npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-024-00381-1

Extending PROXIMAL to predict degradation pathways of phenolic compounds in the human gut microbiota

Check for updates

Francesco Balzerani[1], Telmo Blasco[1], Sergio Pérez-Burillo[1], Luis V. Valcarcel[1,2,3], Soha Hassoun[4,5] ✉ & Francisco J. Planes[1,2,3] ✉

Despite significant advances in reconstructing genome-scale metabolic networks, the understanding of cellular metabolism remains incomplete for many organisms. A promising approach for elucidating cellular metabolism is analysing the full scope of enzyme promiscuity, which exploits the capacity of enzymes to bind to non-annotated substrates and generate novel reactions. To guide time-consuming costly experimentation, different computational methods have been proposed for exploring enzyme promiscuity. One relevant algorithm is PROXIMAL, which strongly relies on KEGG to define generic reaction rules and link specific molecular substructures with associated chemical transformations. Here, we present a completely new pipeline, PROXIMAL2, which overcomes the dependency on KEGG data. In addition, PROXIMAL2 introduces two relevant improvements with respect to the former version: i) correct treatment of multi-step reactions and ii) tracking of electric charges in the transformations. We compare PROXIMAL and PROXIMAL2 in recovering annotated products from substrates in KEGG reactions, finding a highly significant improvement in the level of accuracy. We then applied PROXIMAL2 to predict degradation reactions of phenolic compounds in the human gut microbiota. The results were compared to RetroPath RL, a different and relevant enzyme promiscuity method. We found a significant overlap between these two methods but also complementary results, which open new research directions into this relevant question in nutrition.

Metabolism is defined as the whole set of chemical reactions that take place in organisms[1]. In particular, metabolic pathways represent chemical transformations where a substrate becomes a product, typically with the aid of other molecules such as cofactors[2,3]. Recent advances in sequencing technologies have significantly increased the coverage of metabolic pathways in dozens of organisms. Much effort has been done to integrate these metabolic pathway into genome-scale metabolic models (GEMMs)[4], which aim to accurately define the stoichiometry of all reactions in a particular organism, their associated genes, enzymes or transporters, their compartment localisation and other relevant biological information. GEMMs allow us to analyse the metabolic capabilities of both unicellular and multicellular systems with computational tools developed in the field of constraint-based modelling[5]. Despite these advances, the understanding of cellular metabolism in many organisms is still incomplete, with significant gaps and metabolites that have no links to any reaction in available metabolic models[6,7]. For example, even in the well-annotated KEGG database[8], there remains 10,000 metabolites that are still not linked to a known biochemical reaction[6].

Different computational tools have been developed to fill in metabolic gaps by means of improving the functional annotation of enzymes. The number of annotated enzymes across databases is much lower than that of metabolic reactions, which could be pointing towards enzymes carrying out more than one biochemical reaction. In KEGG, while we have 11,822 reactions, we can only find 8012 enzymes (https://www.kegg.jp/kegg/docs/statistics.html). The same pattern is found in BRENDA[9], where the numbers are 21,665 and 8332, respectively. Similarly, in AGORA[10], a repository of metabolic networks that include 818 organisms from the human gut microbiota, we found a total count of 1438.1 reactions on average per organism. In the case of enzymes, considering the genomic annotation from GenBank[11]

[1]University of Navarra, Tecnun School of Engineering, Manuel de Lardizábal 13, 20018 San Sebastián, Spain. [2]University of Navarra, Biomedical Engineering Center, Campus Universitario, 31009 Pamplona, Navarra, Spain. [3]University of Navarra, Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI), Campus Universitario, 31080 Pamplona, Spain. [4]Department of Computer Science, Tufts University, Medford, MA 02155, USA. [5]Department of Chemical and Biological Engineering, Tufts University, Medford, MA 02155, USA. ✉e-mail: soha.hassoun@tufts.edu; fplanes@tecnun.es

and Ensembl[12], we obtained a total of 900.7 enzymes on average per organism in AGORA. Considering this evidence, annotating enzyme promiscuity seems a promising strategy to improve metabolic networks.

Promiscuous activity of enzymes lies on their capacity to bind to non-canonical substrates and catalyse novel reactions[13,14]. Annotating such capabilities provides the possibility to understand underground metabolism, which is not represented in current databases[15,16]. The growing relevance on enzyme promiscuity in different fields has led to the development of a number of algorithms and computational methods[6,17–19]. In brief, these algorithms rely on a set of generic enzymatic reaction rules, which define the chemical transformations that occur to a substrate, describing its reactive site and atomic rearrangement as a result of the reaction[6,20]. These reactions rules describe an abstraction of known reactions and permit a certain degree of flexibility of the substrates involved[21], potentially leading to new reactions and products.

Several methods use manually curated reaction rules[22–25]. Although these rules integrate the best knowledge about enzymes, they are limited to a reduced number of reactions[21]. For this reason, the development of computational tools able to automatically extract reaction rules from known transformations has received much attention. Considerable progress has been made in recent years[19,21,26,27]. RetroRules[26] provides thousands of rules that are extracted from public databases and constitutes the core of different algorithms for predicting novel metabolic pathways, such as RetroPath RL[20], the latest version of a series of works developed by the same authors[20,26,28,29].

In a previous effort to elucidate the metabolism of phenolic compounds in the human gut microbiota, we applied RetroPath RL to predict phenolic degradation pathways in the human gut microbiota. As more than 2/3 of phenolic compounds in the Phenol Explorer database[30] are not included in universal metabolic databases, it was necessary to employ computational approaches to uncover microbial phenol metabolism. Despite identifying degradation pathways for 80 compounds in the Phenol Explorer database that were not present in previous gut microbiota reconstructions, RetroPath RL could not find candidate pathways for 180 out of 372 of phenolic compounds in the Phenol Explorer database. Continuing this early work, we focus here on PROXIMAL[27], an enzyme promiscuity algorithm that follows a different strategy to build reaction rules and, thus, could potentially complement the results obtained with RetroPath RL.

The PROXIMAL algorithm has been successfully applied to create extended metabolic models in different organisms and to annotate cellular products[7,31]. PROXIMAL makes use of the KEGG database to predict possible transformations[8]. In particular, it is based on RPAIRS[32], a database available in KEGG that provides the necessary alignment between paired substrates and products to define the modified sub-structures. The main limitation of this algorithm is that it cannot be used for transformations not included in KEGG. Moreover, RPAIRS was discontinued in 2016 (https://www.genome.jp/kegg/kegg1a.html), which hampers the application of PROXIMAL to more recent updated versions of KEGG, that is in continuous development. Overall, these limitations restrict the application of PROXIMAL to our problem of phenolic compound degradation in the human gut microbiota, since we rely on AGREDA[33], a metabolic reconstruction that contains relevant reactions not included in KEGG.

To address these issues, we present here a completely new pipeline, called PROXIMAL2, which overcomes the dependency on KEGG data. Moreover, PROXIMAL2 extends the previous methodology for the automatic reaction rule generation, which was unable to correctly model complex reactions involved in the phenolic compound metabolism. In particular, PROXIMAL2 introduces two relevant improvements with respect to the former version: i) correct treatment of multi-step reactions and ii) tracking of electric charges in the transformations. We show that PROXIMAL2 substantially extends the chemical space of PROXIMAL, and it correctly generates a higher number of reaction rules in KEGG. We also present the application of PROXIMAL2 to predict degradation pathways of phenolic compounds in the human gut microbiota and compare the results with RetroPath RL.

