Computational Analysis of the Gut Microbiota-Mediated Drug Metabolism

Sammie Chum*, Alberto Naveira Montalvo*, Soha Hassoun^
Department of Computer Science
Tufts University, Medford, MA 02155

Abstract

The gut microbiota, an extensive ecosystem harboring trillions of bacteria, plays a pivotal role in human health and disease, influencing diverse conditions from obesity to cancer. Among the microbiota's myriad functions, the capacity to metabolize drugs remains relatively unexplored despite its potential implications for drug efficacy and toxicity. Experimental methods are resource-intensive, prompting the need for innovative computational approaches. We present a computational analysis aimed at predicting gut microbiota-mediated drug metabolism (MDM). This computational analysis incorporates data from diverse sources, e.g., UHGG, MagMD, MASI, KEGG, and RetroRules. An existing tool, PROXIMAL2, is used iteratively over all drug candidates from experimental databases queried against biotransformation rules from RetroRules to predict potential drug metabolites along with the enzyme commission number responsible for that biotransformation. These potential metabolites are then categorized into gut microbiota-mediated drug metabolites by cross referencing UHGG. The analysis' efficacy is validated by its coverage on each of the experimental databases in the gut microbial context, being able to recall up to 74% of experimental data and producing a list of potential metabolites, which an average of about 65% are relevant to the gut microbial context. Moreover, explorations into ranking metabolites, iterative applications to account for multi-step metabolic pathways, and potential applications in experimental studies showcase its versatility and potential impact beyond raw predictions. Overall, this study presents a promising computational analysis for further research and application in the fields of gut MDM, drug development and human health.

*Equal contributions

^Corresponding Author: soha.hassoun@tufts.edu, Phone: 617 627 5177

Keywords:

Drug-microbe interactions, Microbial drug metabolism; Gut microbiota; Biotransformation Rules, Enzyme Prediction

1. Introduction

The human gastrointestinal (GI) tract is colonized by ~10^{14} bacteria belonging to at least several hundred species that are collectively termed the intestinal or gut microbiota [1]. This bacterial community is important for digesting dietary nutrients and extracting energy by fermenting carbohydrates indigestible by human enzymes [2]. The intestinal microbiota also impacts a wide range of functions in the GI tract [3] [4]. Beyond the GI tract, gut-brain [5], gut-lung, [6] and gut-liver [7] links have also been identified, thus highlighting the importance of the microbiota to human health and disease. Further, as microbiota-derived metabolites have been detected in circulation [8], alterations in the intestinal microbiota composition have been correlated to several diseases including obesity [9, 10], diabetes [11], cancer [12], and asthma [6].

A topic that is less explored, however, is characterizing the capacity of gut microbes to metabolize drugs, leading to drug activation, inactivation, and toxicity [13-15]. There are already documented studies on how gut microbes interfere with specific drugs. In some cases, such interference is by design, where a prodrug (non-active form) is transformed via microbial enzymes. An example is azo-bonded prodrugs of 5-aminosalicylic acid (5-ASA), including sulfasalazine, balsalazide, and olsalazine, that are used in the treatment of ulcerative colitis [16]. The azo bond links 5-ASA to a carrier molecule. When reaching the colon, azoreductases encoded by the gut microbiome cleave the azo bond to release 5- ASA. 5-ASA degradation rates are a function of gut composition and specificity on carrier molecules [17]. Other cases of interference are non-intentional and stomp researchers and clinicians for decades. One such example is cardiac drug digoxin [18], which was first isolated in the 1930s from the foxglove plant and became a drug in the 1960s. While the microbial species (*Eggerthella lenta*, formerly *Eubacterium lentum*) responsible for the degradation was identified in 1982 [19], the cardiac glycoside reductase genes encoding for this enzymatic degradation was only identified in 2013 and only in a few of *E. lenta's* bacterial strains. These two examples and many others, e.g., ref [20-23], reflect the gut microbes' enzymatic capacities for drug reduction, hydrolysis, acetylation/deacylation decarboxylation, dehydroxylation, demethylation, dehalogenation, and other reactions.

Concerns over microbiota-mediated drug metabolism (MDM) have prompted experimental testing the activity of drugs on panels of bacteria in isolation [24] and measuring isolate activities against a large panel of drugs [25]. Recently, the drug-metabolizing capacity of human fecal communities was challenged with 438 different compounds [26, 27]. A compounding challenge is that drug metabolization may occur through multiple steps that span microbes and the host [28], thus necessitating more complex experimentation. Complex experimental setups can now mimic the entire GI tract [29]. A further challenge is that microbial responses to drugs are a function of microbial community composition [30]. There are now databases that catalog drug-microbial interactions. The MDAD database provides a collection of microbe-drug associations [31], while the MASI (microbiota––active substance interactions) database offers a broad view of interactions between human microbiota and active substances such as drugs, dietary components, herbal products, probiotics, and environmental chemicals [32]. The MagMD (Metabolic action of gut Microbiota to Drugs) database was developed to specifically catalog interactions between gut microbiota and drugs, including detailed annotations on microbial enzymes, metabolites, and the effects on drugs [33]. MagMD data is collected from existing databases (e.g., PharmacoMicrobiomic [14], MASI [32]) and literature searches in PubMed using keywords related to gut microbiota and drug metabolism. To increase the coverage and annotation of MagMD, drug-microbe interactions were established by identifying microbes that had gene sequences with high similarity to sequences of known drug-metabolizing enzymes. For example, beta-glucosidase was responsible for 16,088 interactions between drugs and microbial species.

A complementary approach to experimental work and database curation is using computational approaches to analyze and predict MDM. Microbial promiscuity on drugs have been studied prior based on chemical similarity of drugs [34, 35], and by considering differences between substrate-product pairs of MetaCyc reactions [36]. A recent focus has been predicting microbe-drug associations. LCASPMDA combines a learnable graph convolutional attention network and iterative sampling ensemble strategy to infer latent microbe-drug associations [37], while GARFMDA combines graph attention networks and bilayer random forest to accomplish a similar task [38]. MDASAE uses stacked autoencoders to predict microbe-disease-drug associations [39]. These models, however, do not identify the enzymes responsible for the microbe activity on the drug, nor the metabolic product. There remains a need therefore to identify microbial metabolic products and the corresponding metabolizing enzymes, which in turn will guide experimentations and expedite MDM discoveries.

