Computational methods for the identification of drug-metabolizing species and enzymes in the human gut microbiome.

by
Annamarie Bustion

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Peter J. Turnbaugh
8C4E22B44AE0441...

Peter J. Turnbaugh
Chair

DocuSigned by:
[illegible]E6...

Katherine S. Pollard

DocuSigned by:
Russ B. Altman
4D71342F629148B...

Russ B. Altman

Committee Members

Dedicated to the Super8.

Acknowledgements

First, I would like to acknowledge my mentor and advisor, Katie Pollard, for supporting me over the past five years. In Katie, I found an unparalleled intellectual role model and unwavering support as I conducted my thesis work and other scientific projects. Like many students, I experienced highs and lows both in my studies and in my personal life during graduate school, and Katie always responded with patience and understanding. Likewise, I am thankful to the Pollard lab for their scientific advice and camaraderie, especially Maureen Pittman, Calla Martyn, Abigail Lind, Chunyu Zhao, Xiaofan Jin, Patrick Bradley, and Byron Smith. I am so proud to be a member of such a friendly, rockstar group of scientists.

Next, I would like to thank Peter Turnbaugh, my thesis committee chair, for his incredible support and interest in my academic journey. I worked on a dream project for my thesis, and it would not have been possible without his expert knowledge and keen curiosity. Each conversation and writing draft we shared made me a better scientist and thinker, and his passion for science is infectious. I also thank the friends, collaborators, and role models I met in the Turnbaugh lab, especially Benjamin Guthrie (ultimate frisbee player extraordinaire), Lindsey Pieper, Renuka Nayak, Jordan Bisanz, and Elizabeth Bess.

I also thank my thesis committee member, Russ Altman, and his former student, Emily Mallory. Their research on chemical reaction vector embeddings was the impetus for the work presented in this thesis, and Russ has continued to be a source of exciting ideas and suggestions for the project.

In the Pharmaceutical Sciences and Pharmacogenomics graduate program and with Deanna Kroetz specifically, I found my first home and mentorship at UCSF. The training, coursework, and friendship I experienced in PSPG will stay with me my entire scientific career.

I thank my first academic research mentor, Stephanie Gardner, and my first industry research mentor, Kelsey Satterly for taking a chance on me. They both changed the trajectory of my life by taking the time to mentor me well during my first independent scientific projects.

To all my friends near and far, I would not have stayed sane during this process without you. Will Shaffer, I would not be in graduate school if it wasn't for you encouraging me to apply. You kept me accountable and have never lost confidence in me; thank you. Jay Warrick, thank you for sitting next to me and keeping a smile on my face as I sent in applications. To my partner Aaron, thank you for your emotional support and strength as I finish my graduate school journey. You are an endless source of goofiness, reality checks, and late-night snacks. To the Super8 (Ashli, Brittany, Danielle, Sarah, and Zoe), you are my rocks, and I dedicate this work to you.

Lastly, I thank my family. To Olivia and Papa, thank you for being my role models of curiosity and compassion. To my grandmother, Odie, and my godmother, Auntie Anne, both of whom I lost during graduate school: thank you for being mothers to me; your unconditional love and pride remain with me always.

Contributions

Certain chapters of this dissertation contain material in review at a peer-reviewed journal, and the content captured here may differ from the final published form. The bacterial strain collection screen for methotrexate metabolism described in Chapter 3 was designed and conducted by Renuka Nayak. The protein expression and purification experiments described in Chapter 3 were co-designed and co-conducted by me (Annamarie Bustion), Benjamin Guthrie, and Renuka Nayak. All other work in this thesis is my own.

Chapters 2 and 3 contain work in review:

Bustion, A. E., Agrawal, A., Turnbaugh, P. J., Pollard, K. S. A novel in silico method employs chemical and protein similarity algorithms to accurately identify chemical transformations in the human gut microbiome. *Manuscript in review*. bioRxiv. doi:10.1101/2022.08.02.502504

Computational methods for the identification of drug-metabolizing

species and enzymes in the human gut microbiome.

Annamarie Bustion

Abstract

It is well established that the composition of a human's microbiome can contribute to variations in drug metabolism. Indeed, from cumulative research efforts over the past sixty years, hundreds of drugs are now known to be altered in the presence of the gut microbiome, and many of these changed therapeutic profiles are due to direct metabolism by bacterial enzymes. While studies exploring chemical transformations within the human gut microbiome increasingly employ high-throughput methods, determining metabolite identities and the genetic elements responsible for their production is still a low-throughput process. What results are large knowledge gaps that must be overcome before the physiological and clinical relevance of a given bacterial drug metabolism event can be determined. In this thesis, I demonstrate that computational techniques can help alleviate such knowledge gaps. Attempts have been made in the past to computationally predict which bacterial species and enzymes are responsible for chemical transformations in the gut environment, but with limited utility and low accuracy.

This dissertation begins with an overview of current computational approaches for exploring single-step xenobiotic transformations in the microbiome. I highlight the strengths and weakness of current approaches and make architectural recommendations for future tools. Based on these observations and recommendations, I present an *in silico* approach that employs chemical and protein Similarity algorithms that Identify MicrobioMe Enzymatic Reactions (SIMMER). I show that SIMMER predicts the chemistry and responsible species and enzymes for a queried reaction with high accuracy, unlike previous methods. I demonstrate SIMMER use cases in the context of drug metabolism by predicting previously uncharacterized enzymes for 88 drug transformations known to occur in the human gut. Bacterial species containing these enzymes are enriched within human donor stool samples that metabolize the query compound. After demonstrating its utility and accuracy, I chose to make SIMMER available as both a command-line and web tool, with flexible input and output options for determining chemical transformations within the human gut. Lastly, we demonstrate an experimental use-case of SIMMER by performing the first species-level characterization of methotrexate metabolism in the microbiome of Rheumatoid Arthritis patients.

I present SIMMER as a computational addition to the microbiome researcher's toolbox, enabling them to make informed hypotheses before embarking on the lengthy laboratory experiments required to characterize novel bacterial enzymes that can alter human ingested compounds. Beyond pharmaceutical applications, SIMMER can additionally be employed to determine bacterial enzymes responsible for breaking down non-therapeutics, such as dietary compounds or environmental pollutants. The method can also be extended in the future to make predictions on microbes in other body sites or environments.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AP: Atom-pair, as in atom-pair fingerprint

BLAST: Basic Local Alignment Search Tool

BRENDA: BRaunschweig ENzyme DAtabase

DAMPA: Deoxyaminopteroic acid

DAS28: Disease activity score across 28 joints for Rheumatoid Arthritis patients

EC: Enzyme Commission

ECFP: Extended Connectivity Fingerprint

ESM: Evolutionary scale modeling

GSEA: Gene set enrichment analysis

IGC: Integrated Gene Catalog

iHMP: Integrative Human Microbiome Project

IMG: Integrated Microbial Genomes System

IPTG: Isopropyl β-D-1-thiogalactopyranoside

KEGG: Kyoto Encyclopedia of Genes and Genomes

MACCS: Molecular ACCess System keys

MASI: Microbiota-Active Substance Interaction database

MDAD: Microbe-Drug Association Database

MIDAS: Metagenomic Intra-species Diversity Analysis System

MTX: Methotrexate

NORA: New-onset Rheumatoid Arthritis

PANTHER: Protein analysis through evolutionary relationships database

pHMM: Profile Hidden Markov Model

PK: pharmacokinetics

PREDICT: Personalised Responses to Dietary Composition Trial

RA: Rheumatoid Arthritis

SEA: Similarity ensemble approach

SIMMER: Similarity algorithms that Identify MicrobioMe Enzymatic Reactions

SMARTS: SMILES arbitrary target specification

SMILES: Simplified molecular-input line-entry system

TB: Terrific Broth

TT: Topological torsion, as in topological torsion fingerprint

UHGG: Unified Human Gastrointestinal Genome Collection

UHGP: Unified Human Gastrointestinal Protein Collection

Chapter 1

An introduction to computational approaches for exploring single-step xenobiotic transformations in the microbiome.

1.1 Introduction

Pharmacogenomics research has revealed that host genetic variants lead to interindividual variation in human response to drugs. To ensure that patients receive an effective therapeutic at an appropriate dosage, clinicians now consider such genetic variants when prescribing drugs (Thorn et al., 2013). While some of this variation is explained by human gene variants, all therapeutics still exhibit pharmacokinetic and/or pharmacodynamic idiosyncrasies that cannot be explained by host-variants alone (Artacho et al., 2020). From cumulative research within the recently minted, yet nearly century-old, field of pharmacomicrobiomics, however, we now know that the bacterial organisms residing within human hosts also contribute to altered drug disposition profiles (Rizkallah et al., 2010).