Results

PROXIMAL2 is a rule-based method for the prediction of metabolic products. As in the previous version of the algorithm, PROXIMAL[27], PROXIMAL2 defines the set of reaction rules as look-up tables derived from substrate-product pairs. These look-up tables comprise 2 parts: 'key' and 'value'. In brief, key tables specify the modified substrate structure, particularly defining the reaction centre, atom where the chemical transformation occurs; and value tables describe the modifications resulting in the product. Once these look-up tables are defined for every substrate-product pair, PROXIMAL searches for subgraphs in the query compound matching with key tables and applies the associated transformation defined in the value tables in order to generate putative products.

As detailed in Methods section, for the automatic generation of reaction rules, PROXIMAL2 implements an open chemoinformatic strategy, mainly based on the Python package RDKit[34], which overcomes the dependency of PROXIMAL on the discontinued database RPAIRS and reactions in KEGG. In particular, for the identification of reaction centres, a crucial step in the definition of look-up tables, PROXIMAL2 conducts the necessary atomic alignment of substrates and products via the Maximum Common Substructure algorithm available in RDKit, which additionally allows us to deal with compounds involving rings and charged atoms, features of particular interest in phenolic compound metabolism. Moreover, PROXIMAL2 extends the scope of PROXIMAL, which is limited to reactions with a single reaction centre, making possible to deal with complex transformations involving multiple reaction centres that are typically found in phenolic compound metabolism. Full details of the PROXIMAL2 pipeline can be found in Methods section.

To illustrate the improvement of PROXIMAL2 over PROXIMAL in the automatic generation of reaction rules, we present below a side-by-side comparison between both approaches. Then, we apply PROXIMAL2 to predict degradation pathways of phenolic compounds in the human gut microbiota and compare the results obtained with a different enzyme promiscuity approach, RetroPath RL.

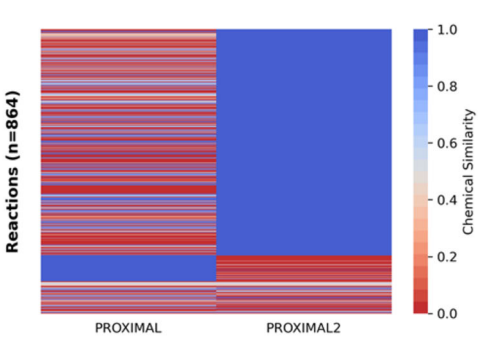
Comparison between PROXIMAL and PROXIMAL2

We first evaluated the accuracy of PROXIMAL2 in recovering annotated products from substrates in KEGG reactions, in comparison with the previous PROXIMAL algorithm (https://hassounlab.cs.tufts.edu/proximal/). Therefore, we applied PROXIMAL2 to the same set of KEGG reactions used by PROXIMAL. This subset of reactions involves 8819 reactions and 4983 associated metabolites (Supplementary Tables 1 and 2). In brief, the validation strategy followed here consists of generating the key and value tables for every substrate (S)-product (P) pair, following the methodology presented in the Methods section, and then assess if the product P is obtained when the same substrate S is applied as a query compound to their associated key and value tables (Supplementary Fig. 1). Rules that satisfy this requirement correctly captures the underlying chemistry of their associated reactions. We compared the accuracy of PROXIMAL2 with PROXIMAL under 3 different scenarios (see Fig. 1a).

First, we analysed the performance of PROXIMAL2 with the same limitations as in PROXIMAL. In particular, we only considered those reactions with a single reaction centre, reducing the study to 5991 reactions, and neglected the atom charge information from the predictions of PROXIMAL2. Under this scenario (Case 1 in Fig. 1a), PROXIMAL2 was able to generate the proper product for 5574 out of 5991 reactions (accuracy: 93%), while the predictions were correct for 5290 reactions in PROXIMAL (accuracy: 88.3%). This result illustrates that our chemoinformatic strategy in PROXIMAL2 (Steps 1–3), which overcomes the dependence on RPAIRS, produces more accurate results than PROXIMAL. In order to evaluate the effect of considering atom charge in PROXIMAL2, we updated the previous comparison and included this feature in our analysis. We now reached the proper product for 5814 reactions, obtaining an accuracy of 97% (Case 2 in Fig. 1a). This shows that the effect of atom charge in PROXIMAL2 further increase the accuracy of PROXIMAL2. Moreover, these results imply a relevant reduction of false positives in PROXIMAL2 with respect to

a

	KEGG reactions with single reaction center (5991)		Whole list of KEGG reactions (8119)
	Case 1	Case 2	Case 3
PROXIMAL	88,30%	88,30%	65,20%
PROXIMAL2	93,10%	97,10%	95,40%



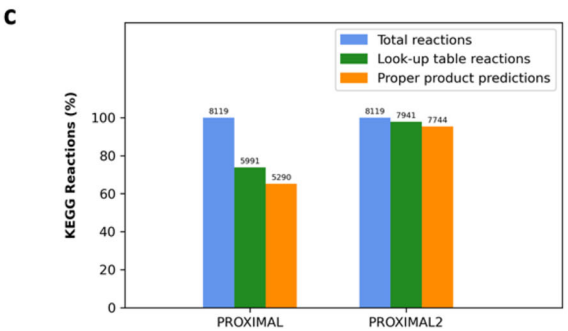


Fig. 1 | Comparison between PROXIMAL and PROXIMAL2. a Percentage accuracy in recovering annotated products in KEGG for PROXIMAL and PROXIMAL2 in three different scenarios: Case1, Case2 and Case3. In Case1 and Case2, we consider reactions with a single reaction centre. We do not consider atom charge information in Case1, but we do in Case2. In Case3, we consider all the reactions in the KEGG version used in PROXIMAL. The number of reactions used in each of the cases is shown in parenthesis. **b** Heatmap representing the chemical similarity of the predicted and annotated products in KEGG in Case 1 and 2 where at least one of the two algorithms fail to predict the annotated product. **c** Barplot representing the reaction coverage with look-up tables and those correctly predicting the annotated product in KEGG reactions. The y- axis shows the coverage in percentage. The total number of reactions are indicated over the bars.

PROXIMAL, since the number of incorrect predictions of annotated products is reduced by 8.7%.

To visualise the differences between PROXIMAL and PROXIMAL2 in Case 2, we extracted the molecules for which at least one of the two algorithms were not able to produce the correct product and computed the chemical similarity values between the predicted and annotated product (Fig. 1b). It can be observed that PROXIMAL2 is substantially more accurate than PROXIMAL, finding the annotated product (chemical similarity = 1) in many more cases. Note here that in both cases, Case 1 and Case 2, the improvement of PROXIMAL2 is highly significant (two proportions z-test p-value ≤ 2e−16).

In the last scenario, we included multi-centre reactions in our analysis and considered the whole set of 8119 reactions. PROXIMAL could not improve the accuracy, since it is not able to model multi-centre reactions, obtaining an accuracy of 65.2%. PROXIMAL2, instead, generated the correct product for 7744 reactions (accuracy: 95.4%), which illustrates the clear advance brought by PROXIMAL2 (Case 3 in Fig. 1a). Note here that we could only generate look-up tables for 7941 out of 8119 reactions with PROXIMAL2 (Fig. 1c) and, thus, the accuracy is even higher for these subset of reactions (97.5%). In both cases, the improvement of PROXIMAL2 is statistically significant (two proportions z-test p-value ≤ 2e−16). PROXIMAL2 was not able to generate look-up tables for the remainder 178 reactions mainly due to the incapacity to deal with stereochemical information and the restrictions of atom charge imposed in the definition of MCS (see Methods section).