To address this need, this paper explores the use of computational analysis to predict putative products due to the activity of a specific microbial enzyme on a query drug. To identify microbial enzymes, we crosslink KEGG Modules cataloged in the Unified Human Gastrointestinal Genome (UHGG) collection [40] to enzyme commission (EC) numbers within the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [41, 42]. To predict MDM products, we utilize a rule-based method, PROXIMAL2 [43, 44] to predict such products. When specifying the enzyme's Enzyme Commission (EC) number, PROXIMAL2 first identifies biotransformation rules relevant to the enzyme, and then applies them to the query drug to generate enzymatic products. The biotransformation rules are derived from reaction rules cataloged in the RetroRules database [45], which over 40,000 reaction rules extracted from public databases and encompass many drug-metabolizing reactions. For evaluation, we curate several datasets from MagMD, MASI, and DrugBank [46] , with and without known catalyzing enzymes, and assess the validity of using PROXIMAL2 to identify cataloged drug-derivative products and the corresponding microbial enzyme responsible for acting on the query drug. Further, we explore using all biotransformation rules in RetroRules to allow for the discovery of putative, yet currently undocumented, drug microbial metabolism. The contributions of the paper are:

- Curating a list of known microbial enzymes by crosslinking KEGG Modules and the UHGG (see Supplementary Data 1).
- Curating two datasets to validate MDM predictions with gut-specific cases and two datasets with non-gut specific cases (see Supplementary Data 2). The datasets include 30 gut-specific, drug-product cases from MASI, MagMD, and the literature, and 465 non-gut specific cases of drugs from DrugBank metabolized by biological enzymes other than explicitly stated CYP enzymes.
- Demonstrating accuracy of MDM predictions using PROXIMAL2 with product recall rates of 73.91% on known gut-specific datasets, and 42.00% on non-gut specific datasets.
- Assigning microbial enzymes to 112 cases of documented drug-product relationships with unknown enzymes.
- Discovery that 29.86% of the non-gut specific dataset can be putatively metabolized by gut enzymes.

2. Methods

Interaction known to be gut-specific	Drug	Enzyme known	Product

E-Microbial	☒	☒	☒	☒
E-Biological	☐	☒	☒	☒
NE-Microbial	☒	☒	☐	☒
NE-Biological	☐	☒	☐	☒

Table 1. Four datasets were curated for validation and exploration of our methods. The datasets are characterized by the type of available information.

2.1 Data

Microbial enzymatic activity on drugs is cataloged in the MagMD and MASI databases. To explore if gut microbes can act on drugs utilizing enzymes whose function is similar to that of other biological enzymes, we curated additional datasets of drug-products from DrugBank. In total, we curated four datasets (Table 1). Datasets with known enzymes include an E prefix, whereas datasets with no known enzymes include an NE prefix.

E-Microbial is a validation set for our method, consisting of drug-microbe-enzyme-product relationships that are known for microbial enzymes from MagMD and MASI and data from the literature that have not yet been added to these databases; E-Biological presents drug-enzyme-product relationships from DrugBank that are known for any biological enzyme. As data involving drug, microbe, metabolizing enzyme, and metabolic derivative products are limited, we expanded our curation efforts to include microbe-drug interactions with no known metabolizing enzyme. NE-Microbial and NE-Biological datasets consist of drug-product relationships from microbial and biological metabolism respectively, without a documented metabolizing enzyme.

The datasets were curated as follows. E-Microbial (Table 2A) was compiled from MagMD, MASI, and manual literature search. Each entry comprises a drug, microbe, enzyme name and corresponding EC number, and a metabolic derivative product. From MagMD and MASI, 16 unique cases of drug metabolism were extracted, with 5 overlapping cases between the two databases. Subsequently, manual literature search was utilized to identify cases that are not currently documented in the MagMD or the MASI databases. Seven additional cases of MDM were identified. E-Microbial therefore contains data for 20 unique drugs, with 23 unique drug-product pairs. The E-Microbial set is therefore used to validate the model's ability in predicting metabolizing enzymes and their products.

E-Biological was curated using DrugBank's database of 3,703 drug metabolism reactions. Most participating enzymes of the DrugBank reactions are non-gut microbiome specific, with many of the enzymes to be known for liver metabolism. To remove some such reactions for a more balanced dataset, entries were filtered to exclude cases with keywords such as "liver", "cytochrome", and "CYP". E-Biological was composed of 100 known drug-enzyme-product relationships. The E-Biological set is used to validate the model's predictive ability of the appropriate metabolizing enzymes and products, and we also assert if the predicted enzymes are microbial.

The next two datasets, NE-Microbial and NE-Biological provide only drug-product pairs with an unknown metabolizing enzyme. NE-Microbial (Table 2B) was curated with cases specific to metabolism by the gut microbes, but unknown enzymes, from MASI and MagMD. Five cases were extracted from MASI and 2 from MagMD. Altogether, there are 7 drugs and 7 unique drug-metabolite pairs. Similarly, NE-Biological

is composed of 365 DrugBank drug-product pairs with undocumented enzymes. In addition to validating if our method suggests the documented metabolic products, we report on gut microbiota enzymes that can metabolize these drugs.

(A) E-Microbial

Drug	Species	Enzyme	KEGG EC	Product	Source
Balsalazide	Clostridium; Eubacterium	Azo reductases	1.6.5.-	5-aminosalicylic acid	MagMD
Clonazepam	*Gut microbes*	Nitroreductase	1.7.1.16	7-aminoclonazepam	MagMD
Digoxin	Eubacterium Lentum	Digoxin reductase	1.3.2.-	Dihydrodigoxin	MagMD
Genistin	Absiella	Beta-glucosidase	3.2.1.21	Geinstein	MagMD
Irinotecan	Bacteroides oleiciplenus; Bacteroides ovatus; Bacteroides salyersiae; Bacteroides stercorirosoris; etc.	Beta-glucosidase	3.2.1.31	SN-38	MagMD
Isoquercetin	Clostridia; Acutalibacteraceaea; Christensenellales; etc.	Beta-glucosidase	3.2.1.21	3,4-dihydroxyphenylacetic acid	MagMD
	Clostridia; Acutalibacteraceaea; Christensenellales; etc.	Beta-glucosidase	3.2.1.21	Butyrate	MagMD