Many of the early discoveries in this space were low throughput, deep characterizations of single reactions, such as digoxin's reduction by *Eggerthella lenta* Cgr2 (Haiser et al., 2014; Koppel et al., 2018; Pollet et al., 2017; Spanogiannopoulos et al., 2016; Zimmermann et al., 2019a). In recent years, however, high-throughput studies using large drug libraries applied to either human *ex vivo* stool samples or mono-cultures of gut isolates greatly expanded the number of microbiome affected drugs from under 100 to 273 (Javdan et al., 2020; Zimmermann et al., 2019a). The majority of these xenobiotics are assumed to be directly metabolized by bacterial enzymes, but only 110 of these compounds are associated with an identified bacterial metabolite

(Figure 1.1, Supplementary File 1). Furthermore, only 31 of these reactions have been linked to a characterized bacterial enzyme in human gastrointestinal tracts (Figure 1.1, Table 1.1). Thus, high-throughput experimental pharmacomicrobiomics research has greatly increased our knowledge of the number of drugs altered by bacteria in the human gut, but has also led to various knowledge gaps.

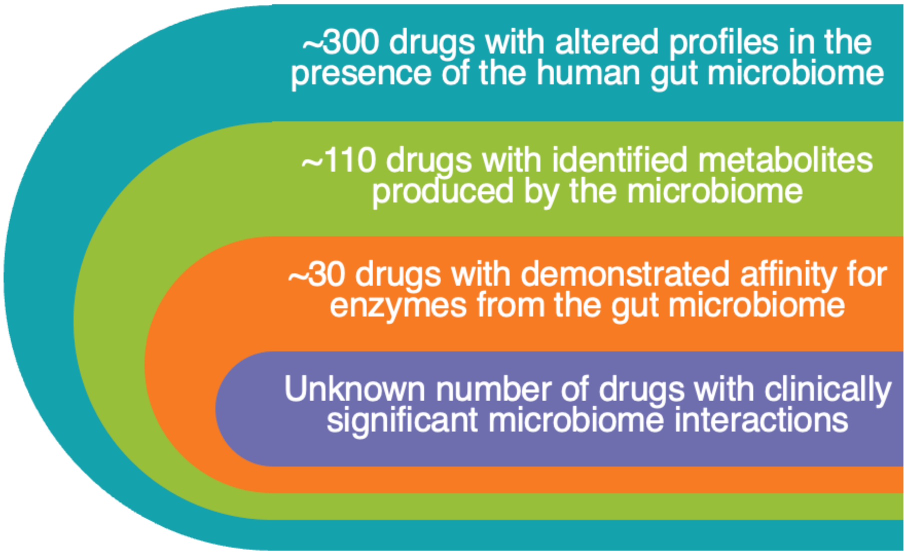

~300 drugs with altered profiles in the presence of the human gut microbiome
~110 drugs with identified metabolites produced by the microbiome
~30 drugs with demonstrated affinity for enzymes from the gut microbiome
Unknown number of drugs with clinically significant microbiome interactions

Figure 1.1 Existing bottlenecks in pharmacomicrobiomics research.
Knowledge of the number of drugs altered by the microbiome far outpaces knowledge of the metabolites formed, and bacterial enzymes responsible.

My previous review sought to guide the microbiome-based pharmacologist in the appropriate experimental directions (Bisanz et al., 2018), so here, I guide the researcher on which computational methods can complement these workflows. In addition, I will comment on the state and comprehensivity of pharmacomicrobiomics, microbiome gene catalog, and microbiome reaction databases.

Table 1.1 Drugs metabolized by characterized bacterial enzymes in the human gut.

Drug	Microbial metabolite	Microbial enzyme	EC
hydrocortisone (cortisol)	20A-dihydrocortisone	20A-HSDH	1.1.1.1
hydrocortisone (cortisol)	20B-dihydrocortisone	20B-HSDH	1.-
balsalazide	5-aminosalicylic acid, 4-aminobenzoyl-beta-alanine	AzoR	1.7.1.6
olsalazine	5-asa	AzoR	1.7.1.6
prontosil	triaminobenzene, sulphanilamide	AzoR	1.7.1.6
nicardipine	aminonicardipine	AzoR	1.5.1.34
sulfasalazine	sulfapyridine, 5-ASA	AzoR, BT_0217, BT_1429	1.7.1.6
ezetimib glucuronide	ezetimib	Beta-glucuronidase	3.2.1.31
diclofenac glucuronide	diclofenac	Beta-glucuronidase	3.2.1.31
indomethacin glucuronide	indomethacin	Beta-glucuronidase	3.2.1.31
ketoprofen glucuronide	ketoprofen	Beta-glucuronidase	3.2.1.31
morphine 6-glucuronide	morphine	Beta-glucuronidase	3.2.1.31
SN38-G	SN-38	Beta-glucuronidase	3.2.1.31
diltiazem	desacetylditiazem	BT_4096	-
pericyazine	acetylpericyazine	BT_2367	-
pericyazine	propionylpericyazine	BT_2367	-
digoxin	dihydrodigoxin	Cgr2	1.3.2.-
gemcitabine	2′, 2′-difluorodeoxyuridine	Cytidine deaminase	3.5.4.5
dopamine	m-tyramine	DadhR506	1.1.-.-
4-ASA	N-acetyl-4-aminosalicylic acid	N-acetyltransferase	2.3.1.5
5-ASA (mesalazine)	N-acetyl-5-aminosalicylic acid	N-acetyltransferase	2.3.1.5
clonazepam	nitroreduction	NfsB	1.5.1.34
flunitrazepam	nitroreduction	NfsB	1.5.1.34
nitrazepam	7-aminonitrazepam	NfsB	1.5.1.34
5-FU	5-fluorodihydropyrimidine-2,4(1H,3H)-dione	PreTA	1.3.1.1

Drug	Microbial metabolite	Microbial enzyme	EC
brivudine	bromovinyluracil	BT_4554	2.4.2.2
sorivudine	bromovinyluracil	BT_4554	2.4.2.2
capecitabine	deglycocapecitabine	Thymidine phosphorylase, Uridine phosphorylase	2.4.2.-
doxifluridine	5-fu	Thymidine phosphorylase, Uridine phosphorylase	2.4.2.-
trifluridine	trifluorothymine	Thymidine phosphorylase, Uridine phosphorylase	2.4.2.-
levodopa	dopamine	TyrDC	4.1.1.25

1.2 Computational approaches for discovering bacterial drug metabolizing enzymes

Knowledge of a bacterial drug metabolism event can stem from many types of experimental techniques (Bisanz et al., 2018), but most examples of drug metabolism were learned from high throughput methods such as drug library screens against strain collections or against patient stool sample incubations (Javdan et al., 2020; Zimmermann et al., 2019a). While these methods are powerful in their ability to quickly find communities and species capable of depleting a parent compound, they do not yet comprehensively address the genetic determinant of metabolism, enzymes. Furthermore, of the few species and enzymes identified experimentally in these studies, it remains unknown how far the demonstrated functions extend beyond the bacterial tree studied.

A computational approach one can employ to narrow in on genetic elements is through metatranscriptomics studies that differentiate changes in gene expression between drug-exposed or drug-naive bacterial communities, with the rationale that many enzymes are under substrate upregulation (Maini Rekdal et al., 2020, 2019; Maurice et al., 2013). Drawbacks of this method are that availability of RNA-seq data in previously published datasets varies, and not all enzymes undergo substrate induction (Bisanz et al., 2018). Related, one could choose to use comparative genomics approaches such as ElenMatchR or PhenoLink, that can determine the genes or SNPs that differ in presence between a metabolizer and non-metabolizer strain/sample (Bayjanov et al., 2012; Bisanz et al., 2020). Thus far, these studies appear most helpful when comparing strains within a species.

Due to the rising popularity of high-throughput LC-MS/MS techniques in pharmacomicrobiomics research, researchers also attempt to identify responsible enzymes by attempting to correlate genetic and metabolomic data. The concept here is that as a particular gene rises in abundance across samples, metabolites in association will also rise. What results is linked clusters of metabolites and genetic elements that may then be further screened for direct associations. This methodology can be cumbersome due to the difficulty of finding associations in such large datasets, and by the fact that abundance measures may be inappropriately calculated or not directly related to rising metabolite levels (Cao et al., 2019; Gloor et al., 2017; Melnik et al., 2017; Yan et al., 2022).