In the construction of the Maximum Common Substructure (MCS) in PROXIMAL2 we fixed 2 optional parameters related atoms belonging to ring and atom charges (see Step 2 in Methods section). We carried out a sensitivity analysis to evaluate the effect of these heuristic choices in the outcome of PROXIMAL2. In particular, we considered the 4 possible cases, namely when both parameters are fixed, the two cases where only one of the parameters is fixed and none of the parameters are fixed. Supplementary Fig. 2 shows that best performance is obtained in the case that both parameters are considered, which justifies the use of these parameters in PROXIMAL2.

Finally, we assessed the robustness of the metabolic space generated in PROXIMAL and PROXIMAL2 by means of a Leave-One-Out strategy. In particular, for each annotated substrate (S)-product (P) pair, we evaluated whether P is recovered when S is applied as a query compound to any reaction rule except from the one they generate. PROXIMAL2 was able to recover the annotated products in 36% of the whose set of reactions, whereas PROXIMAL only 11%, which again emphasizes the improvement of PROXIMAL2 over PROXIMAL. This relevant difference is observed due to the greater capacity of PROXIMAL2 over PROXIMAL to generate correct reaction rules, as observed in Fig. 1c, which benefits its ability to recover leave out cases.

Application of PROXIMAL2 to predict the degradation of phenolic compounds in the human gut microbiota

Phenolic compounds are potent antioxidants that are derived from foods of plant origin and are mainly metabolised by the human gut microbiota. Despite increasing interest in the health and nutrition literature on phenolic compounds, their metabolism remains largely unknown. Universal metabolic databases, such as KEGG[8] or the Model SEED database[35], store reactions from species not present in the gut microbiota, and pathway extraction is not direct. For this reason, automatic reconstruction pipelines, such as AGORA[10] or CarveMe[36], include a limited number of phenolic compounds in curated genome-scale models of the human gut microbiota. Previously, we addressed this problem using a combination of computational methods, manual annotation and expert knowledge, leading to an improved version of AGORA, called AGREDA[33]. Moreover, we extended AGREDA using RetroPath RL, a popular enzyme promiscuity algorithm[37]. However, we could not find candidate pathways for 180 out of 372 of phenolic compounds in the Phenol Explorer database (Rothwell et al.[30]) and, thus, complementary approaches are required. In this sub-section, we continue

our previous works by applying PROXIMAL2 to predict degradation pathways of phenolic compounds in the human gut microbiota and comparing the results obtained with RetroPath RL.

RetroPath RL is a rule-based method that makes use of the RetroRules database[26] to investigate enzyme promiscuity. RetroRules defines the reaction centre based on an atom-atom mapping between substrates and product atoms and compute reaction rules using the reaction SMARTS (SMILES Arbitrary Target Specification) formalism. The level of specificity of reaction rules in RetruRules can be tuned according to the diameter parameter, i.e. the size of a hypothetic sphere around the reaction centre. Moreover, RetroPath RL allows the user to fix the threshold of chemical similarity between the query compound and the substrate of the rule. To apply RetroPath RL in the most similar conditions to PROXIMAL2, we fixed the diameter at 4 and the chemical similarity threshold at 0.6. Moreover, we defined the same metabolic space for both approaches, namely the one introduced in our previous work[37], which involves 5087 reactions taken from AGREDA and the Model Seed Database (Supplementary Table 3).

A high proportion of reactions considered in the metabolic space of AGREDA were present in the RetroRules database; however, we also included some manually curated reactions important for the metabolism of phenolic compounds[33,37]. PROXIMAL2 only requires the biochemical reaction equation to generate the look-up tables (reaction rules). For the reactions present in RetroRules, we extracted the equations from MetaNetX database, whereas the equations for the manually curated reactions were obtained from AGREDA. Following the complete pipeline of PROXIMAL2, we could generate look-up tables (reaction rules) for 4860 reactions (Fig. 2a). Reaction rules for RetroPath RL, in contrast, were directly obtained from the RetroRules database at diameter 4. For manually curated reactions, we had to create them one by one using the RetroRules webpage (https://retrorules.org/diy). As a result, we generated reaction rules for 5064 reactions with RetroPath RL (Fig. 2a). Overall, PROXIMAL2 shows a higher level of automatisation to generate reaction rules.

We compared the reactions rules obtained with PROXIMAL2 and RetroPath RL, finding that 4837 were present in both cases, whereas 23 and 227 were unique to PROXIMAL2 and RetroPath RL, respectively (Fig. 2a). The main differences between both approaches in the generation of rules are due to several reasons: (i) different treatment of cofactors (29.5%); (ii) ability of RetroPath RL to deal with stereochemistry (25.1%); (iii) reactions no longer present in the MetaNetX database, i.e. impossibility to download the reaction equation from MetaNetX and, consequentially, they were not considered with PROXIMAL2 (16.7%); iv) differences in the definition of MCS and reaction centres, e.g. the mandatory condition in the extraction of MCS in PROXIMAL2 (but not in RetroPath RL) that a match between atoms included in rings can happen only if the atoms are part of rings in both substrate and product (15%). As a result, RetroPath RL appears to cover a slightly wider area of the chemical space than PROXIMAL2.

Then, we applied the generated rules with both approaches to 372 phenolic compounds (Supplementary Table 4) obtained from Phenol-Explorer database[30]. Since RetroPath RL provides all reaction products that can be generated from a query compound and PROXIMAL2 generates one product at a time, we filtered out the results for which PROXIMAL2 was not able to predict the whole set of putative products. We obtained results for 354 out of 372 phenolic compounds using PROXIMAL2, whereas RetroPath RL generated products for 323 (Fig. 2a). Specifically, 319 of them were in common, while 35 were specific for PROXIMAL2 and 4 for RetroPath RL (Fig. 2b). The differences observed between PROXIMAL2 and RetroPath RL are mainly caused by two factors: i) a different filter of chemical similarity for the query compound, namely PROXIMAL2 uses the Dice coefficient and RetroPath RL the Tanimoto coefficient; ii) differences in the definition of MCS, noted above, which determines a different reaction centre and local neighbourhood.

These differences are even higher if we concentrate on the predicted reactions with these methods. In this case, we found an overlap between PROXIMAL2 and RetroPath RL of only 30%. For this comparison, we considered the same number of predicted products for each phenolic compound in both cases. In particular, we extracted the number of predicted reactions by RetroPath RL for each phenolic compound and fixed this same number of predicted reactions for PROXIMAL2, according to chemical similarity with the annotated products. We assumed that both approaches reached the same product from phenolic compounds when the chemical similarity (Dice coefficient) was equal to 1. This result, in our opinion, emphasizes the complementarity of these approaches.

In our previous work[37], we applied RetroPath RL to the same 372 phenolic compounds and extracted putative products for 303 of them using the recommended diameters by the authors (more than 6). Here, PROXIMAL2 could generate putative products for 53 additional phenolic compounds (301 were in common between PROXIMAL2 and RetroPath RL under this scenario). These 53 phenolic compounds were connected to 430 metabolites involved in AGREDA. Interestingly, we found that *Lignans* and *Isoflavonoids* were highly represented in this set of metabolites, namely 16 and 9 metabolites, respectively (Fig. 2c). In addition, according to the food composition provided by Phenol-Explorer, the 53 phenolic compounds are part of 8 sub-groups of food, with *Soy and soy products* and *Fruits – Berries* the most annotated (Fig. 2d). Moreover, we analysed the taxonomies of predicted reactions for the degradation of this subset of 53 phenolic compounds. Figure 2e shows the contribution of different taxonomic classes to the reactions involved in the degradation of lignans and isoflavonoids (see Supplementary Table 5 for details). It can be observed that the most relevant classes are Bacilli, Clostridia, Bacteroidia, Gammaproteobacteria and Actinobacteria, in line with other phenolic compounds previously annotated in AGREDA. The same analysis at the species level can be found in Supplementary Table 6.

