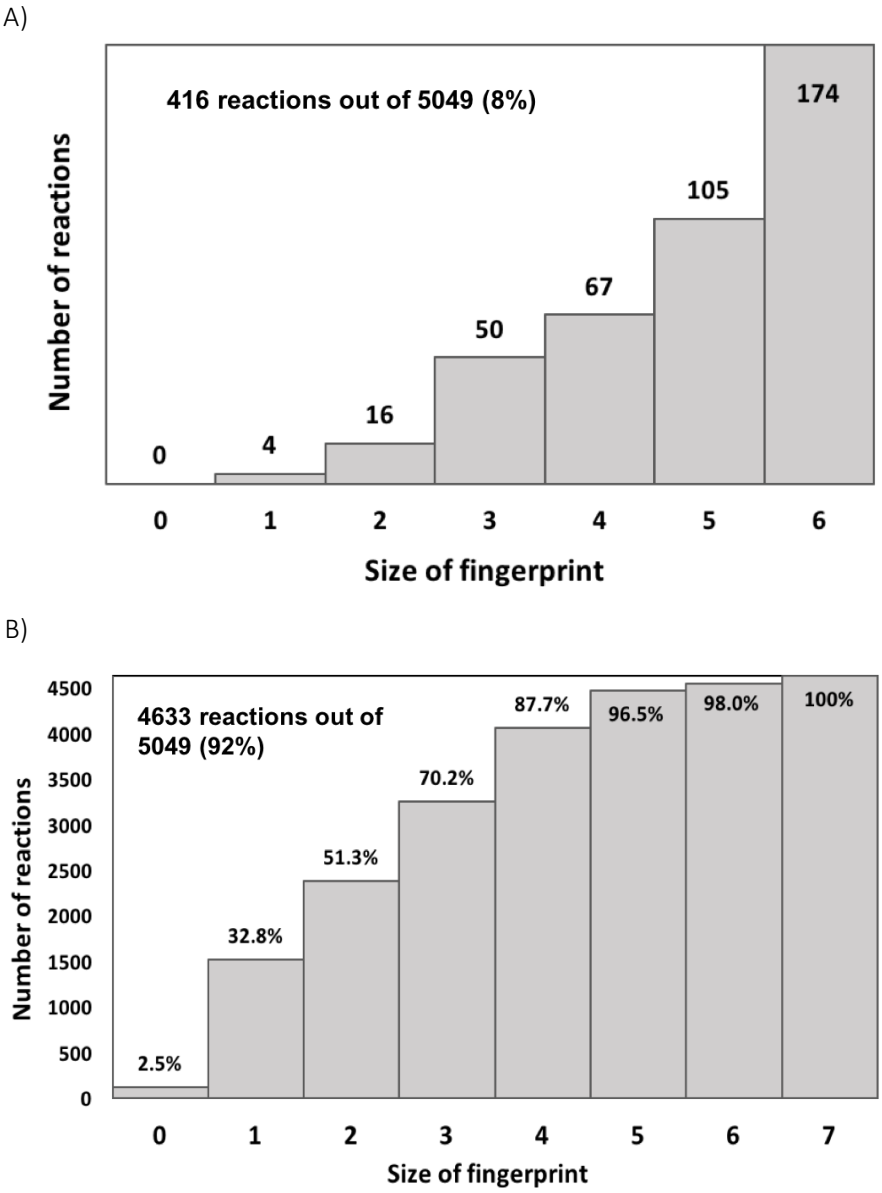

Figure 7. Panel A: the non-orphan reactions that can be completely described with the fingerprints up to six description layers. Panel B: percent of correctly matched reactions as a function of the BridgIT fingerprint size.