	Clostridia; Acutalibacteraceae; Christensenellales; etc.	Beta-glucosidase	3.2.1.21	Acetate	MagMD
Naringin	Bacteroides cutis; Bacteroides fluxus; Bacteroides fragilis, etc.	Beta-glucosidase	3.2.1.21	Naringenin	MagMD
Neoprontosil	Gut microbes	Azo reductases	1.6.5.-	Sulfanilamide	MagMD
Nitrazepam	Gut microbes	Nitroreductase	1.7.1.16	7-aminonitrazepam	MagMD
Olsalazine	Clostridium; Eubacterium	Azo reductases	1.6.5.-	5-aminosalicylic acid	MagMD
Prontosil	Gut microbes	Azo reductases	1.6.5.-	Sulfanilamide	MagMD
Sulfasalazine	Alistipes	Azo reductases	1.6.5.-	Sulfapyridine	MagMD
	Alistipes	Azo reductases	1.6.5.-	5-aminosalicylic acid	MagMD
Levodopa	Enterococcus faecium	Tyrosine decarboxylases	4.1.1.25	Dopamine	MASI
5-fluorouracil	Proteobacteria; Firmicutes	Uracil phosphoribosyltransferase	2.4.2.9	5-fluorouridine monophosphate	Clarke et al., 2019
Albiflorin	Bifidobacteria	Esterases/Hydrolases	3.1.1.1	Benzoic acid, B2, B4, B3, BL esterases	Peng et al., 2022
Caffeic acid	Pseudomonas; Lactobacillus; Clostridium	Bacterial acyl-CoA dehydrogenase	1.3.8.-	3,4-dihydroxyphenyl propionic acid	Guo, Lee, Jeong, 2020
Glycyrrhizin	Streptococcus LJ-22	Beta-glucuronidase	3.2.1.31	18β-Glycyrrhetinic Acid-3-O-β-D-Glucuronide	Park et al., 2005
Sulindac	Gut microbes	Methionine sulfoxide reductase	1.8.4.13	Sulindac sulfide	Hirosawa, 2022

Chloramphenicol	Gut microbes	Nitroreductase	1.7.1.16	Amino-chloramphenicol	Guo, Lee, Jeong, 2020
Flunitrazepam	Gut microbes	Nitroreductase	1.7.1.16	7-aminonitrazepam	Guo, Lee, Jeong, 2020

(B) NE-Microbial

5-fluorocytosine	Intestinal microflora			5-fluorouracil	MASI
Nizatidine	Unclassified gut microbiota			Nitroso-deoxygenated metabolite	MASI
Resveratrol	Bifidobacterium longum; Lactobacillus acidophilus; Bacillus cereus			Piceid	MASI
Sorivudine	Enterobacteriaceae; Bacteroidetes			(e)-5-(2-bromovinyl) uracil	MASI
Tacrolimus	Faecalibacterium prausnitzii			M1	MASI
Sennosides	Gut microbes			Rheinanthrone	MagMD
Ipsalazide	Gut microbes			Local 5-ASA	MagMD

Table 2. Gut-microbiome drug metabolism cases curated from MagMD, MASI, and manual search through the literature. **(A)** Cases with documented information on the drug, enzyme, and resulting metabolite. **(B)** Cases with documented information on the drug and resulting metabolite, but unknown acting enzyme or EC class.

2.2. PROXIMAL2 for predicting MDM

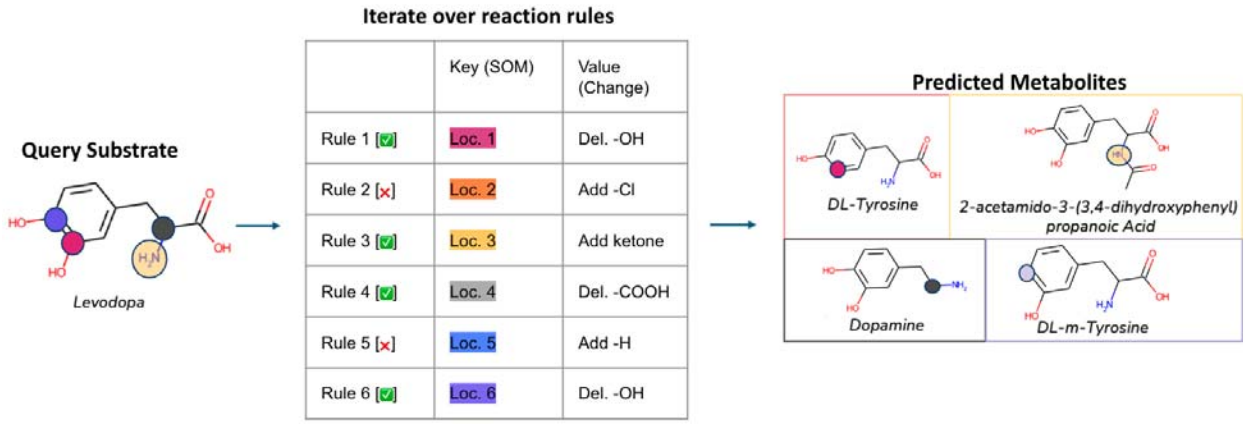

Figure 1. Schematic representation of PROXIMAL2 workflow. PROXIMAL2 iterates over atoms in the query substrate and applies any rule or change that has a corresponding key location that matches in the list of reaction rules. Each reaction rule is described by a key-value pair that encodes the site of metabolism (SOM) and the change that occurs, respectively.

PROXIMAL2, a tool for the prediction of xenobiotic metabolism for CYP enzymes [43] and later to explore a wider range of enzymatic biotransformations [47], was repurposed for the prediction of gut MDM (**Figure 1**). Given a set of biotransformation rules from a database such as RetroRules [45], PROXIMAL2 creates lookup tables to catalog the biotransformation. Each entry in the table consists of a key and a value. The key consists of a putative reaction center (site-of-metabolism, SOM) and its adjacent neighboring atoms and distant adjacent atoms, i.e., a 2-hop neighborhood, all coded in KEGG atom types [41, 42]. The value describes the change to the reaction center and the neighbors, if applicable. Given a query drug and a specific enzyme, PROXIMAL2 will first identify the reaction rules associated with the enzyme. Then, PROXIMAL2 will iterate through the reaction rules and attempt to apply them to each atom within the query. PROXIMAL2 predicts several putative products, as shown in the example in **Figure 1**.

2.3. Associating enzymes with microbial species

RetroRules biotransformations are associated with enzyme commission numbers (ECs) but are not linked to any species. We identify the subset of RetroRules biotransformations that can be attributed to microbial enzymes. We obtain a listing of known microbial enzymes as follows. The UHGG database [40] lists associations between microbes and their corresponding KEGG Modules, where each such module is manually curated to reflect a set of genes that operates as a functional unit. From the KEGG database, we extracted a set of enzymes (EC numbers) associated with these modules. We observed that of all the 42,307 biotransformations listed in RetroRules, 42.21% (17,858) are associated with enzymes relevant to the gut. As such, if an enzyme is contained within this set, then we can classify the corresponding biotransformation as due to the gut MDM. This list of microbial biotransformations, their relevant EC numbers, and corresponding species is available in **Supplementary Data 1 and 2.**

(A) Biotransformation predictions due to known catalyzing enzymes	**(B) Biotransformation predictions due to RetroRules**