Finally, there also exist methods to directly predict enzymes responsible for a drug-transformation event based on shared chemistry with characterized bacterial reactions from the literature. The two existing tools in this category are DrugBug and Microbe FDT (Guthrie et al., 2019; Sharma et al., 2017); the rest of this introduction will focus on such methods' merits and drawbacks, and what improvements can be made to create more accurate enzyme prediction tools.

1.3 Use of chemical and protein similarity to identify new bacterial enzyme functions

While different from each other in design and scope, both DrugBug and MicrobeFDT leverage chemistry-enzyme linked databases to predict human gut microbiome enzymes responsible for degrading a query compound. As has been discussed previously in the literature (Aziz et al., 2018) and will be further examined in Chapter 2, however, both tools remain preliminary and exhibit low accuracy species and enzyme predictions. Here, I will discuss remedies to the pitfalls of these existing methods, and in Chapter 2 I will present an alternative method designed based on the recommendations that follow.

1.3.1 Chemical and protein information must be drawn from comprehensive databases

Even though the human gut microbiome protein space remains largely (~40%) unannotated (Almeida et al., 2020; Thomas and Segata, 2019), microbiome enzyme predictions are possible by comparing sequence information from the microbiome to protein information found in bacterial reaction and pathway databases. For best results, any predictive tool must be built upon the most comprehensive databases possible for 1. Microbiome sequencing data, and 2. Literature curated metabolic pathway databases.

Human gut microbiome sequencing data is well captured in the Integrated Gene Catalog (IGC), and more recently, the comprehensive Unified Human Gastrointestinal Protein (UHGP) catalog (Almeida et al., 2020; Li et al., 2014). The former comprises 9,879,896 genes found from large metagenomic sequencing studies available at the time of creation, while the latter comprises 170,602,708 genes of 286,997 genomes garnered from existing human gut microbiome datasets. Of these two resources, UHGP is currently the most comprehensive, and preferable because it retains readily available information on from which genome(s) a given gene stems. Neither

DrugBug (built on a custom database of 491 bacterial genomes) nor MicrobeFDT (built on IMG's 3,008 bacterial genomes) utilized these resources, thus limiting the completeness of their predictions (Markowitz et al., 2012; Sharma et al., 2017).

MetaCyc and KEGG are currently the most informative literature-curated pathway databases, both containing more bacterial reactions than BRENDA, Reactome, or Rhea (Altman et al., 2013). In terms of single-step reactions, MetaCyc (version 24.1) exceeds KEGG (Release 104.0) from a numbers standpoint (16,810 reactions versus KEGG's 11,841). From a protein perspective too, MetaCyc exceeds KEGG in reaction representation (237,506 bacterial enzymes to KEGG's 10,249 bacterial enzymes). Interestingly, the overlap between MetaCyc and KEGG is not as high as one would assume and was last calculated as only 1,961 reactions in common (Altman et al., 2013). Combining both bacterial chemistry databases, therefore, would be the most comprehensive representation of bacterial chemistry. Interestingly, neither DrugBug nor MicrobeFDT utilized reaction databases, instead opting for the simplicity of substrate databases. This decision drastically affected both methods' accuracy as will be touched on in Section 1.3.3 and explored in Chapter 2.

Any microbiome enzyme prediction tool, even one drawing on the most comprehensive information available for bacterial reactions, will suffer from the underexplored nature of microbiome chemistry. For this reason, populating and updating MetaCyc and KEGG with characterized microbiome chemistry as it emerges is of critical importance. As it stands, both databases only contain a handful of single-step reactions specific to bacteria in the human gut microbiome. To circumvent this problem currently, researchers may turn to pharmacomicrobiomic-specific databases, though these still leave room for improvement in terms of completeness and updates (Aziz et al., 2018; Rizkallah et al., n.d.; Sun et al., 2018; Zeng et al., 2021).

1.3.2 Functional transfer between databases is sensitive to protein search algorithms

When drawing connections between existing bacterial reaction databases and microbiome catalogs, it is necessary to use appropriate protein searches that do not rely on functional annotations, as annotations are missing ~40% of the time in microbiome gene collections (Almeida et al., 2020; Thomas and Segata, 2019). When comparing database entries instead based on sequence similarity, the algorithm employed plays a critical role. Homology searches in pharmacomicrobiomics research are often conducted using pairwise search algorithms such as BLAST (Altschul et al., 1990). A limitation of this method is that substitutions, deletions, and insertions are penalized by a set amount, regardless of where in the alignment they occur. For a given collection of functional enzymes, however, sequence conservation varies at different sites in the protein, as a result of differing strengths of selection pressures on different residues (i.e. high conservation at active sites versus low conservation in disordered domains). This position-specific information can be leveraged by performing homology searches with profile Hidden Markov Models (pHMMs), which encode protein family evolutionary patterns present in a multiple sequence alignment (Eddy, 1998). In the antibiotic resistance protein space, for example, pHMMs that incorporate position specific information have found distant homologs with retained function not recovered via pairwise search methods (Gibson et al., 2015). pHMM searches are an improvement over BLAST from the standpoint of finding distant homologs and from the standpoint of finding targets with retained activity, as previous research has shown that global sequence identity does not necessarily map to similar function (Gerlt et al., 2012).

Again, current enzyme prediction methods do not employ best practices for sequence search algorithms, with both DrugBug and MicrobeFDT utilizing functional annotation transfer based on Enzyme Commission (EC) codes (Guthrie et al., 2019; Sharma et al., 2017). EC codes are four-digit identifiers of enzyme-driven reactions, where each digit describes the reaction with increasing chemical granularity (McDonald et al., 2009). The first digit, the EC class, describes broad chemistry such as whether the reaction is an oxidoreduction, hydrolysis, etc. event. The second and third digits, the EC sub-class and sub-sub-class, describe more detailed chemical information such as electron donor or transfer group identity. The fourth and final digit, the EC serial designation, often describes a reaction's substrate specificity. All in all, while EC codes are helpful for describing reaction types, they are not sufficient for functional predictions of microbiome orthologues due to the paucity of EC annotations in this dataspace. EC entries themselves are sometimes incomplete, with about 36% of assigned EC numbers lacking either a gene or protein sequence (Pouliot and Karp 2007). Lastly, the EC fails to adequately express the complexity of peptidase enzymes (EC 3.4, about 10% of the enzymes classified by EC). All peptidases catalyze a nearly identical reaction, hydrolysis of a peptide bond, and as a result, the EC has lumped peptidases of diverse functions into only a few low resolution sub-classes (McDonald and Tipton 2021; McDonald and Tipton 2014). For this reason, chemists instead turn to the MEROPS database that classifies peptidases based on structural features and evolutionary relationships (McDonald and Tipton 2021; McDonald and Tipton 2014; Barrett et al. 2001).

1.3.3 The importance of choosing appropriate chemical representations

Another foundational aspect of building predictive tools from existing chemistry-protein databases is the use of appropriate chemical representations, as any query compound will be computationally compared to database reaction entries. Chemical representations for the problem at hand need to be considered in two ways, first—how individual molecules are represented, and second—whether and how to represent an entire reaction. To the first point, many types of chemical fingerprints (boolean arrays or bitmaps representing patterns found in a molecule) exist, each with pros and cons that must be weighed depending on the chemical search in which they will be used (Capecchi et al., 2020; Duan et al., 2010; Schneider et al., 2015). Table 1.2 is a non-exhaustive list, but describes the most common 2D chemical fingerprint options found in chemoinformatics software packages. Various benchmarking studies have demonstrated that structural keys (like MACCS) exhibit the worst performance, being unable to discriminate between two molecules when highly similar (Capecchi et al., 2020; Duan et al., 2010; Wild and Blankley, 2000). Of the three hashed representations in Table 1.2, all perform better than structural keys in benchmarking studies, but of note, ECFP fingerprints do not capture global molecular details such as size and shape of a chemical species (Capecchi et al., 2020). To date, all microbiome metabolism tools have employed structural key fingerprints (Guthrie et al., 2019; Mallory et al., 2018; Sharma et al., 2017).

Table 1.2 Chemical fingerprint options

Fingerprint Type	Description
Structural keys	Boolean array representing the presence/absence of 155 (MACCS) or 881 (PubChem) predefined structural features (e.g., “at least one Nitrogen”).
Atom-Pair (AP)	Hashed representation of all pairwise atoms (plus number of heavy neighbors and number of pi electrons) in a molecule and the shortest topological distance separating them.
Topological Torsion (TT)	Hashed representation of each length 4 (atoms) linear path and all atoms along the path.

Fingerprint Type	Description
Extended connectivity fingerprint (ECFP)	Hashed representation of all atoms and bonds within nested fragments grown radially from a heavy atom center. Also known as Morgan fingerprints.