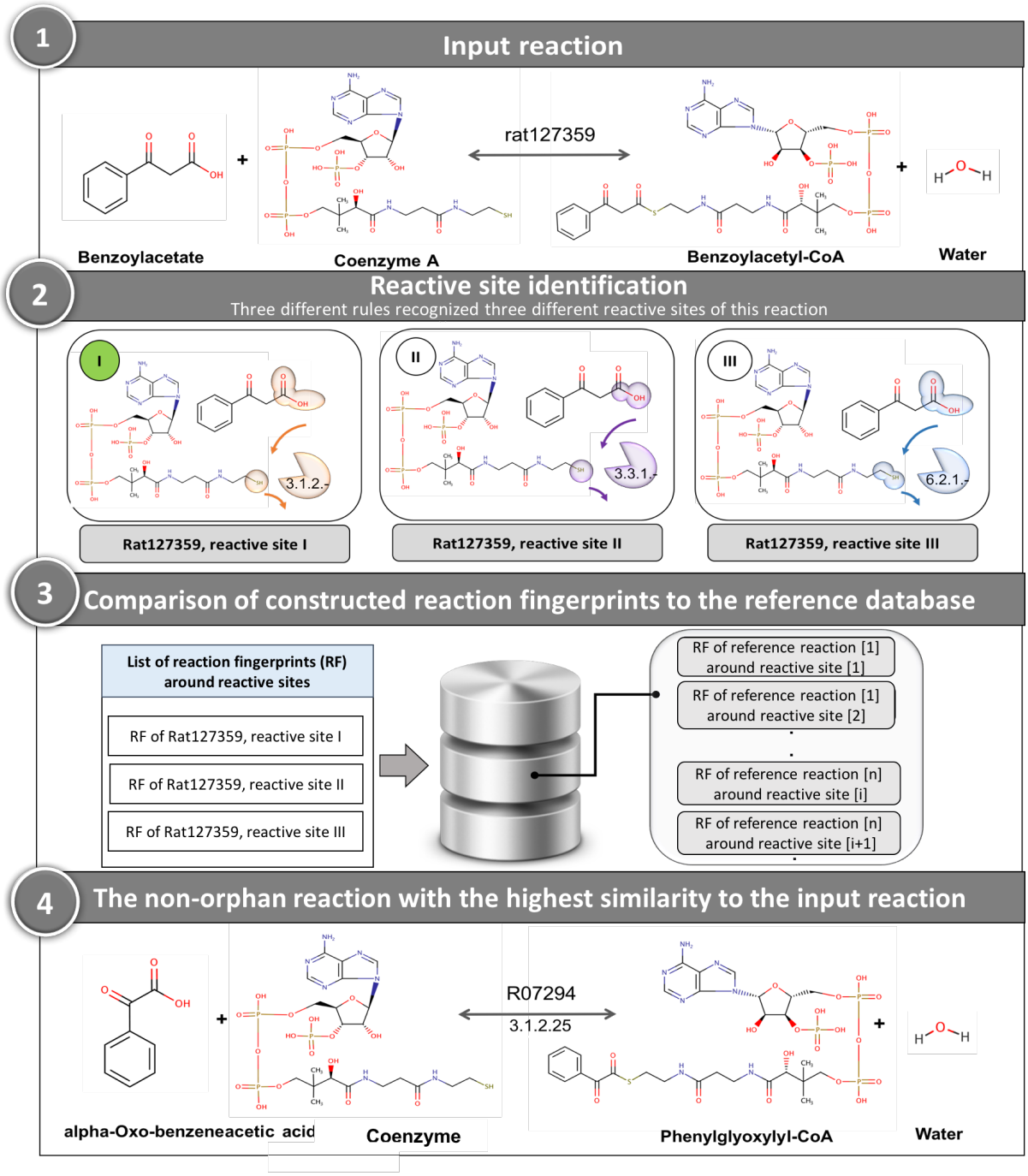

Figure 8. Details of the BridgIT procedure applied to a novel ATLAS reaction.