		Product Recall (%)	Enzyme Recall (%)	Both Recall (%)	Percent in gut (%)	Product Recall (%)	Enzyme Recall (%)	Both Recall (%)	Percent in gut (%)
E-Microbial	**any enzyme**					73.91	39.13	21.74	64.77
	enzymes associated with gut microbes	47.83	39.13	21.74	100.00	65.22	39.13	21.74	100.00
E-Biological	**any enzyme**	36.00	35.00	14.00	56.41	42.00	35.00	14.00	60.26
	enzymes associated with gut microbes	31.00	24.00	9.00	100.00	34.00	24.00	9.00	100.00
NE-Microbial	**any enzyme**					57.14			72.34
	enzymes associated with gut microbes					42.86			100.00
NE-Biological	**any enzyme**					38.08			61.32
	enzymes associated with gut microbes					29.86			100.00

Table 3. Results of applying PROXIMAL2 to predict microbial drug metabolism, reported using the recall metric on the product predicted by PROXIMAL2, on the enzyme (if specified for our curated dataset), and the recall of both product and enzyme, if specified. We also report the percentage of cases where the enzyme responsible for the biotransformation was associated with the gut microbes according to our UHGG analysis. **(A)** Using biotransformations that are *only* associated with the enzymes reported in our curated datasets. **(B)** Using all biotransformations available in RetroRules.

3. Results

3.1. Evaluating product and enzyme recall

We assess the capabilities of our approach in predicting microbial drug metabolism by evaluating its ability to recall known products, and enzymes, if known of microbial drug metabolism. Hence, PROXIMAL2 was applied to each drug present in the curated datasets using biotransformations associated with enzymes reported as responsible for such transformation (Table 3A). The resulting products are then compared against the products as reported in each of the corresponding datasets. Performance is measured using product recall, which is defined as the percentage of cases where the drug derivative product was reported by PROXIMAL2. We report the enzyme recall, which is the percentage of cases where the reported enzyme in the curated datasets led to the correct product. We also report on "both recall", the percentage of cases where the correct product is produced due to the biotransformation rules associated with the documented enzyme. Indeed, there are products that PROXIMAL2 predicts due to a documented enzyme that was not identical to the one in our curated dataset. An example is provided in **Figure 2**. Naringin, a flavonoid with anti-inflammatory properties, is correctly predicted by PROXIMAL2 to metabolize into naringenin when catalyzed by EC 3.2.1.21 (**Figure 2A**) through a bond cleavage. However, PROXIMAL2 also generates another metabolism prediction catalyzed by 3.2.1.21 with the introduction of a ring structure (**Figure 2B**).

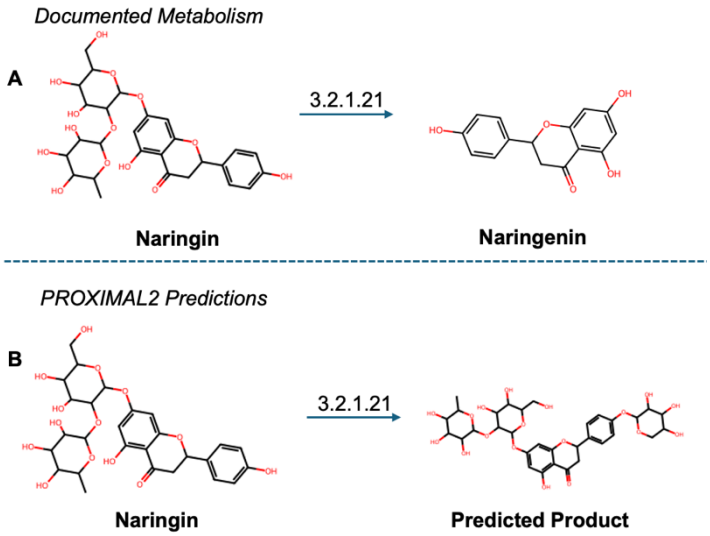

Figure 2. The metabolism of query drug naringin. **(A)** The documented transformation of naringin to naringenin catalyzed by EC 3.2.1.21. **(B)** The predicted metabolism of naringin to an undocumented metabolite generated by PROXIMAL2 and catalyzed by the same EC 3.2.1.21.

To assess the efficacy of PROXIMAL2 in recalling information, the first of the PROXIMAL2 runs we completed utilized the E-Microbial dataset run against the subset of RetroRules that contained enzymes included within this dataset. In this run, 47.83% of products were recalled, 39.13% of enzymes were recalled and 21.74% of cases with the documented enzyme predicting the documented product were observed (Table 3A).

When expanding the set of reference biotransformations on PROXIMAL2 to include the full RetroRules set, we see an increase in product recall to 73.91% for the E-Microbial set (Table 3B). However, only 64.77% of the predicted enzymes were relevant to gut MDM and so, when the products were filtered to

only include these enzymes, the product recall decreased to 65.22%. Despite not being able to recall a large majority of products with just the enzymes listed in the dataset, when expanded to include all of RetroRules, PROXIMAL2 is able to recall 65.22% of documented products in the E-Microbial dataset as a result of gut MDM. The increase in product recall when accounting for all biotransformation in RetroRules illustrates the ability of PROXIMAL2 to annotate products correctly beyond the biotransformations associated with the documented enzymes.

To assess the efficacy of PROXIMAL2 on non-gut MDM specific data, we evaluate the performance of PROXIMAL2 against the E-Biological dataset. On the run with biotransformation predictions due to enzymes documented in the E-Biological dataset, we observe a product recall of 36.00%, enzyme recall of 35.00% and both recall of 14.00% (Table 3A). However, pertaining to these predictions, only 56.41% of predicted enzymes were classified as relevant to gut MDM. As such, when we filter the results to only include gut MDM relevant enzymes the product, enzyme and both recall decreased to 31.00%, 24.00% and 9.00%, respectively.

Expanding the set of reference biotransformations to the full extent of RetroRules, we see an increase in product recall for the E-Biological set to 42.00% (Table 3B). However, only 60.26% of the predicted enzymes were relevant to gut MDM and when the products were filtered to only include microbial enzymes, the product recall decreases to 34.00%. Similar to the E-Microbial dataset, we observe an increase in the product recall of the E-Biological dataset when accounting for all biotransformation in RetroRules. However, the recall metrics on this dataset were consistently lower than the corresponding metrics for the E-Microbial dataset. This result suggests that the biotransformations in the E-Biological dataset are not as clearly represented in the biotransformation rules within RetroRules.

As an additional evaluation for PROXIMAL2's predictions, we assess the efficacy of PROXIMAL2 on the NE-Microbial and NE-Biological datasets which lack enzymatic information pertaining to each of the documented biotransformations. Specifically, we observe that PROXIMAL2 can recall 57.14% and 38.08% of products in the NE-Microbial and NE-Biological dataset respectively when run with the set of reference biotransformations in the full extent of RetroRules (Table 3B). Of these predicted biotransformations, 72.34% of enzymes for the NE-Microbial dataset and 61.32% of enzymes for the NE-Biological dataset were classified as relevant to gut MDM. When only considering these gut MDM specific enzymes, the observed product recall decreased to 42.86% and 29.86%. Consequently, we see that through these two examples, PROXIMAL2 is able to provide gut MDM specific enzyme recommendations for up to about 45% of biotransformations with unknown enzymes.