Overall, these results show that PROXIMAL2 complements our previous analysis with RetroPath RL and opens new research directions to understand the metabolism of phenolic compounds in the human gut microbiota. A complete biological validation of the reactions and metabolites predicted by PROXIMAL2 requires further analysis and data. However, for illustration of the relevance of the results obtained, we present below some specific cases for which PROXIMAL2 (and not RetroPath RL) leads to biologically meaningful hypotheses.

First, we found different *glucuronide* species where PROXIMAL2 predicts the loss of *glucuronic acid*. One example can be observed in Fig. 2f, where the degradation of *phloretin_2'-O-glucuronide* into *phloretin* and *glucuronic acid* is shown. This metabolite is present in apples and derived foods, such as apple juices[38]. The output product, *phloretin*, has been shown to exert anti-inflammatory activity in different diseases via the gut microbiota, e.g. ulcerative colitis[39]. The template glucuronidase enzyme, from which this reaction is predicted, can be found in 774 metabolic models in AGREDA, particularly among the following classes: *Bacilli*, *Clostridia* and *Bacteroides*.

Another relevant phenolic compounds for which PROXIMAL2 predicts a degradation reaction is *4-hydroxyhippuric acid*, which is converted into *glycine* and *4-hydroxybenzoic acid* (Fig. 2f). Our predicted reaction has been proposed elsewhere in the literature but in the opposite direction[40]. This could be a plausible hypothesis because *4-hydroxybenzoic acid* is part of a wider set of foods and *4-hydroxyhippuric acid* is more often found in urine samples, according to PhenolExplorer. Interestingly, in contrast with other phenolic compounds, *4-hydroxyhippuric acid* seems to be pro-inflammatory by increasing the level of cytokine TNF-α[41]. The template enzyme from which this reaction is predicted takes part in 51 metabolic models in AGREDA, being *Gammaproteobacteria*, *Alphaproteobacteria* and *Betaproteobacteria* the main relevant classes.

Finally, PROXIMAL2 predicted the degradation of the lignan *cyclolariciresinol*, also called *isolariciresinol*. This phenolic compound is found in different fruits, including blackberry, blackcurrant or strawberries, and sesame seeds, according to PhenolExplorer; however, it has been abundantly identified in wines[42] and soy-based supplements[43], among others. Previous works in the literature have hypothesised the antioxidant activity of *isolariciresinol*[44]. This is consistent with findings that other lignan metabolites, such as *secolariciresinol*, have a protective role against a variety of

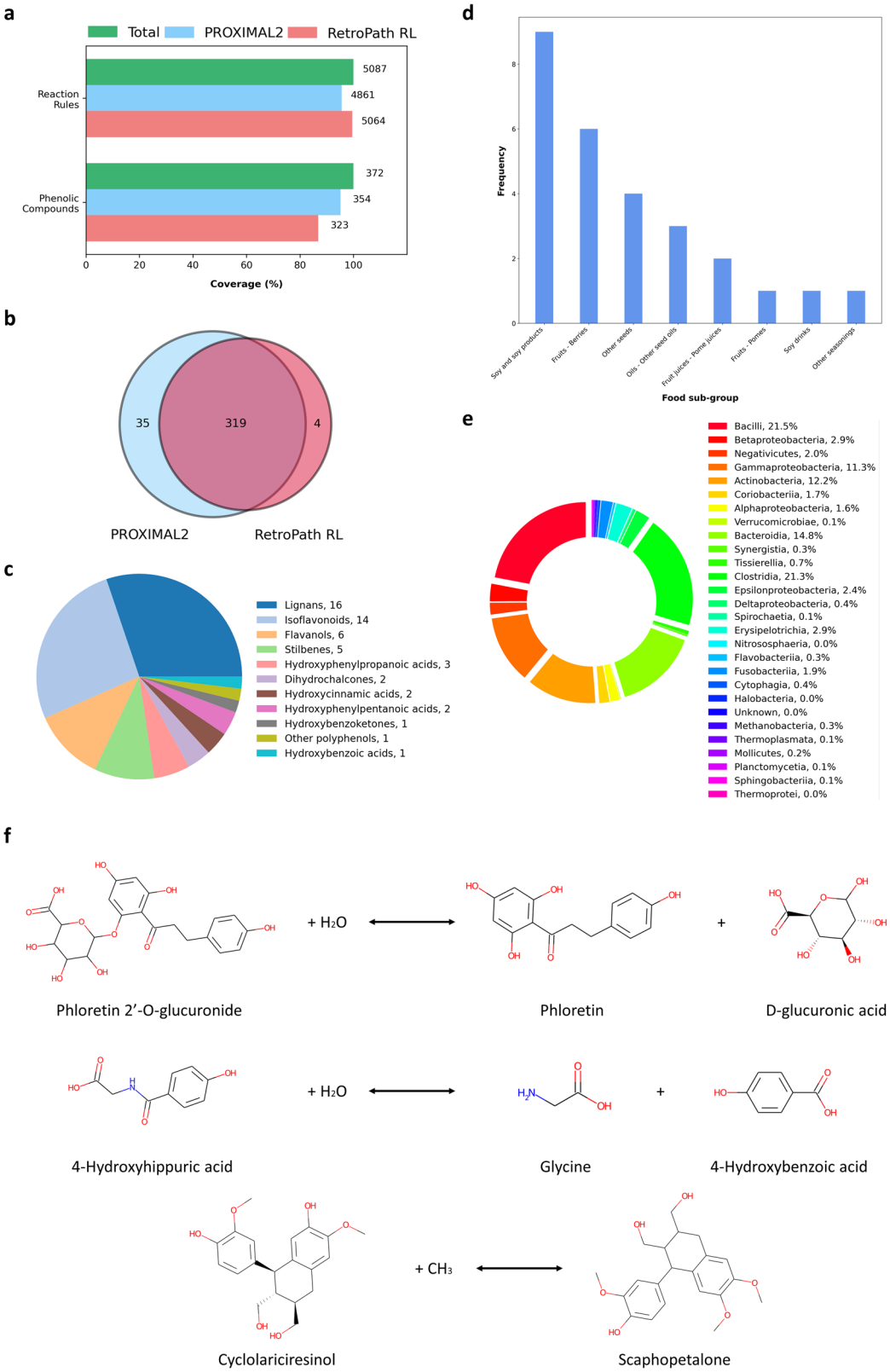


Fig. 2 | Comparison between PROXIMAL2 and RetroPath RL. a Representation of reaction coverage regarding the ability to generate rules by the algorithms and the coverage of phenolic compounds to produce promiscuous products. The number to the right of the bars represents the absolute number of reactions and compounds; **b** Venn diagram of the phenolic compounds that can be potentially degraded by PROXIMAL2 and RetroPath RL; **c** Representation of the different sub-classes of the 53 phenolic compounds captured by PROXIMAL2. The number of compounds belonging to the sub-class is expressed in the legend, e.g. 'Lignans, 16'; **d** Frequency of sub-groups of foods associated with the 53 phenolic compounds exclusively predicted by PROXIMAL2 in comparison with[37]; **e** Taxonomic classes involved in the predicted reactions for lignans and isoflavonoids. The number in the legend represents the contribution of each class to the predicted reactions; **f** 3 reactions predicted by PROXIMAL2 in the subset of 35 phenolic compounds for which RetroPath RL did not reach a solution.

diseases[45]. With respect to the metabolism of *isolariciresinol*, there is limited evidence in the literature. Here, we propose the transformation of *isolariciresinol* into *scaphopetalone* (Fig. 2f). This metabolite has been identified in different plants[46] but not in human samples so far. However, the activity of *scaphopetalone* derivatives against Herpes simplex and Herpes zostera has been recently demonstrated[47]. The template enzyme from which this reaction is predicted takes part in all metabolic models in AGREDA.