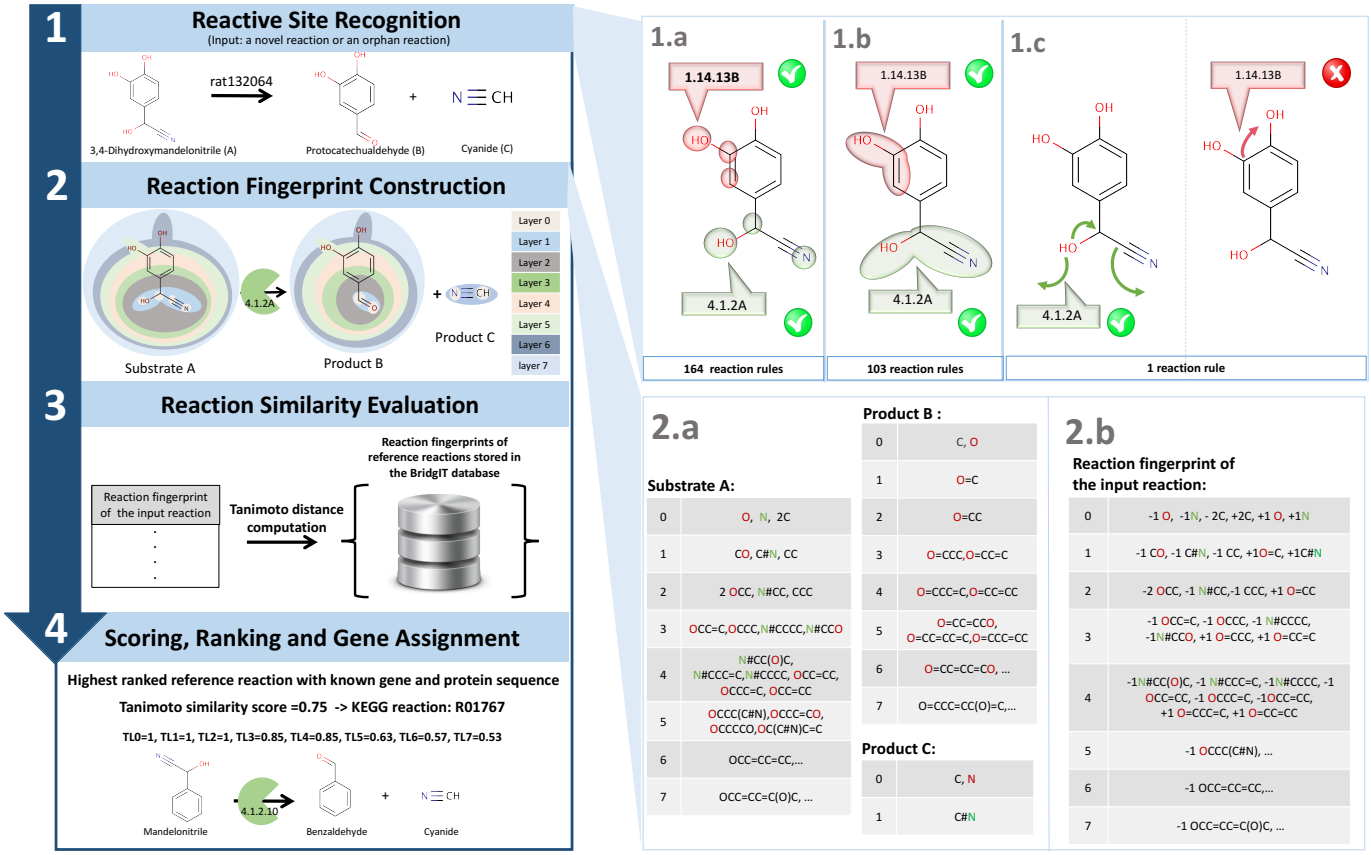

Figure 9. Main steps of the BridgIT workflow: (1) reactive site recognition for an input reaction (*de novo* or orphan); (2) reaction fingerprint construction; (3) reaction similarity evaluation; and (4) sorting, ranking and gene assignment. Panels 1.a to 1.c illustrate the procedure of the identification of reactive sites for the *de novo* reaction rat 132064. Panel 1.a: Two candidate reactive sites of 3,4-dihydroxymandelonitrile (substrate A) that were recognized by the rules 4.1.2. (green) and 1.14.13 (red). Panel 1.b: Both rules recognized the connectivity of atoms within two candidate reactive sites. Panel 1.c: Only reaction rule 4.1.2. can explain the transformation of substrate A to products. Panel 2.a shows the fragmentation of reaction compounds, whereas panel 2.b illustrates the mathematical representations of the corresponding BridgIT reaction fingerprints.