The ability to generate MDM predictions can also be leveraged to identify novel compounds measured in ex vivo studies, as was demonstrated prior for the CHO cell [48]. The application of PROXIMAL2 allows for the generation of a library of potential metabolites for experimentalists to compare chemically derived structures. Additional applications of comparing the similarity across derived structures with predicted structures provides confirmation, insight, and a potential route of metabolism with the generated EC prediction. There were two such structures in our datasets that we attempted to identify using PROXIMAL2. The derivative structures of the drug tacrolimus were experimentally determined utilizing MS and NMR, with the primary metabolite named M1, as the major metabolite, by researchers [33]. When applying PROXIMAL2, a library of 32 potential metabolites was generated, but none were identical to the experimental M1 structure. We also attempted to annotate the derivative product of the drug nizatidine. This derivative was classified as a nitroso-deoxygenated metabolite using HPLC/MS [14].

The application of PROXIMAL2 generated 3 potential metabolites, with none matching the experimentally derived structure.

3.3. Case study 1: Iterative application of PROXIMAL2 to uncover the microbial metabolism for sennosides

Drugs are often metabolized through multiple enzymatic steps. For example, the drug levosimendan has been documented to first metabolize into an amino phenol pyridazinone metabolite, and then subsequently to an N-acetylated conjugate [49]. Here, we explore how the iterative application of PROXIMAL2 can elucidate multi-step drug metabolism.

The drug sennosides, used for treating constipation, is documented in the MagMD database to metabolize to the product rheinanthrone, as listed in NE-Microbial (Table 2B), but was not successfully predicted by PROXIMAL2. However, as shown in **Figure 3** and as documented prior [50], sennosides is metabolized in a series of biotransformations that first metabolizes to sennidin A, then to rheinanthrone (the active metabolite) and lastly rhein. We examine the iterative application of PROXIMAL2 to uncover this metabolizing pathway by considering each metabolite in this pathway as a query substrate to predict the correct subsequent product.

The first product of this pathway, sennidin A, was correctly predicted by PROXIMAL2 as a product of sennosides through a hydrolysis reaction catalyzed by beta-glucosidase. Subsequently, sennidin A was expected to metabolize into rheinanthrone through a reductive carbon-carbon cleavage more commonly performed by Cytochrome P450 enzymes found in the liver. This product was not generated with PROXIMAL2 due to the lack of a similar transformation rule in the RetroRules database. Although carbon-carbon cleavages exist in RetroRules, none reflect similar molecular neighborhoods and therefore were not applied to the substrate. Lastly, when rheinanthrone was utilized as a query substrate, PROXIMAl2 was successful in generating the product rhein. In addition to product prediction, PROXIMAL2 was able to identify 3 corresponding enzymes likely to catalyze rheinanthrone: 1.13.12.20, 1.13.12.21, and 1.13.12.22. These enzymes share the first three digits of their EC number, 1.13.12.x, representing a class of oxidoreductases that incorporate one oxygen atom. This is illustrated by rheinanthrone's transformation to rhein (**Figure 3**) with the introduction of a double bonded oxygen atom, further supporting the ability to predict an appropriate catalyzing enzyme.

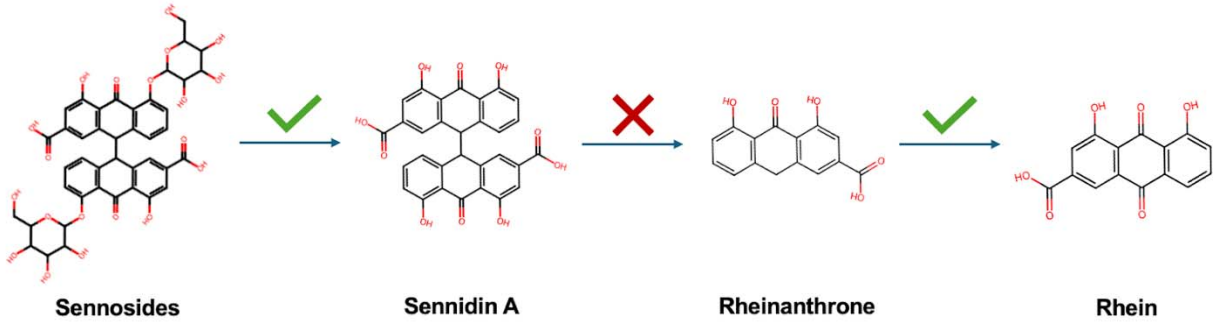

Figure 3. The biotransformation steps that the drug sennosides undergo to transform into product rheinanthrone, as documented in MagMD, and then to rhein. PROXIMAL2 was successful in predicting the metabolism of sennosides to sennidin A and rheinanthrone to rhein, but unable to predict the transformation from sennidin A to rheinanthrone.

3.4. Case study 2: Ranking predicted products of microbial metabolism on tamoxifen and 5-fluorocytosine

The application of biotransformation rules to a query drug results in many putative products. As some enzymes are more promiscuous than others, the number of products varies. For example, as shown in **Figure 4**, enzyme 1.14.14.1 acting on estradiol resulted in 8 products when applying either the subset of rules or all of RetroRules. The variation in predicted products is a result of 6 explored sites of metabolism in estradiol by PROXIMAL2, in addition to different combinations of the reactive sites. The catalyzing enzyme, 1.14.14.1, is a form of monooxygenase within the class of oxidoreductases known for the incorporation or reduction of an oxygen atom. The large range of reactions this enzyme is known to catalyze, including hydroxylation, deamination, dealkylation, oxidation, and epoxidation, further contributes to the many putative products of estradiol metabolism. Excluding this case, there was an average of 2.36 products predicted per correct documented enzyme for E-Biological and 9.0 products for E-Microbial, emphasizing the need to discern the correct product even with knowledge of the catalyzing enzyme. When using all biotransformations within RetroRules, the average number of products is 7.0 for E-Microbial and 2.5 for E-Biological.