The second consideration for representing chemistry in an enzyme prediction tool, is the use of substrate versus reaction fingerprints. Current microbiome enzyme prediction tools solely consider substrates when making their predictions, resulting in low accuracy output further discussed in Chapter 2 (Guthrie et al., 2019; Sharma et al., 2017). Biotransformations (described in databases like MetaCyc and KEGG) involve the relationship between substrate(s), cofactor(s), and an enzyme to yield a particular product(s). As one substrate can exhibit affinity for multiple enzymes, resulting in multiple unique products, sole employment of substrates in a chemical fingerprint does not achieve the resolution necessary to make relevant predictions. Recent research (Mallory et al., 2018) appropriately employed both substrate and product chemistry, by representing reactions as a single fingerprint vector (resulting from the difference between a product's and substrate's vector); these fingerprints were used to compare bacterial-drug metabolism events to primary reactions in the MetaCyc database, but without the end-goal of enzyme identity prediction. From a reaction description standpoint, the published method was still limited in that it only included a description of one substrate and one product per reaction, precluding it from utilizing cofactors and from accurately describing transformations that employ multiple substrates and/or produce multiple products. To circumvent this pitfall, chemical representations should describe multiple inputs and outputs for a single reaction (Schneider et al., 2015).

1.4 The future of microbiome metabolism prediction tools

In my thesis work that follows, I address the above issues with current microbiome metabolism prediction tools, and implement my recommendations. What results is a novel computational tool for exploring species and enzymes capable of performing any query reaction of interest. The architecture and computational validation of this tool is described in Chapter 2. I demonstrate the utility of my method by using it to corroborate previously collected information on xenobiotic transformations occurring in the human gut. Finally, in Chapter 3, I present preliminary experimental findings demonstrating my method's ability to predict bacterial species responsible for the degradation of methotrexate (MTX), an anti-arthritic compound which only exhibits efficacy in half of the prescribed population (Scher et al., 2020). In sum, this work represents a clear departure from and improvement over previous methods, and lays the foundation for future characterizations of xenobiotic degradation in the human gut, demonstrated here with the first species-level characterization of Methotrexate metabolism by bacteria in the human gut microbiome.

Chapter 2

A novel *in silico* method employs chemical and protein similarity algorithms to accurately identify chemical transformations in the human gut microbiome.

2.1 Introduction

Humans consume a large array of foods, therapeutics, and other xenobiotics that are processed, in part, by enzymes of bacteria residing within the gut. While some bacterial enzymes are orthologous to the human metabolism repertoire, many bacteria possess metabolic capabilities distinct from our own (Zimmermann et al., 2019a). It is important to ascertain the extent of microbial capacity for chemical transformation because it has implications for the bioavailability, toxicity, and efficacy of the compounds humans ingest (Koppel et al., 2017; Spanogiannopoulos et al., 2016). Additionally, because the human gut microbiome differs from individual to individual, knowledge of the prevalence and abundance of bacterial enzymes must be determined before beneficial clinical and dietary decisions can be made (Javdan et al., 2020).

While experimental methods can be employed to expand what we know of bacterial enzymatic capabilities in the gut, the scientific community lacks genetic tools for nearly all bacterial species of the human microbiota, and heterologous expression in model organisms can fail for a plethora of reasons (Bisanz et al., 2020; Patel et al., 2022). When experimental methods are tractable, the time required is often so extended that knowledge is gained in a low-throughput manner. For these reasons, attention should turn to the employment of *in silico* computational methods that can guide experimentalists in their hypothesis-building process by aiding in the prioritization of substrates, species, and genes worth studying.

As described in Chapter 1, recent attempts have been made to create computational descriptions of chemical transformation by human gut bacteria, but none can be expanded to predict the metabolic capabilities of bacterial proteins with unknown function or to explore the capacity of microbial enzymes to degrade novel substrates. Two previously published methods aimed to predict known drug metabolism events within the human gut microbiome, but the accuracy of their predictions was limited due to the fact that both tools only consider substrates, rather than a full chemical description of substrate(s), cofactor(s), and product(s) formed in a reaction. Both tools were also limited by the use of small databases that do not fully capture the diversity of the human gut microbiome (Guthrie et al., 2019; Sharma et al., 2017).

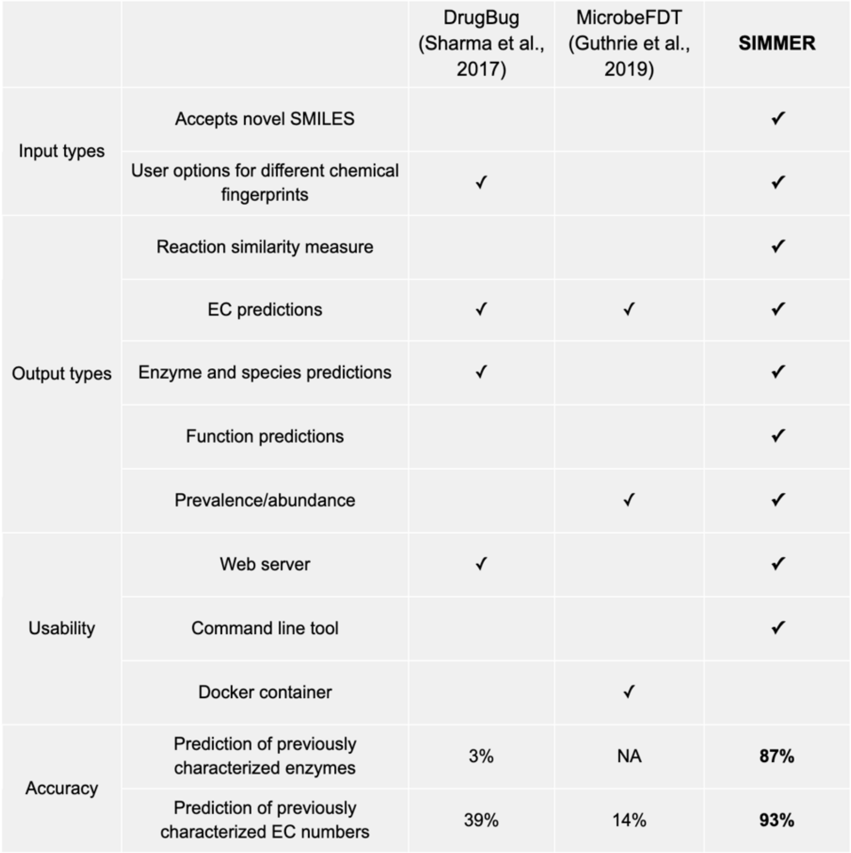
To address these gaps in accurate predictive software for bacterial chemical transformations, I present SIMMER, a tool that combines chemoinformatics and metagenomics approaches to accurately predict bacterial enzymes capable of metabolism events. Given an input reaction, SIMMER predicts an Enzyme Commision (EC) code that describes the chemical nature of the query. SIMMER additionally predicts specific bacterial enzyme sequences, functions, prevalence, and abundance for the reaction. Our key innovations are the use of full chemical representations that include cofactors employed and products produced in a reaction, the use of statistically informed sequence searches of a comprehensive human gut microbiome gene catalog, and the development of a novel EC predictor based on reaction rather than gene sequence. As a use-case, I present evidence that SIMMER provides high accuracy predictions of bacterial enzymes responsible for known drug metabolism events, and I identify the likely bacterial enzyme for 88 drugs known to be metabolized by the gut microbiome for which the enzyme was previously unknown.