Discussion

Cellular metabolism involves the totality of chemical transformations that can occur in organisms and, though our knowledge is continuously growing, many metabolic pathways are still incomplete. A relevant case in the field of nutrition is the metabolism of phenolic compounds in the human gut microbiota, which remains largely unknown. Phenolic compounds, which are mainly derived from foods of plant origin[48,49], are converted into bioactive metabolites that appear to limit the risk of several major diseases, such as coronary heart disease[50], cancer[51] or diabetes[52]. This fact has stimulated research to complete the knowledge about the degradation pathways of these nutrients in the human gut microbiota.

Recently, several methodologies have been developed to fill in metabolic gaps. The computational analysis of the metabolic space through enzyme promiscuity has received much attention. Specifically, rule-based methods have grown in number and quality in the last years[21]. In a previous work[37], we applied a well-known rule-based enzyme promiscuity algorithm, RetroPath RL, to predict the degradation pathways of 372 phenolic compounds from Phenol-Explorer. Here, we explore a different rule-based algorithm, PROXIMAL, to address the same question. Given the current limitations and KEGG dependencies of PROXIMAL, we developed PROXIMAL2, which can automatically generate rules for a wider spectrum of reactions and make more reliable and comprehensive predictions for the degradation of phenolic compounds in the human gut microbiota.

As detailed in Methods section, without KEGG dependencies, PROXIMAL2 automatically extracts reactions rules without relying on the KEGG database, and replicates the look-up tables defined in PROXIMAL for predicting novel reactions. Further, PROXIMAL2 includes new features that were not part of PROXIMAL and expands its scope of application. In particular, PROXIMAL2 is able to capture complex transformations that involves multi step reactions through the development of multi-centres look-up tables. In addition, a detection of possible atom charges was implemented in PROXIMAL2, which allows tracking the charges present in substrates and products and adding or removing charges depending on the transformation. These new features of PROXIMAL2 significantly improved the performance of PROXIMAL, as described in the Results section with the comparison with KEGG reactions. PROXIMAL2 increased the coverage of KEGG reactions that can be potentially used for predicting enzyme promiscuity, namely 1950 KEGG reactions can be considered with PROXIMAL2 but not with PROXIMAL. Second, PROXIMAL2 shows higher accuracy than PROXIMAL in recovering annotated products from the substrates in KEGG reactions in the different scenarios considered in Fig. 1a.

PROXIMAL2 was applied to predict novel degradation routes of phenolic compounds in the human gut microbiota. Note here that this study had not been feasible with PROXIMAL, given its KEGG dependences and the complexity of annotated reactions in AGREDA for phenolic compounds. We compared the results of PROXIMAL2 with RetroPath RL. With respect to rule generation, we found that PROXIMAL2 is more practical in the automatic extraction of reaction rules, as illustrated with the subset of manually curated reactions included in the metabolic space. On the other hand, RetroPath RL covers a wider range of chemical space, obtaining reaction rules for 227 reactions for which PROXIMAL2 could not obtain look-up tables. Although the difference is limited, we expect to address these limitations in future developments.

In addition, we found that PROXIMAL2 and RetroPath RL have a significant overlap in the subset of phenolic compounds for which they could find a putative product. Both algorithms obtained a promiscuous product in 319 out of 372 compounds (85.7%). However, we found PROXIMAL2 and RetroPath RL complementary; PROXIMAL2 predicted new output products for 35 phenolic compounds that were not captured by RetroPath RL, which represents an 9% increase in identification of phenolic compounds. We found that these differences are dependent on the choice of specific parameters, e.g. diameter in RetroPath RL or chemical similarity coefficient used in both approaches, which emphasizes the importance of investigating the optimal set of parameters for different algorithms. Importantly, our results open new research directions in the metabolism of phenolic compounds in the human gut microbiota. We expect to further analyse and experimentally validate the results obtained from this study.

Finally, beyond the problem of phenolic compound degradation in the human gut microbiota, PROXIMAL2 is a general-purpose algorithm and can be applied to predict novel associations of other molecules of interest to putative products. In fact, the previous version of PROXIMAL2, PROXIMAL, has been previously applied to a variety of biological questions, including the prediction of xenobiotic metabolism[27] to create extended metabolic model of *E. coli*[7], and to suggest biological molecular candidates when annotating metabolomics data[31]. As PROXIMAL2 has the same features than PROXIMAL, it can be complementarily used with other existing rule-based methods, such as RetroPath RL, to provide novel insights into metabolic gaps for different applications.

Methods

PROXIMAL[27] is a rule-based method for the prediction of metabolic products. PROXIMAL defines look-up tables for a database of substrate-product pairs. These look-up tables link specific molecular substructures with their associated chemical transformations (Fig. 3a). In particular, 'keys' in the look-up tables specify the modified substrate structure, including: i) the reaction centre, atom where the chemical transformation occurs; ii) adjacent neighbours, atoms connected to the reaction centre at distance 1; iii) distant neighbours, atoms connected to the reaction centre at distance 2. In other words, the modified substrate structure induces a subgraph of neighbours within radius 2 starting from the reaction centre. In addition, 'values' in the look-up tables describe the modifications resulting in the product, including: i) reaction centre, ii) adjacent neighbours and iii) added/removed functional group, which defines the modified part of the substrate. Note here that the atoms in the look-up tables follow the nomenclature of 'KEGG atom types', which are defined according to their functional groups and microenvironment, e.g. C8x or C8y.

Once these look-up tables are defined for all substrate-product pairs involved in the database of reactions of interest, PROXIMAL lists the different subgraphs of neighbours within radius 2 for a given query compound, searches for the ones matching with the sub-structures stored in the key tables and applies their associated transformation defined in the value tables in order to generate putative products. Figure 3b illustrates that the subgraph of neighbours centred at atom number 8 in the query compound L-Tyrosine matches with the key table defined in Fig. 1a for the pair 4-Hydroxybenzoate and 3,4-Dihydroxybenzoate, leading to the compound 3,4-Dihydroxy-L-phenylalanine.

The identification of the reaction centre for each substrate-product pair is the critical step in the definition of look-up tables. This is done following different steps that require i) the alignment of the substrate and product atoms and ii) the classification of different atoms into different 'KEGG atom types' (Fig. 3a). Reaction centres are defined as any specific substrate atom that aligns with a product atom of different KEGG atom type. In Fig. 1a, the reaction centre corresponds with the atom number 8 of 4-Hydroxybenzoate and the atom number 8 of 3,4-Dihydroxybenzoate, whose KEGG atom types are C8x and C8y, respectively.