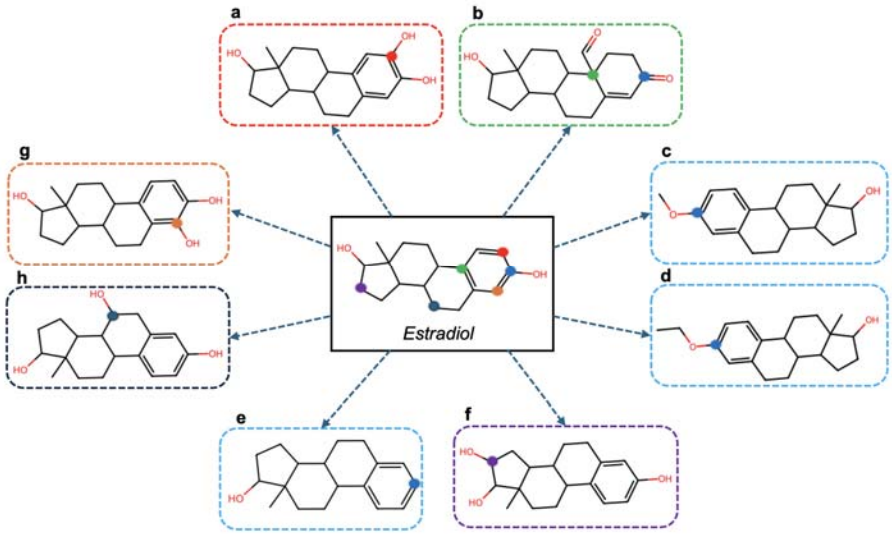

Figure 4. The eight products generated by the application of PROXIMAL2 to the query drug estradiol. The boxes around the molecules and the reaction centers are highlighted with the matching color to the query drug to indicate the site of metabolism, with an exception to product b containing two sites of metabolism.

Identifying the most likely product can be achieved via the application of GNN-SOM [28], a graph-based deep learning technique that can be leveraged to determine the most likely metabolites to be found in vivo. GNN-SOM is a graph neural net trained on all reactions from the KEGG RPAIR database by using RDM encodings, signifying the reaction center, difference, and match in each reactant pair. The final GNN-SOM consists of ten trained networks using the Chebyshev convolutional operator, with 2 or 3 layers, and a layer size of 512. GNN-SOM accepts a molecular fingerprint and EC number as input and returns a likelihood score at each atom of the compound, reflecting its likelihood to serve as a potential site-of-metabolism. By applying GNN-SOM to the predicted products of PROXIMAL2, the score at the reaction center corresponding to each product can be used to denote the likelihood of the product.

To assess the value of GNN-SOM, two ranking approaches can be taken by varying its application. First, GNN-SOM can be utilized on only the putative products generated by PROXIMAL2 that were paired with the experimentally documented EC number, to generate enzyme-specific rankings. Second, all metabolites can be ranked by their likelihood scores, regardless of the EC number pairing generated by PROXIMAL2. This method determines the most likely in vivo candidates without an experimentally documented EC. Therefore, the first method allows for an analysis on the ability of GNN-SOM to discern the most likely site of metabolism with a consistent pair of input information, the drug and the correct metabolism-responsible enzyme. The second simulates how GNN-SOM may be used for the identification of the most likely metabolic products given a query drug without a known catalyzing enzyme.

The application of GNN-SOM in one query drug use case, tamoxifen from E-Biological, was explored following the first method described above **(Figure 5).** Tamoxifen is used to treat breast cancer by blocking estrogen activity to mitigate tumor growth [51]. From PROXIMAL2, 5 products with 5 unique reaction centers were generated with the correct documented EC of 1.14.14.1. The documented metabolite was ranked first with a likelihood score of 0.58, and the second highest ranked metabolite scored 0.08. Reaction centers with higher likelihood scores than the known center, such as the nitrogen atom with a score of 0.78, were also labeled by GNN-SOM but not generated as predictions by PROXIMAL2. Furthermore, GNN-SOM provides a method to select the most likely metabolite for cases with documented promiscuous enzymes.

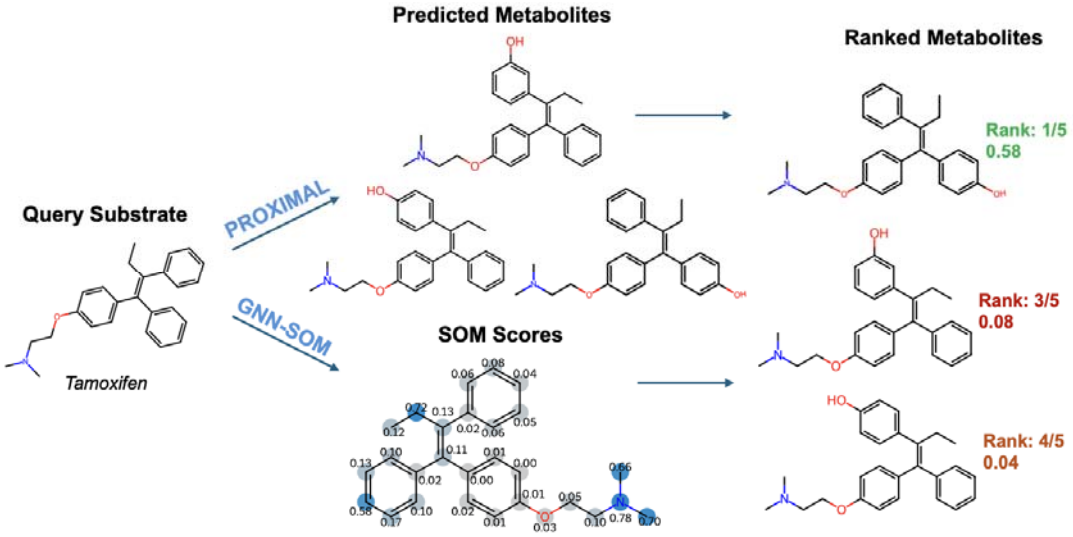

Figure 5. The analysis of tamoxifen from E-Biological with PROXIMAL2 and GNN-SOM to rank all predicted candidate metabolites. Three of the predicted metabolites are showcased, with the correct product ranked first out of five products with the documented EC.

Another application of GNN-SOM in a query drug use case, 5-fluorocytosine from NE-Microbial, used as treatment for fungal infections [52], was explored following the second method described above **(Figure 6).** PROXIMAL2 generated 12 unique products with 6 unique corresponding enzymes and 3 potential reaction centers. Upon analysis of all likelihood scores generated by drug and EC pairs, the top ranked product corresponded with the documented product of 5-fluorouracil with a likelihood score of 0.96 and was the only predicted product with the correct reaction center. The second and third ranked products exhibited significantly lower likelihood scores of 0.104 and 0.009, reflecting unlikely sites of metabolism

and GNN-SOM's potential in determining the most likely products without a known EC. Although the current model of GNN-SOM is not trained on drug-specific candidates, the model demonstrates promising results if repurposed as a tool for processing PROXIMAL2 predictions.