2.2 Results

2.2.1 SIMMER pipeline to predict xenobiotic metabolizing enzymes

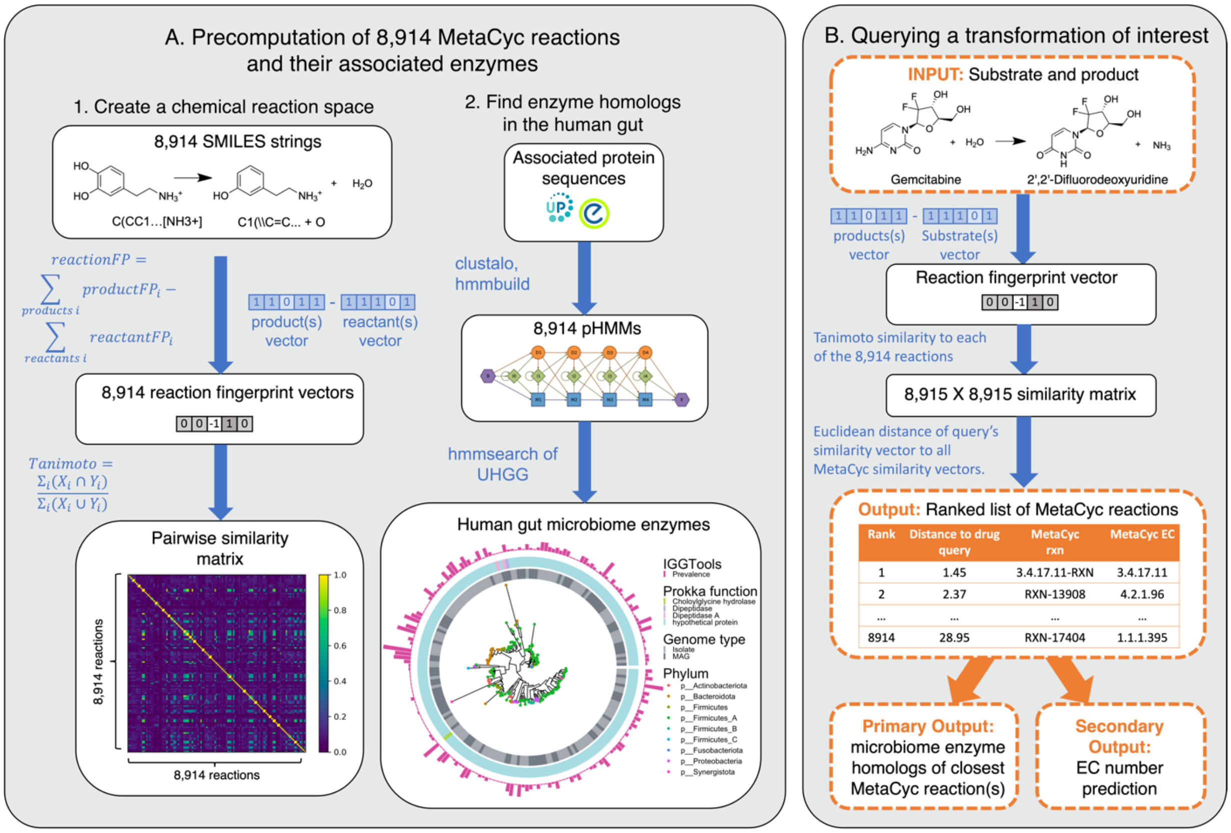
There are many desiderata for a bacterial drug metabolism predictor (Table 2.1). Such a tool must be able to, based on quantified chemical similarity, predict EC annotations, specific enzyme sequences, and the prevalence and abundance across human samples of those predicted sequences. I developed SIMMER, a tool that leverages chemical and protein similarity to identify enzymes in the human microbiome that could perform a queried chemical reaction (Figure 2.1). Given input substrate(s), metabolite(s), and any known cofactors, SIMMER predicts bacterial enzymes capable of performing the reaction and quantifies their prevalence and abundance in the human gut. SIMMER accomplishes this by chemically fingerprinting an input reaction, and then comparing it to all reactions in MetaCyc. Enzyme annotations from the most similar MetaCyc reactions

Table 2.1 Desired elements of a microbiome chemical transformation predictor.



		DrugBug (Sharma et al., 2017)	MicrobeFDT (Guthrie et al., 2019)	**SIMMER**
Input types	Accepts novel SMILES			✓
	User options for different chemical fingerprints	✓		✓
Output types	Reaction similarity measure			✓
	EC predictions	✓	✓	✓
	Enzyme and species predictions	✓		✓
	Function predictions			✓
	Prevalence/abundance		✓	✓
Usability	Web server	✓		✓
	Command line tool			✓
	Docker container		✓	
Accuracy	Prediction of previously characterized enzymes	3%	NA	**87%**
	Prediction of previously characterized EC numbers	39%	14%	**93%**

are then used as queries for a protein similarity search to find homologs in the genomes of gut bacteria. To decrease the runtime of a SIMMER query, I precomputed chemical descriptions and protein similarity searches for all reactions in MetaCyc.


A. Precomputation of 8,914 MetaCyc reactions and their associated enzymes
1. Create a chemical reaction space
8,914 SMILES strings
HO
HO
NH_3^+
HO
NH_3^+
+ H_2O
C(CC1...[NH3+]
C1(\\C=C... + O
$reactionFP = \sum_{products\ i} productFP_i - \sum_{reactants\ i} reactantFP_i$
1 1 0 1 1 - 1 1 1 0 1
product(s) vector
reactant(s) vector
8,914 reaction fingerprint vectors
0 0 -1 1 0
$Tanimoto = \frac{\Sigma_i (X_i \cap Y_i)}{\Sigma_i (X_i \cup Y_i)}$
Pairwise similarity matrix
8,914 reactions
8,914 reactions
1.0
0.8
0.6
0.4
0.2
0.0
2. Find enzyme homologs in the human gut
Associated protein sequences
clustalo, hmmbuild
8,914 pHMMs
hmmsearch of UHGG
Human gut microbiome enzymes
IGGTools
Prevalence
Prokka function
Choloylglycine hydrolase
Dipeptidase
Dipeptidase A
hypothetical protein
Genome type
Isolate
MAG
Phylum
p__Actinobacteriota
p__Bacteroidota
p__Firmicutes
p__Firmicutes_A
p__Firmicutes_B
p__Firmicutes_C
p__Fusobacteriota
p__Proteobacteria
p__Synergistota
B. Querying a transformation of interest
INPUT: Substrate and product
F
OH
OH
F
H_2N
N
N
O
O
+ H_2O
F
OH
OH
F
O
N
N
H
O
O
+ NH_3
Gemcitabine
2',2'-Difluorodeoxyuridine
1 1 0 1 1 - 1 1 1 0 1
products(s) vector
Substrate(s) vector
Reaction fingerprint vector
0 0 -1 1 0
Tanimoto similarity to each of the 8,914 reactions
8,915 X 8,915 similarity matrix
Euclidean distance of query's similarity vector to all MetaCyc similarity vectors.
Output: Ranked list of MetaCyc reactions
Primary Output: microbiome enzyme homologs of closest MetaCyc reaction(s)
Secondary Output: EC number prediction

Rank	Distance to drug query	MetaCyc rxn	MetaCyc EC
1	1.45	3.4.17.11-RXN	3.4.17.11
2	2.37	RXN-13908	4.2.1.96
...	...	...	...
8914	28.95	RXN-17404	1.1.1.395

Figure 2.1 SIMMER architecture.
(A) Precomputation on 8,914 gene annotated bacterial reactions downloaded from MetaCyc. Chemical fingerprints representing each MetaCyc reaction were created from SMILES descriptors. A latent chemical space was then created via a pairwise reaction similarity matrix based on Tanimoto coefficients. For each reaction, relevant gene sequences were retrieved from UniProt and Entrez database linkouts and used to create multiple sequence alignments and subsequent pHMMs using ClustalO and HMMER3 respectively. pHMMs were used to retrieve homologs in a catalog of human gut microbiome genes. **(B)** Running a SIMMER query. After receiving a reaction query (input compound, co-factors, products), SIMMER fingerprints the reaction and compares it to the precomputed chemical space from 1A. MetaCyc reactions are sorted by similarity to the query. Significantly enriched EC identities are reported, and from the most similar reaction, human gut microbiome enzymes are reported along with their abundance and prevalence in gut microbiomes.

SIMMER's underlying data was drawn from the MetaCyc reaction database because its small-molecule reaction descriptions each possess at least one experimentally validated enzyme annotation and often include a description of the reaction type via EC code. To build a precomputed chemical search space for SIMMER queries (Figure 2.1A), I created two-dimensional fingerprint representations for 8,914 enzyme driven reactions in MetaCyc (Caspi et al., 2008; Schneider et al., 2015). Using these fingerprints, I estimated the similarity between all pairs of reactions based on Tanimoto coefficients. To build the enzyme backbone of SIMMER, I compiled the Uniprot and/or Entrez gene identifiers linked to each MetaCyc transformation into a profile hidden Markov model (pHMM) that represents the diversity of the enzyme family for a respective reaction. The resulting pHMMs were then used to query the Unified Human Gastrointestinal Genome (UHGG) collection of 286,997 isolate genomes and metagenome assembled genomes from the human gut environment (Almeida et al., 2020). Additionally, prevalence and abundance of all pHMM search hits were assessed in stool metagenomes from the PREDICT human cohort using MIDAS2, an implementation of Metagenomic Intra-Species Diversity Analysis System (MIDAS) designed for use with the UHGG catalog (Almeida et al., 2020; Nayfach et al., 2016; Zhao et al., 2022).