The look-up tables in PROXIMAL were built using the RPAIRS database, where substrate-product pairs are defined for each reaction, atomic species are classified into KEGG atom types, and substrate atoms are aligned to product atoms. Unfortunately, as noted above, RPAIRS database was discontinued in 2016, which restricts its application to a limited subset of reactions and, thus, metabolic space (i.e. set of potential biochemical

a

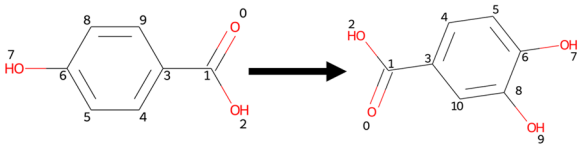
Look-up table definition

key	Reaction centre	Adjacent neighbours	Distant neighbours
	C8x	C8x, C8y	C8x, C8y, C8x, O1a, C8x
value	Reaction centre	Adjacent neighbours	Added functional group
	C8y	C8x, C8y	O1a

Substrate and product alignment

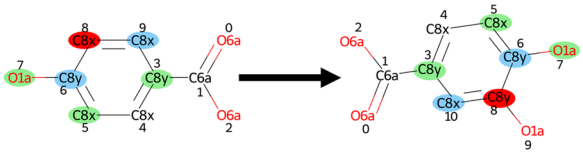
Substrate		Product	
Atom number	Atom type	Atom number	Atom type
0	O6a	0	O6a
1	C6a	1	C6a
2	O6a	2	O6a
3	C8y	3	C8y
4	C8x	4	C8x
5	C8x	5	C8x
6	C8y	6	C8y
7	O1a	7	O1a
8	C8x	8	C8y
9	C8x	10	C8x

Substrate-product pairs



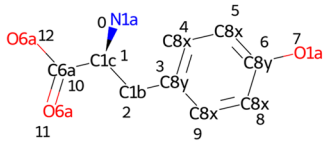
4-Hydroxybenzoate 3,4-Dihydroxybenzoate

KEGG atom type definition



b

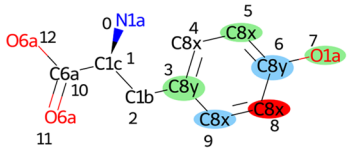
Query representation



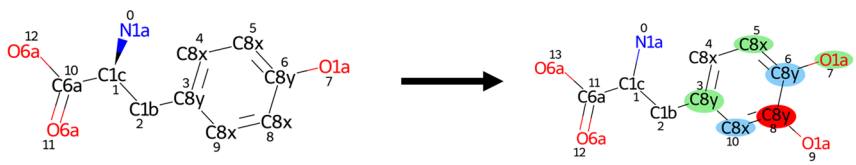
L-Tyrosine

Query			
Atom number	Reaction centre	Adjacent neighbours	Distant neighbours
0	N1a	C1c	C6a, C1b, N1a
1	C1c	N1a, C6a, C1b	C1c, O6a, O6a, C1c, C8y, C1c
2	C1b	C1c, C8y	C6a, N1a, C1b, C8x, C8x, C1b
3	C8y	C1b, C8x, C8x	C1c, C8y, C8y, C8x, C8y, C8x
4	C8x	C8y, C8x	C1b,C8x, C8x, C8y, C8x
5	C8x	C8y, C8x	C8x, O1a, C8x, C8y, C8x
6	C8y	C8x, O1a, C8x	C8x, C8y, C8y, C8x, C8y
7	O1a	C8y	C8x, O1a, C8x
8	C8x	C8x, C8y	C8x, C8y, C8x, Oa1, C8x
...			

Matched key in the query compound

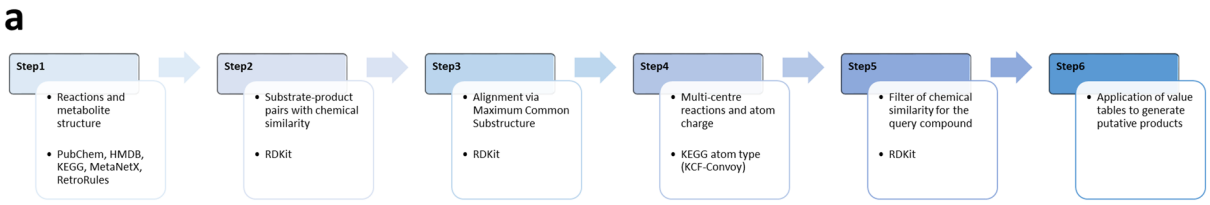


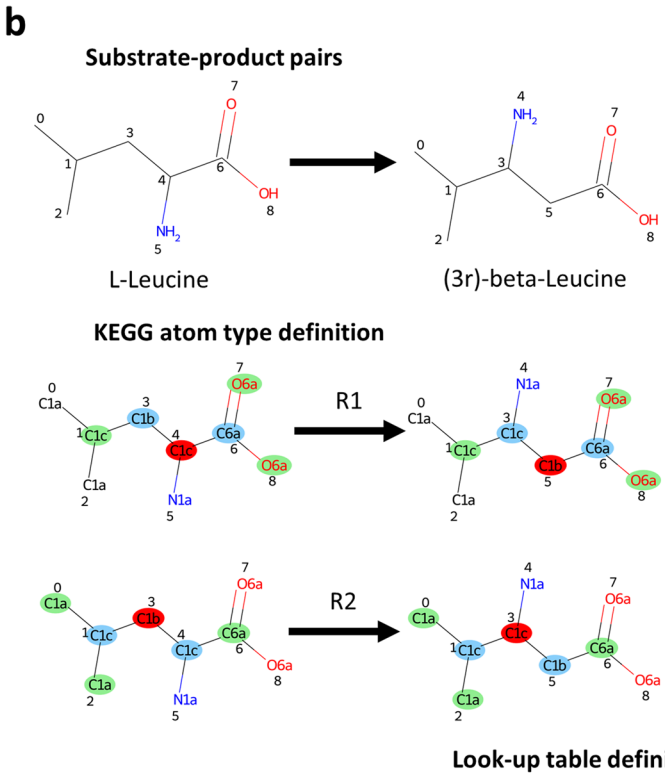
Product generation



L-Tyrosine 3,4-Dihydroxy-L-phenylalanine

Fig. 3 | Schematic representation of the PROXIMAL workflow. a Definition of an example look-up table in PROXIMAL. After the pair selection, the substrate and product are represented in KEGG atom type and then aligned to permit the definition of the reaction centre and its neighbours up to distance 2. The starting reaction is (*R01296*): *4-Hydroxybenzoate+O2+NADH+H(+) -> 3,4-Dihydroxybenzoate +H2O+NAD(+)*. **b** Application of the look-up table to an example selected query. The latter is represented in KEGG atom type and the key of the look-up table is searched within the molecule. Finally, the product is generated based on the look-up table value. The red, blue and green circles represent the reaction centre, adjacent and distant neighbour, respectively. The atoms O and N are marked in red and blue, respectively, as conventionally used in chemistry by the CPK colouring rules.





Substrate and product alignment

Substrate		Product	
Atom number	Atom type	Atom number	Atom type
0	C1a	0	C1a
1	C1c	1	C1c
2	C1a	2	C1a
3	C1b	3	C1c
4	C1c	5	C1b
6	C6a	6	C6a
7	O6a	7	O6a
8	O6a	8	O6a

R1	key	Reaction centre	Adjacent neighbours	Distant neighbours
		C1c	C1b, C6a	C1c, O6a, O6a
	value	Reaction centre	Adjacent neighbours	Removed functional group
		C1b	C1c, C6a	O1a

R2	key	Reaction centre	Adjacent neighbours	Distant neighbours
		C1b	C1c, C1c	C1a, C1a, C6a
	value	Reaction centre	Adjacent neighbours	Added functional group
		C1c	C1c, C1b	N1a

Fig. 4 | PROXIMAL2 workflow and illustration of multi-centre reactions. **a** PROXIMAL2 pipeline summarising main changes with respect to the previous version. **b** Example of multi-centre reactions. The alignment provides two changing KEGG atom types (atom numbers 3 and 4), leading to two reaction centre that are labelled as *R1* and *R2*. In the look-up table, the information about the reaction centres and respective neighbour is stored. The reaction is (*R01091*): *L-Leucine -> (3R)-beta-Leucine*.

transformations and molecules[17]). Therefore, we present here a completely new pipeline, called PROXIMAL2, summarised in Fig. 4a.