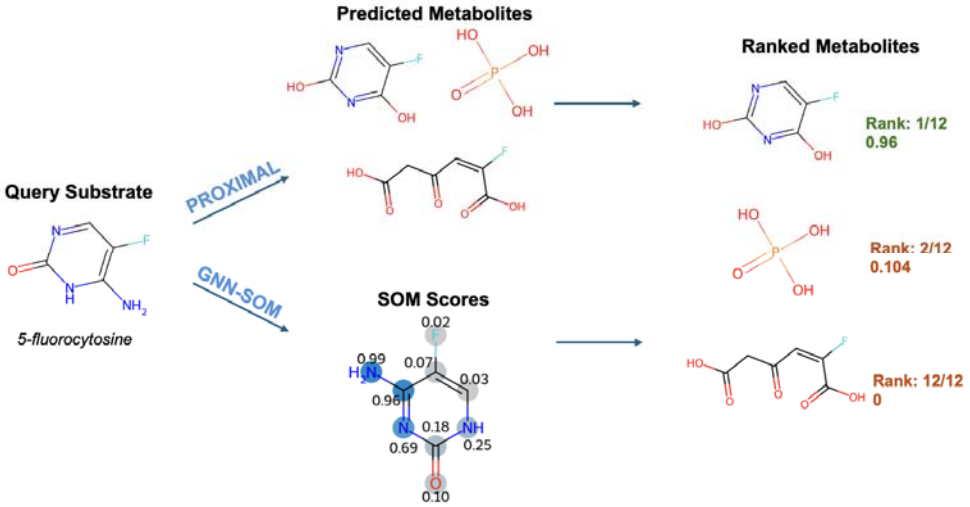

Figure 6. The analysis of 5-fluorocytosine from NE-Microbial with PROXIMAL2 and GNN-SOM to rank all predicted candidate metabolites. Three of the predicted metabolites are showcased, with the correct product, 5-fluorouracil, ranked first with a score of 0.96.

4. Discussion & Conclusion

This study explored gut MDM using computational methodologies, leveraging the PROXIMAL2 tool and integrating data from sources like the UHGG and KEGG. We developed a framework for predicting putative products of microbiota-drug-metabolism, validated against curated datasets, and analyzed for product/enzyme recall and enzyme prediction. Our findings elucidate PROXIMAL2's efficacy in predicting both products and relevant enzymes in the gut microbial context. We accomplished the development of a robust prediction methodology, assessing PROXIMAL2's performance across diverse datasets. Our analyses reveal insights into the complex interactions of gut MDM. Furthermore, by exploring iterative applications of PROXIMAL2 and envisioning potential applications in ex vivo experimental studies, we demonstrate the versatility and potential impact of our MDM pipeline in advancing drug discovery. These findings contribute to understanding gut MDM and pave the way for innovative approaches in deciphering this complex ecosystem. This has significant implications for human health and drug development, offering new avenues for research and application in the field.

While databases such as MASI and MagMD curated known microbial activities on drugs, there remains a need to uncover novel interactions between microbes and drugs. We investigated how rule-based methods like PROXIMAL2 can perform when challenged with a query drug and a microbial enzyme, or when challenged with a known drug-product pair to recover the microbial enzyme responsible for such transformation. The product recall rate of 73.91% on E-Microbial when using all biotransformations within RetroRules indicate either that the enzymatic transformation to produce the product is not available, or that multiple enzymatic steps are needed to generate the product thus reflecting inconsistencies in recording products within the database sources. Further, when comparing the 73.91% recall rate with the low recall rate of 47.83% for E-Microbial when using the named enzyme indicates

that there are potential inaccuracies in recording the named enzyme in the literature or that the biotransformation rules associated with that enzyme are not complete within RetroRules. In either case, PROXIMAL2 may have shortcomings in applying complex transformations to generate the products, calling for improvements in product prediction

5. Acknowledgements

Research reported in this paper was supported by the Office of the Director of the National Institutes of Health under awards number R03OD036490 and R35GM148219. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Tiihonen, K., A.C. Ouwehand, and N. Rautonen, *Human intestinal microbiota and healthy ageing.* Ageing Res Rev, 2010. **9**(2): p. 107-16.
2. Arora, T. and R. Sharma, *Fermentation potential of the gut microbiome: implications for energy homeostasis and weight management.* Nutr Rev, 2011. **69**(2): p. 99-106.
3. Hooper, L.V., D.R. Littman, and A.J. Macpherson, *Interactions between the microbiota and the immune system.* Science, 2012. **336**(6086): p. 1268-73.
4. Kamada, N., et al., *Role of the gut microbiota in immunity and inflammatory disease.* Nat Rev Immunol, 2013. **13**(5): p. 321-35.
5. Cryan, J.F. and S.M. O'Mahony, *The microbiome-gut-brain axis: from bowel to behavior.* Neurogastroenterol Motil, 2011. **23**(3): p. 187-92.
6. Russell, S.L., et al., *Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma.* EMBO Rep, 2012. **13**(5): p. 440-7.
7. Compare, D., et al., *Gut--liver axis: the impact of gut microbiota on non alcoholic fatty liver disease.* Nutr Metab Cardiovasc Dis, 2012. **22**(6): p. 471-6.
8. Wikoff, W.R., et al., *Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites.* Proc Natl Acad Sci U S A, 2009. **106**(10): p. 3698-703.
9. Shen, J., M.S. Obin, and L. Zhao, *The gut microbiota, obesity and insulin resistance.* Mol Aspects Med, 2013. **34**(1): p. 39-58.
10. Greenblum, S., P.J. Turnbaugh, and E. Borenstein, *Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease.* Proc Natl Acad Sci U S A, 2012. **109**(2): p. 594-9.
11. Nicholson, J.K., et al., *Host-gut microbiota metabolic interactions.* Science, 2012. **336**(6086): p. 1262-7.
12. Arthur, J.C., et al., *Intestinal inflammation targets cancer-inducing activity of the microbiota.* Science, 2012. **338**(6103): p. 120-3.
13. Wallace, B.D., et al., *Alleviating cancer drug toxicity by inhibiting a bacterial enzyme.* Science, 2010. **330**(6005): p. 831-5.
14. Saad, R., M.R. Rizkallah, and R.K. Aziz, *Gut Pharmacomicrobiomics: the tip of an iceberg of complex interactions between drugs and gut-associated microbes.* Gut Pathog, 2012. **4**(1): p. 16.
15. Taylor, M.R., et al., *Vancomycin relieves mycophenolate mofetil-induced gastrointestinal toxicity by eliminating gut bacterial beta-glucuronidase activity.* Sci Adv, 2019. **5**(8): p. eaax2358.