After creating SIMMER's precomputed chemical and pHMM search space, I next made SIMMER queryable (Figure 2.1B). When queried with a chemical transformation, SIMMER computes the chemical similarity of the input to all precomputed MetaCyc reactions, and sorts all MetaCyc reaction fingerprints according to their ascending Euclidean distance from the query. From this sorted list, SIMMER outputs enzymes (i.e. the precomputed pHMM search hits for the closest reactions) responsible for the query reaction and an EC code (i.e. reaction type) prediction. I implemented and validated a novel method to predict EC codes by extending a common approach to gene set enrichment analysis (GSEA) (Figure 2.2—figure supplement 1) (Subramanian et al., 2005). With this enrichment method, SIMMER predicted reaction types for queries with high recall, precision, and accuracy for EC classes, sub-classes, and sub-sub-classes (Figure 2.2B, Figure 2.2—source data, Figure 2.2—figure supplement 1).

I hypothesized that SIMMER accurately predicted chemical transformations due to its use of a full reaction that includes reactants, cofactors, and products, rather than just substrates. I assessed this hypothesis by demonstrating that SIMMER groups similar reactions together in chemical space. MetaCyc reactions possess EC annotations that describe the chemical class of a reaction (e.g. oxidoreduction, hydrolysis, intramolecular rearrangement, etc). I queried SIMMER with all EC annotated MetaCyc reactions and demonstrated that queries group significantly closer to other reactions within their EC class than they do to reactions of a different class (Figure 2.2A). I determined that SIMMER's ability to group similar reactions in chemical space is resilient to different fingerprinting methods (Figure 2.2—figure supplement 2), but not to loss of products created and cofactors consumed in a reaction (Figure 2.2—figure supplement 3). Thus I showed that similar reactions only cluster together in chemical space when a full reaction description (i.e. SIMMER's representation method) is employed.

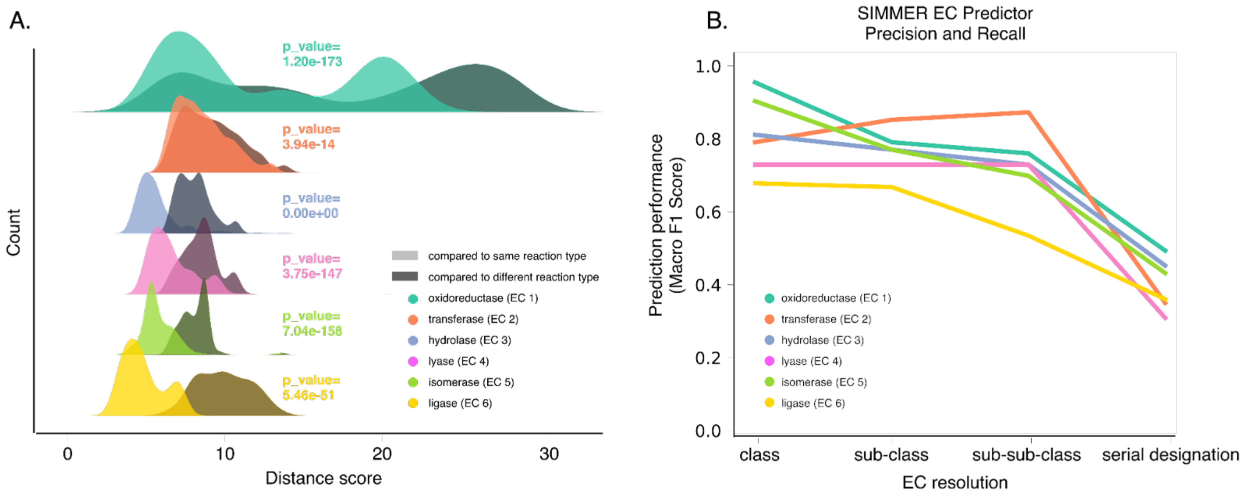


Figure 2.2 SIMMER's chemical representations capture information relevant to enzymatic reactions.
(A) SIMMER clusters similar reactions together in chemical space. To analyze SIMMER's ability to group chemically similar reactions, I examined reaction similarity within versus between EC classes using the procomputed MetaCyc reaction dataset (N=8,914 reactions). A silhouette-like euclidean distance score was created by determining for each reaction its euclidean distance to all reactions within its EC class versus outside its EC class. For all EC classes, scores were smaller within versus between EC classes using SIMMER's chemical representation, indicating that SIMMER can detect reaction similarity within EC classes. From the pairs of distributions I computed a Kolmogorov-Statistic to determine if the distributions significantly ($p<0.05$) differed. **(B)** The F1-score, or harmonic mean of SIMMER's precision and recall, when predicting EC numbers on a subset of the MetaCyc database (N=576 reactions total; 96 per EC class). The score is high for EC classes, and it generally decreases as an EC number's resolution increases.

Because SIMMER was created with the assumption that chemically similar reactions are mediated by sequence similar enzymes, I next ensured that similarity within SIMMER's chemical space could be used to find shared, responsible enzymes. First, for all MetaCyc enzymes associated with multiple reactions, one reaction was used as a SIMMER query, and the second reaction searched for in the ordered reaction list output. As a negative control, these reaction similarity results were then compared to all possible pairwise combinations of reactions not conducted by the same enzyme. SIMMER predicted high similarity between reactions conducted by a shared enzyme, and low similarity for those reactions without a shared enzyme, (Figure 2.3A). I also found a negative association between chemical reaction distance and sequence similarity of MetaCyc enzyme annotations, indicating that reactions with similar chemistry are conducted by sequence similar enzymes, though there is much variation in this relationship (Figure 2.3B). This reflects the known association between sequence similarity and similarity in chemical function, as well as reports that this relationship can often be overestimated (Tian and Skolnick, 2003). Together, these analyses demonstrated that sequence similar enzymes do indeed mediate chemically similar reactions, strengthening the logic of combining chemical and protein similarity in a microbiome enzyme prediction tool.

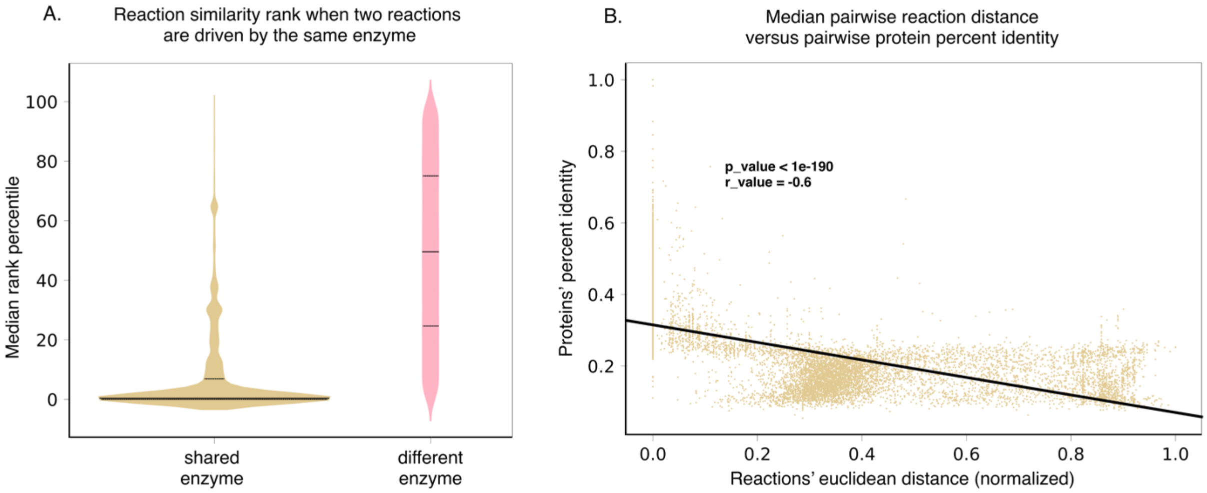


Figure 2.3 SIMMER's chemical representations can be used to find shared, responsible enzymes.
(A) When SIMMER was queried with a MetaCyc reaction, other reactions driven by the same enzyme are returned as the most similar. As a contrast, reactions driven by a different enzyme yield a more uniform rank distribution. Solid lines of the violin plots depict median reaction similarity rank and dashed lines represent lower and upper quartile ranges. **(B)** Similarly, there is a negative association between pairwise reaction euclidean distance and pairwise protein identity, demonstrating that SIMMER can capture the known, albeit weak relationship between sequence identity and similar reaction chemistry (Tian and Skolnick, 2003).