Step 1: Definition of the database of reactions, metabolites and structural information

First, we created the metabolic space that is required to apply our enzyme promiscuity algorithm. Starting from an input database of reactions, we first removed the most common cofactors involved within the transformations and deleted the reactions if they only included cofactors. Then, the metabolites involved in the different transformations were extracted. We obtained SMILES and InChI for each metabolite from different public databases: PubChem[53], KEGG[8], HMDB[54], MetaNetX[55] and RetroRules[26]. Metabolites with no available structure were filtered out from the reactions; similarly, input reactions only involving metabolites without structure were deleted. As a result, we obtained a list of simplified reactions.

Step 2: Definition of substrate-product pairs

As noted above, PROXIMAL works with substrate-product pairs. However, the output (simplified) reactions from Step1 can include more than one substrate or product. To identify the best matching pairs for a given reaction,

we calculated the chemical similarity between each substrate-product pair and paired them according to such value. The chemical similarity was determined using the RDKit package[34] and the Morgan Fingerprint[56]. In reactions with a single substrate (or product), the latter is associated to each product (substrate) to form more than one pair. The output of this step is the list of substrate-product pairs for the generation of look-up tables. Note that with this chemoinformatic approach we recovered 97.5% of pairs annotated in RPAIRS.

Step 3: Alignment of substrate-product pairs

As noted above, the identification of the reaction centre and look-up tables for each substrate-product pair requires their atomic alignment. To that end, we extracted the maximum common substructure (MCS) between the substrate and product with the function *findMCS* available in the module *rdFMCS* in RDKit package. We fixed two optional parameters in this RDKit function in order to consider the specificity of the atoms belonging to rings and charged atoms. In particular, a match between atoms included in rings can happen only if the atoms are part of rings in both the substrate and product. Similarly, two atoms can match with each other only when both have the same charge (Supplementary Note 1, Supplementary Fig. 3).

Step 4: Definition of reaction centres and look-up tables

As noted above, reaction centres are defined as any specific substrate atom that aligns with a product atom of different KEGG atom type. Here, the classification of substrate and product atoms into different KEGG atom types were done with the package KCF-Convoy[57,58]. With this information and the atomic alignment (Step3), reaction centres were identified.

A relevant improvement included in PROXIMAL2 is the possibility to analyse reactions containing multiple reaction centres. In some cases, it may happen that a reaction represents a multi-step reaction[59], where the intermediate steps are removed for any reason (e.g. inability to measure the intermediate), and they are unified in the transformation that connects the initial substrates directly to the final products. This concept translates into having more than one reaction centre when the changing bonds are analysed. For each reaction centre involved in these transformations, look-up tables were defined, as shown in Fig. 4b.

In addition, in some cases, we found a modification of the bonds between the atoms within the MCS, showing a new arrangement due to the transformation. Specifically, during the rearrangement of the structure, a bond can be introduced, deleted or simply changed (e.g. going from a double to single bond). These modifications were also extracted together and included in the look-up tables as multiple reaction centres (see an illustration in Supplementary Fig. 4). Finally, the information about the possible position of a charged atom is extracted in order to apply that charge when the transformation is applied.

Step 5: Search of matching keys within the query compound

Once the look-up tables representing the chemical transformations are generated, we define the subgraph of neighbours within radius 2 for each atom of the query compound and search for matching keys in the look-up tables. For multi-centre reactions we ensure that the query compound matches with all their associated keys. Note here that a pre-filter was implemented to any operator applied to the query compound. In particular, we discard matches where the chemical similarity between the query compound and the substrate of the matching entry was below 0.6.

Step 6: Product generation

Once an operator in the look-up table matches the query compound, we generated the promiscuous product. Therefore, considering the molecule of interest, the transformation is applied to each reaction centre, adding atoms and bonds coherently according to the original template reaction. When there is more than one reaction centre, it can happen that a substructure, which must be added to the query, is in common between several reaction centres. To avoid the addition of that substructure multiple times, we implemented a tracking system of the atoms and bonds already introduced, rejecting duplicate additions. Then, bonds are added, removed or changed coherently to the operator definition. Finally, information about the charges present in the query is kept and introduced in the generated product molecule. If a charged atom was removed due to the transformation, the charge will not be present in the final product. Similarly, when a functional group introduced in the structure contained a charged atom, that charge was introduced in the predicted molecule.

Once all the changes within the chemical transformation are applied, the generated products are saved as json file, where the information about the final product (Smiles ID and mol block text), the reaction template (ID, EC number and Formula), and the initial molecule (name, ID and structure) is stored. Although the generation of the mol block text was always possible because it was generated manually, the SMILES string was generated using the RDKit package.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data employed in this study can be obtained from the following databases: (i) Metabolic model: AGREDA v1.1.0 (https://github.com/francesco-balzerani/AGREDA_1.1); (ii) Metabolites and Chemical rules: PubChem (https://pubchem.ncbi.nlm.nih.gov/), Human Metabolome Database (https://hmdb.ca/), KEGG (https://www.genome.jp/kegg/), MetaNetX (https://www.metanetx.org/), RetroRules (https://retrorules.org/), Phenol-Explorer (http://phenol-explorer.eu/). The source data underlying Fig. 1, Fig. 2a-e and Supplementary Fig. 2 are provided as a Source Data file.

Code availability

Python code of PROXIMAL2 is available in https://github.com/HassounLab/PROXIMAL2.

Received: 15 May 2023; Accepted: 9 May 2024;
Published online: 27 May 2024

References

1. Blanco, A. & Blanco, G. Metabolism. In *Medical Biochemistry* (eds. Blanco, A. & Blanco, G.) 275–281 (Elsevier, 2017). https://doi.org/10.1016/B978-0-12-803550-4.00013-6.
2. Hafner, J. & Hatzimanikatis, V. NICEpath: Finding metabolic pathways in large networks through atom-conserving substrate–product pairs. *Bioinformatics* **37**, 3560–3568 (2021).
3. Folador, E. L. et al. Protein-Protein Interactions: An Overview. In *Encyclopedia of Bioinformatics and Computational Biology* (eds. Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C.) 821–833 (Elsevier, 2019) https://doi.org/10.1016/B978-0-12-809633-8.20292-6.
4. Thiele, I., Heinken, A. & Fleming, R. M. T. A systems biology approach to studying the role of microbes in human health. *Curr. Opin. Biotechnol.* **24**, 4–12 (2013).
5. Price, N. D., Papin, J. A., Schilling, C. H. & Palsson, B. O. Genome-scale microbial in silico models: The constraints-based approach. *Trends Biotechnol.* **21**, 162–169 (2003).
6. MohammadiPeyhani, H., Hafner, J., Sveshnikova, A., Viterbo, V. & Hatzimanikatis, V. Expanding biochemical knowledge and illuminating metabolic dark matter with ATLASx. *Nat. Commun.* **13**, 1–12 (2022).
7. Amin, S. A., Chavez, E., Porokhin, V., Nair, N. U. & Hassoun, S. Towards creating an extended metabolic model (EMM) for E. coli using enzyme promiscuity prediction and metabolomics data. *Microb. Cell Fact.* **18**, 1–12 (2019).
8. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

9. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: A European ELIXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2019).
10. Magnúsdóttir, S. et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* **35**, 81–89 (2017).
11. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **45**, D37–D42 (2017).
12. Kersey, P. J. et al. Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **46**, D802–D808 (2018).
13. Gupta, R. D. Recent advances in enzyme promiscuity. *Sustain. Chem. Process.* **4**, 1–7 (2016).
14. Copley, S. D. Shining a light on enzyme promiscuity. *Curr. Opin. Struct. Biol.* **47**, 167–175 (2017).
15. Notebaart, R. A., Kintses, B., Feist, A. M. & Papp, B. Underground metabolism: network-level perspective and biotechnological potential. *Curr. Opin. Biotechnol.* **49**, 108–114 (2018).
16. Guzmán, G. I. et al. Enzyme promiscuity shapes adaptation to novel growth substrates. *Mol. Syst. Biol.* **15**, 1–14 (2019).
17. Carbonell, P., Parutto, P., Herisson, J., Pandit, S. B. & Faulon, J. L. XTMS: Pathway design in an eXTended metabolic space. *Nucleic Acids Res.* **42**, 389–394 (2014).
18. Carbonell, P. & Faulon, J. L. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* **26**, 2012–2019 (2010).
19. Kumar, A., Wang, L., Ng, C. Y. & Maranas, C. D. Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.* **9**, 184 (2018).
20. Koch, M., Duigou, T. & Faulon, J. L. Reinforcement learning for bioretrosynthesis. *ACS Synth. Biol.* **9**, 157–168 (2020).
21. Ni, Z., Stine, A. E., Tyo, K. E. J. & Broadbelt, L. J. Curating a comprehensive set of enzymatic reaction rules for efficient novel biosynthetic pathway design. *Metab. Eng.* **65**, 79–87 (2021).
22. Li, C. et al. Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci.* **59**, 5051–5060 (2004).
23. 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.* **5**, 1155–1166 (2016).
24. Hafner, J., Mohammadipeyhani, H., Sveshnikova, A., Scheidegger, A. & Hatzimanikatis, V. Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power. *ACS Synth. Biol.* **9**, 1479–1482 (2020).
25. Jeffryes, J. G. et al. MINEs: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminform.* **7**, 1–8 (2015).
26. Duigou, T., Du Lac, M., Carbonell, P. & Faulon, J. L. Retrorules: A database of reaction rules for engineering biology. *Nucleic Acids Res.* **47**, D1229–D1235 (2019).
27. Yousofshahi, M., Manteiga, S., Wu, C., Lee, K. & Hassoun, S. PROXIMAL: A method for prediction of xenobiotic metabolism. *BMC Syst. Biol.* **9**, 1–17 (2015).
28. Carbonell, P., Parutto, P., Baudier, C., Junot, C. & Faulon, J. L. Retropath: Automated pipeline for embedded metabolic circuits. *ACS Synth. Biol.* **3**, 565–577 (2014).
29. Delépine, B., Duigou, T., Carbonell, P. & Faulon, J. L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab. Eng.* **45**, 158–170 (2018).
30. Rothwell, J. A. et al. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database* **2013**, bat070 (2013).
31. Hassanpour, N. et al. Biological filtering and substrate promiscuity prediction for annotating untargeted metabolomics. *Metabolites* **10**, 160 (2020).
32. Kotera, M. et al. RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Inform.* **15**, P062 (2004).
33. Blasco, T. et al. An extended reconstruction of human gut microbiota metabolism of dietary compounds. *Nat. Commun.* **12**, 1–12 (2021).
34. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Components* **8**, 5281 (2011).
35. Henry, C. S. et al. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977–982 (2010).
36. Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* **46**, 7542–7553 (2018).
37. Balzerani, F. et al. Prediction of degradation pathways of phenolic compounds in the human gut microbiota through enzyme promiscuity methods. *npj Syst. Biol. Appl.* **8**, 24 (2022).
38. Kahle, K. et al. Polyphenols are intensively metabolized in the human gastrointestinal tract after apple juice consumption. *J. Agric. Food Chem.* **55**, 10605–10614 (2007).
39. Wu, M. et al. Phloretin ameliorates dextran sulfate sodium-induced ulcerative colitis in mice by regulating the gut microbiota. *Pharmacol. Res.* **150**, 104489 (2019).
40. Stalmach, A., Edwards, C. A., Wightman, J. D. & Crozier, A. Colonic catabolism of dietary phenolic and polyphenolic compounds from Concord grape juice. *Food Funct.* **4**, 52–62 (2013).
41. Monagas, M. et al. Dihydroxylated phenolic acids derived from microbial metabolism reduce lipopolysaccharide-stimulated cytokine secretion by human peripheral blood mononuclear cells. *Br. J. Nutr.* **102**, 201–206 (2009).
42. Nurmi, T. et al. Lignans in selected wines. *Food Chem.* **83**, 303–309 (2003).
43. Landete, J. M. Plant and mammalian lignans: A review of source, intake, metabolism, intestinal bacteria and health. *Food Res. Int.* **46**, 410–424 (2012).
44. Baderschneider, B. & Winterhalter, P. Isolation and characterization of novel benzoates, cinnamates, flavonoids, and lignans from Riesling wine and screening for antioxidant activity. *J. Agric. Food Chem.* **49**, 2788–2798 (2001).
45. Senizza, A. et al. Lignans and gut microbiota: An interplay revealing potential health implications. *Molecules* **25**, 1–17 (2020).
46. Vardamides, J. C. et al. Scaphopetalone and scaphopetalumate, a lignan and a triterpene ester from Scaphopetalum thonneri. *Phytochemistry* **62**, 647–650 (2003).
47. Andrus, M. B., Johnson, F. B., Greer, M. R. J. & Cates, R. G. Scaphopetalone analogs and their uses. *US Pat.* **1**, 2015–2018 (2017).
48. Scalbert, A., Johnson, I. T. & Saltmarsh, M. Polyphenols: antioxidants and beyond. *Am. J. Clin. Nutr.* **81**, 215–217 (2005).
49. Scalbert, A., Manach, C., Morand, C., Rémésy, C. & Jiménez, L. Dietary polyphenols and the prevention of diseases. *Crit. Rev. Food Sci. Nutr.* **45**, 287–306 (2005).
50. Heim, K. E., Tagliaferro, A. R. & Bobilya, D. J. Flavonoid antioxidants: Chemistry, metabolism and structure-activity relationships. *J. Nutr. Biochem.* **13**, 572–584 (2002).
51. Halliwell, B. Effect of diet on cancer development: Is oxidative DNA damage a biomarker? *Free Radic. Biol. Med.* **32**, 968–974 (2002).
52. Dembinska-Kiec, A., Mykkänen, O., Kiec-Wilk, B. & Mykkänen, H. Antioxidant phytochemicals against type 2 diabetes. *Br. J. Nutr.* **99**, ES109–ES117 (2008).
53. Kim, S. et al. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
54. Wishart, D. S. et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
55. Moretti, S. et al. MetaNetX/MNXref - Reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44**, D523–D526 (2016).

56. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
57. Kotera, M. et al. KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. *BMC Syst. Biol.* **7**, 1–17 (2013).
58. Sato, M., Suetake, H. & Kotera, M. KCF-Convoy: efficient Python package to convert KEGG Chemical Function and Substructure fingerprints. bioRxiv, 2018, p. 452383.
59. Zhou, Y. & Zhuang, X. Kinetic analysis of sequential multistep reactions. *J. Phys. Chem. B* **111**, 13600–13610 (2007).

Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation programme through the STANCE4HEALTH project (Grant No. 816303); and by Fundacion La Caixa with the grant promoting university interchange to young predoctoral researchers [to F.B.].

Author contributions

F.J.P. and S.H. conceived this study. F.B., T.B., S.P-B., L.V., F.J.P. and S.H. developed the algorithm and performed the computational analysis. All authors wrote, read, and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00381-1.

Correspondence and requests for materials should be addressed to Soha Hassoun or Francisco J. Planes.

Reprints and permissions information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024