16. Nikfar, S., et al., *A meta-analysis of the efficacy of sulfasalazine in comparison with 5-aminosalicylates in the induction of improvement and maintenance of remission in patients with ulcerative colitis.* Dig Dis Sci, 2009. **54**(6): p. 1157-70.
17. Sousa, T., et al., *On the colonic bacterial metabolism of azo-bonded prodrugsof 5-aminosalicylic acid.* J Pharm Sci, 2014. **103**(10): p. 3171-5.
18. Mathan, V.I., et al., *Geographic differences in digoxin inactivation, a metabolic activity of the human anaerobic gut flora.* Gut, 1989. **30**(7): p. 971-7.
19. Dobkin, J.F., et al., *Inactivation of digoxin by Eubacterium lentum, an anaerobe of the human gut flora.* Trans Assoc Am Physicians, 1982. **95**: p. 22-9.
20. Dai, L., et al., *Sevelamer Use in End-Stage Kidney Disease (ESKD) Patients Associates with Poor Vitamin K Status and High Levels of Gut-Derived Uremic Toxins: A Drug-Bug Interaction?* Toxins (Basel), 2020. **12**(6).
21. Guo, Y., H. Lee, and H. Jeong, *Gut microbiota in reductive drug metabolism.* Prog Mol Biol Transl Sci, 2020. **171**: p. 61-93.
22. van Kessel, S.P., et al., *Gut bacterial tyrosine decarboxylase associates with clinical variables in a longitudinal cohort study of Parkinsons disease.* npj Parkinson's Disease, 2021. **7**(1): p. 115.
23. Wilson, I.D. and J.K. Nicholson, *Gut microbiome interactions with drug metabolism, efficacy, and toxicity.* Transl Res, 2017. **179**: p. 204-222.
24. Maier, L., et al., *Extensive impact of non-antibiotic drugs on human gut bacteria.* Nature, 2018. **555**(7698): p. 623-628.
25. Zimmermann, M., et al., *Mapping human microbiome drug metabolism by gut bacteria and their genes.* Nature, 2019. **570**(7762): p. 462-467.
26. van de Steeg, E., et al., *An Ex Vivo Fermentation Screening Platform to Study Drug Metabolism by Human Gut Microbiota.* Drug Metab Dispos, 2018. **46**(11): p. 1596-1607.
27. Javdan, B., et al., *Personalized Mapping of Drug Metabolism by the Human Gut Microbiome.* Cell, 2020. **181**(7): p. 1661-1679.e22.
28. Spanogiannopoulos, P., et al., *The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism.* Nature Reviews Microbiology, 2016. **14**(5): p. 273-287.
29. Mesnage, R., et al., *Alterations in human gut microbiome composition and metabolism after exposure to glyphosate and Roundup and/or a spore-based formulation using the SHIME® technology.* bioRxiv, 2021: p. 2021.12.16.472928.
30. Bishai, J.D. and N.W. Palm, *Small Molecule Metabolites at the Host-Microbiota Interface.* J Immunol, 2021. **207**(7): p. 1725-1733.
31. Sun, Y.-Z., et al., *MDAD: A Special Resource for Microbe-Drug Associations.* Frontiers in Cellular and Infection Microbiology, 2018. **8**.
32. Zeng, X., et al., *MASI: microbiota-active substance interactions database.* Nucleic Acids Res, 2021. **49**(D1): p. D776-D782.
33. Zhou, J., et al., *MagMD: Database summarizing the metabolic action of gut microbiota to drugs.* Comput Struct Biotechnol J, 2022. **20**: p. 6427-6430.
34. Sharma, A.K., et al., *A novel approach for the prediction of species-specific biotransformation of xenobiotic/drug molecules by the human gut microbiota.* Scientific Reports, 2017. **7**(1): p. 9751.
35. Guthrie, L., S. Wolfson, and L. Kelly, *The human gut chemical landscape predicts microbe-mediated biotransformation of foods and drugs.* Elife, 2019. **8**.
36. Mallory, E.K., et al., *Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome.* Pac Symp Biocomput, 2018. **23**: p. 56-67.
37. Yang, Z., et al., *LCASPMDA: a computational model for predicting potential microbe-drug associations based on learnable graph convolutional attention networks and self-paced iterative sampling ensemble.* Front Microbiol, 2024. **15**: p. 1366272.

38. Kuang, H., et al., *A novel microbe-drug association prediction model based on graph attention networks and bilayer random forest.* BMC Bioinformatics, 2024. **25**(1): p. 78.
39. Fan, L., L. Wang, and X. Zhu, *A novel microbe-drug association prediction model based on stacked autoencoder with multi-head attention mechanism.* Sci Rep, 2023. **13**(1): p. 7396.
40. Almeida, A., et al., *A unified catalog of 204,938 reference genomes from the human gut microbiome.* Nat Biotechnol, 2021. **39**(1): p. 105-114.
41. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.
42. Kanehisa, M., et al., *KEGG for taxonomy-based analysis of pathways and genomes.* Nucleic Acids Res, 2023. **51**(D1): p. D587-D592.
43. Yousofshahi, M., et al., *PROXIMAL: a method for Prediction of Xenobiotic Metabolism.* BMC Syst Biol, 2015. **9**: p. 94.
44. Balzerani, F., et al., *Extending PROXIMAL to predict degradation pathways of phenolic compounds in the human gut microbiota.* bioRxiv, 2023: p. 2023.05.17.541107.
45. Duigou, T., et al., *RetroRules: a database of reaction rules for engineering biology.* Nucleic Acids Res, 2019. **47**(D1): p. D1229-D1235.
46. Knox, C., et al., *DrugBank 6.0: the DrugBank Knowledgebase for 2024.* Nucleic Acids Res, 2024. **52**(D1): p. D1265-d1275.
47. Balzerani, F., et al., *Extending PROXIMAL to predict degradation pathways of phenolic compounds in the human gut microbiota.* npj Systems Biology and Applications, 2024. **10**(1): p. 56.
48. Hassanpour, N., et al., *Biological Filtering and Substrate Promiscuity Prediction for Annotating Untargeted Metabolomics.* Metabolites, 2020. **10**(4).
49. Antila, S., S. Sundberg, and L.A. Lehtonen, *Clinical pharmacology of levosimendan.* Clin Pharmacokinet, 2007. **46**(7): p. 535-52.
50. Le, J., et al., *Pharmacology, Toxicology, and Metabolism of Sennoside A, A Medicinal Plant-Derived Natural Compound.* Front Pharmacol, 2021. **12**: p. 714586.
51. Fisher, B., et al., *Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study.* JNCI: Journal of the National Cancer Institute, 1998. **90**(18): p. 1371-1388.
52. Sigera, L.S.M. and D.W. Denning, *Flucytosine and its clinical usage.* Ther Adv Infect Dis, 2023. **10**: p. 20499361231161387.