2.2.2 An expanded list of gut bacterial enzymes relevant to known cases of drug metabolism

To assess SIMMER's prediction accuracy for previously characterized reactions and to mount a comparison to existing methods, I used drug metabolism as a use case. First, I curated 298 drug-metabolism events associated with the human gut microbiome from the literature (Supplementary File 1). For 31 of these reactions the responsible bacterial enzyme, characterized metabolite(s), and associated EC annotation are known (Supplementary File 1, Table 1.1). These 31 reactions are conducted by 18 enzymes. Due to orthology and proclivity for genetic transfer between even distantly related bacteria, however, there are likely many as yet undiscovered homologs of these drug-metabolizing enzymes that can catalyze identical drug metabolism events (Pollet et al., 2017). To account for this, I created an expanded database (Figure 2.4, Figure 2.4—figure supplement 1, Figure 2.4—source data) of the 18 characterized enzymes from pHMM and phmmer searches of the UHGG database (Almeida et al., 2020), yielding 52,849 total candidate homologs (a median of 1,087 candidates per enzyme). After filtering enzymes by hmmer significance, alignment length, presence in data from the human jejunum (Zmora et al., 2018) and RNA-sequencing studies (Integrative HMP (iHMP) Research Network Consortium, 2019), and predicted affinity for the substrate in question using the Similarity Ensemble Approach (Keiser et al., 2007), our database contained a

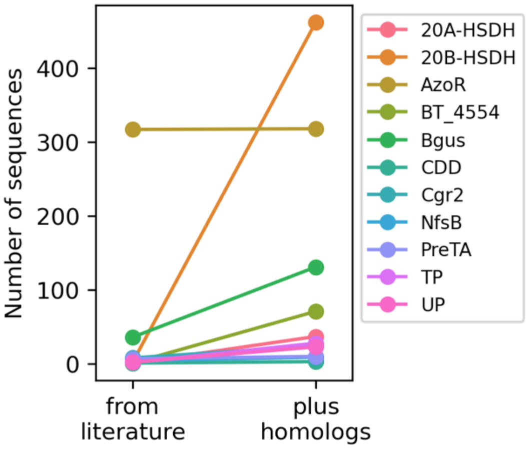


Figure 2.4 An expanded list of gut bacterial enzymes relevant to known cases of drug metabolism. Eleven of the 18 enzymes responsible for positive control drug-metabolism events have high confidence homologs that I gathered by filtering for biological significance.

median of 2 high-confidence homologous sequences per enzyme (range = 0 to 460 across the 18 enzyme families, Figure 2.4, Figure 2.4—figure supplement 1, Figure 2.4—source data). These 741 additional enzyme sequences for 31 reactions formed our positive control set of known gut bacterial enzymes.

2.2.3 SIMMER captures known gut bacterial enzymes involved in drug metabolism

With our expanded database of drug-metabolizing enzymes from the human gut microbiome in hand, I next verified that SIMMER can accurately predict reaction types and responsible enzymes for the 31 known chemical transformations. Only 3 of these 31 reactions are themselves MetaCyc entries (5-ASA, dopamine, and levodopa degradation); if EC codes and enzymes of reactions not described in MetaCyc were also accurately predicted, it would show that SIMMER can discover non-identical yet chemically similar reactions.

Of the 31 drug-metabolism events known to occur via human gut bacterial enzymes, EC annotations exist for 28. SIMMER identified the correct EC class for 26 of these 28 reactions (93%) (Figure 2.5, Figure 2.5—source data). For some queries SIMMER predicted more than one significant EC code, but again, for 26 out of the 28 reactions, the top EC class prediction was a match (Figure 2.5, Figure 2.5—figure supplement 1, Figure 2.5—source data). The two failed EC predictions were for nicardipine reduction (inappropriately predicted as an isomerase reaction) and for brivudine transformation (for which SIMMER made no significant prediction).

In addition to accurate EC (i.e. reaction type) identification, SIMMER also accurately predicted the specific enzymes from the human gut microbiome that conduct the 31 query reactions (Figure 2.5—source data). This enzyme list was populated by the results of the precomputed pHMM searches of human microbiome catalogs with annotated gene sequences from MetaCyc reactions (Figure 2.1A). In 27 cases (87%), the characterized (i.e. positive control) enzyme(s) for a reaction was found in the output enzyme list for the top 20 of the ranked MetaCyc reactions (Figure 2.5—figure supplement 1). Since the positive controls span four EC classes (EC1 oxidoreductases, EC2 transferases, EC3 hydrolases, and EC4 lyases) this result demonstrates SIMMER's ability to accurately predict microbiome based enzymes for a diversity of reaction types. Also, despite inaccurate EC predictions for nicardipine reduction and brivudine transformation, SIMMER was able to respectively predict AzoR and BT_4554 enzymes as responsible for the reactions (Figure 2.5, Figure 2.5—figure supplement 1, Figure 2.5—source data).

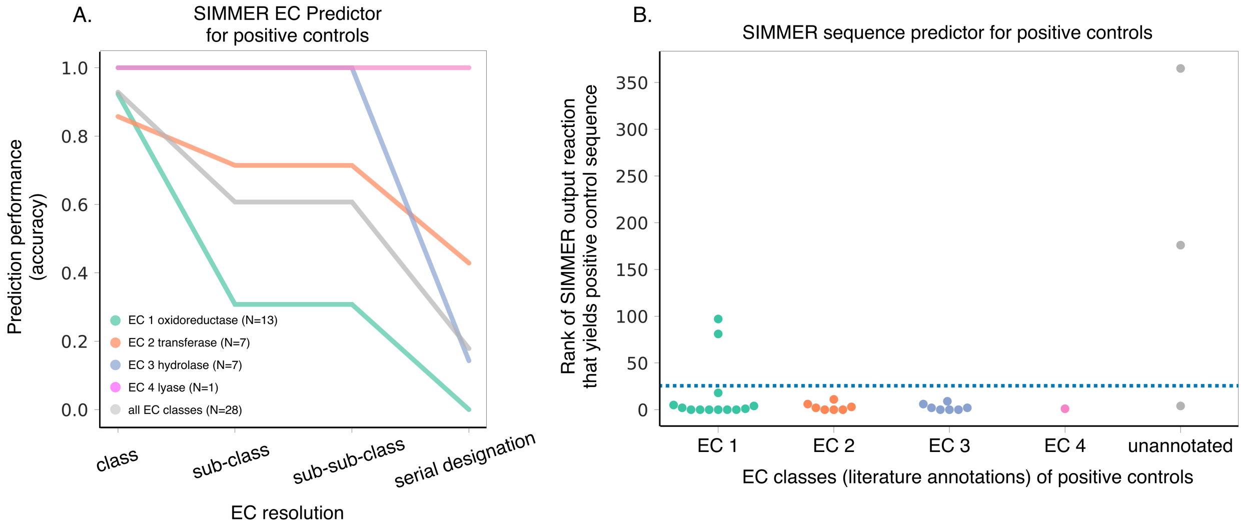


Figure 2.5 SIMMER captures known gut bacterial enzymes involved in drug metabolism. (A) SIMMER accurately predicted EC classes for 28 previously characterized reactions that possess EC annotations. As with the MetaCyc database (Figure 2.2B), accuracy dropped off as EC resolution increased. **(B)** SIMMER predicted bacterial sequences previously shown to drive 31 drug-metabolism events in the gut microbiome. Depicted is the rank (out of N=8,914 reactions) of the MetaCyc reaction that yielded a gut microbiome homolog matching the known positive control sequence. Reported accuracy is based on such a reaction being within the top 25 ranked reactions (dashed blue line).

2.2.4 SIMMER outperforms existing methods

To mount a comparison to the other *in silico* methods that, in part, aimed to describe microbiome drug metabolism, I next queried the 31 positive control reactions using MicrobeFDT and DrugBug (Table 2.1) both of which rely solely on substrate chemical similarity rather than information from a whole reaction (Guthrie et al., 2019; Sharma et al., 2017). For the 28 EC annotated positive control reactions, DrugBug had 39% accuracy in predicting EC classes (in comparison to SIMMER's 93%), and predicted the correct enzyme for a single reaction, SN38 glucuronide deconjugation, despite the presence of chemically similar reactions metabolized by the same enzyme amongst the positive controls (Table 2.1—source data). I additionally queried SIMMER with the four positive controls (ginsenoside Rb1, quercetin-3-glucoside, cycasin, and sorivudine) associated with characterized bacterial enzymes from the original DrugBug publication (Table 2.1—source data). Both DrugBug and SIMMER were able to predict EC classes for sorivudine, but only SIMMER was able to accurately predict the specific enzyme (BT_4554) responsible for the drug's degradation. For ginsenoside Rb1 (3.2.1.192), quercetin-3-glucoside (3.2.1.21), and cycasin (3.2.1.21), SIMMER was able to accurately predict EC codes out to sub-sub-class (3.2.1.-), serial designation (3.2.1.21), and sub-sub-class (3.2.1.-) respectively, which was a resolution improvement over DrugBug (Table 2.1—source data).

I next queried the 28 EC annotated drug-metabolism positive controls against MicrobeFDT (which uses a chemical graph to predict EC codes, but not enzymes) in two ways: first by looking for direct enzyme metabolism events, and second, by looking for enzyme metabolism of compounds that overlap chemically with the positive control in question. When directly queried, MicrobeFDT produced metabolism predictions for four of the 28 positive controls, three of which were correct. When queried with chemically similar compounds according to its graph, MicrobeFDT produced metabolism predictions for 13 of the 28 positive controls, and four were correct (14% overall accuracy in comparison to SIMMER's 93%). It is important to note that MicrobeFDT is reliant on a fixed database that cannot be modified by the user, meaning that a compound cannot be queried if it is not already present in MicrobeFDT's graph. I finished the comparison between SIMMER and MicrobeFDT by querying SIMMER with the metabolism use-case described in the MicrobeFDT publication, altretamine demethylation. In our hands, there was no Cypher query against the MicrobeFDT database that resulted in a demethylase EC code; I determined possible demethylase EC codes by running a query in the Swiss Institute for Bioinformatics Enzyme Nomenclature Database (Bairoch, 2000). I performed queries of direct EC annotation for melamine and altretamine, as well as EC annotation queries for any compound with either substructure or toxicity overlap with altretamine or melamine. The closest result to a demethylase enzyme was a cypher query of toxicity overlap with altretamine that yielded a nitric oxide synthase (EC 1.14.13.39) acting on L-arginine among its results (Table 2.1—source data). For its significant EC (reaction type) prediction, SIMMER identified altretamine demethylation appropriately as an oxidoreductase reaction acting on a CH-NH group of donors (EC 1.5.-), but not significantly as a demethylation event (Table 2.1—source data).

This comparison illustrates SIMMER's enhanced accuracy over other methods for the use case of characterized drug metabolism events by gut bacteria, and also illustrates SIMMER's novel ability to predict chemical transformations not previously described in literature or databases.

2.2.5 SIMMER predicts novel drug-metabolizing enzymes

After establishing SIMMER's accuracy in predicting drug-metabolizing enzymes in the human gut environment, I predicted EC codes, functional annotations, and enzyme sequences for novel microbiome drug metabolism reactions that do not yet possess a responsible, characterized enzyme (Figure 2.6—source data). From our literature curation of 298 non-antibiotic therapeutics affected by the microbiome (Supplementary File 1), I was confident that 88 are directly metabolized by gut bacteria due to their association with an identified bacterial metabolite in the literature. I formatted these 88 reactions in SMILES format and input them as queries to SIMMER.

Of the 88 reactions queried, SIMMER determined significant EC predictions for 75 reactions (86.2%), and 61 (70.1%) of these were out to the serial designation (i.e. highest resolution) EC code (Figure 2.6—source data). This list of 61 transformations presents reactions for which I believe enzyme characterization is worth pursuing as our predictions are significantly similar to enzymes already explored in the literature. SIMMER's EC predictions resulted in expanded and diversified EC class membership for drug-transformations known to occur in the microbiome (Figure 2.6C). Of interest, this analysis resulted in a large expansion of putative hydrolysis, reduction, and isomerization reactions in the human gut microbiome. The number of SIMMER predictions varies widely by reaction, with median output of 372 genes, 286 genomes, and 10 phyla predicted as responsible across the 88 reactions (Figure 2.6—source data, Figure 2.6A). Unsurprisingly, many of these reactions are predicted to occur due to enzymes found in Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria, but there are also SIMMER enzyme predictions in phyla not previously associated with drug metabolism (Figure 2.6B).

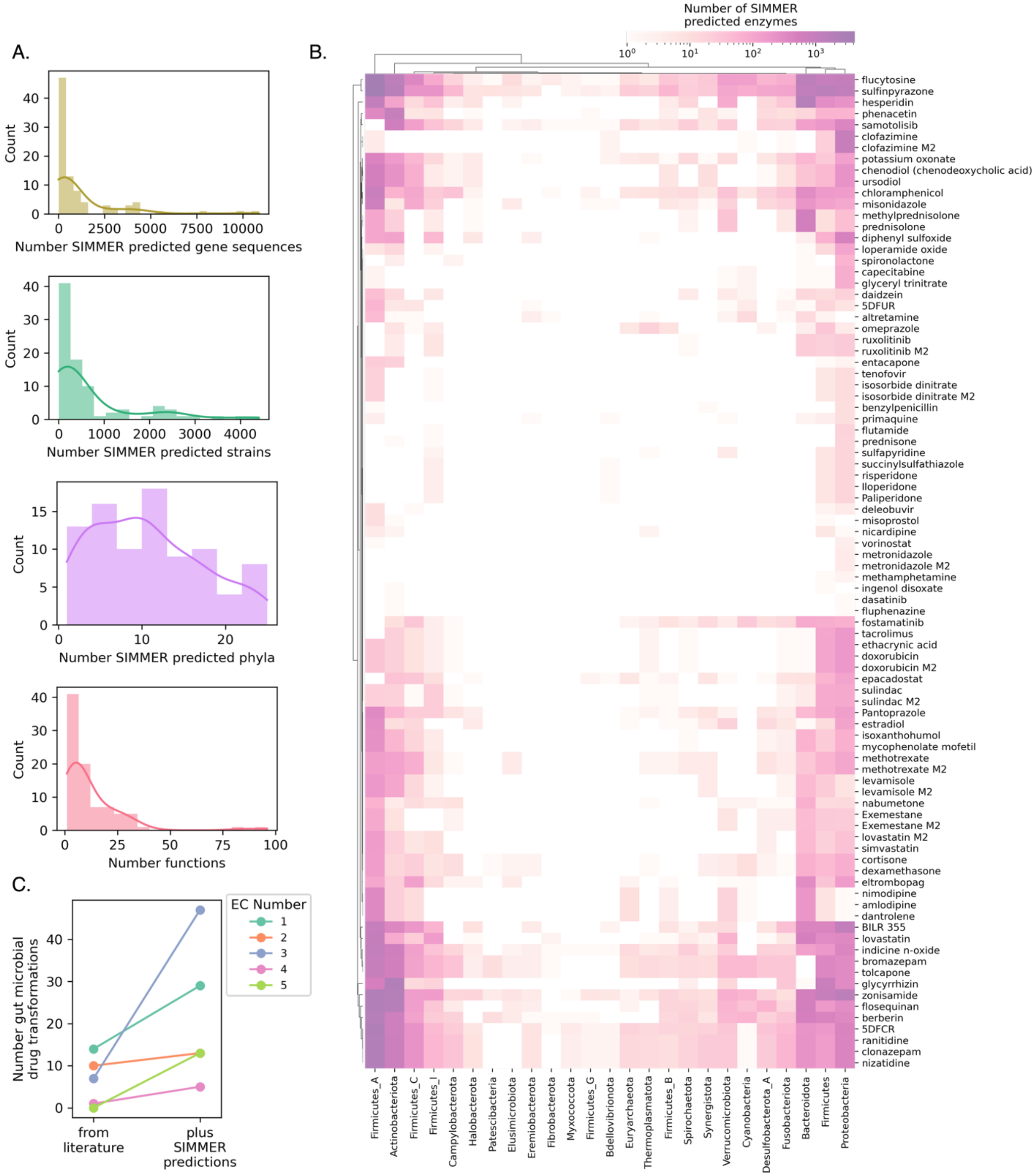


Figure 2.6 SIMMER predicts novel drug metabolizing enzymes.
(A) Distributions depict the unique number of genes, strains, and phyla predicted to be responsible for 88 reported drug transformation reactions, as well as predicted gene functions. **(B)** A heatmap illustrating the number of phyla from the UHGG database capable of performing 88 drug metabolism events. Color intensity refers to the number of unique drug-metabolizing enzymes for a given phylum conducting a given reaction. **(C)** Enzyme Commission Class representation for bacterial transformations of therapeutics before and after the employment of

SIMMER. My predictions greatly expand the number of characterized reduction (EC1), hydrolysis (EC2), and isomerization (EC5) events and modestly increase the number of transferase (EC2) and lyase (EC4) events.

Eight of the 88 novel transformations were among those investigated in a high-throughput study exploring the metabolism of 571 compounds in *ex vivo* stool samples (Javdan et al., 2020). This publication demonstrated bacterial degradation of 57 therapeutics in a single pilot donor stool sample (with associated shotgun sequencing), as well as in 20 human stool samples (with associated 16S rRNA gene sequencing). While this study greatly expanded the number of drugs known to break down in the presence of gut bacteria and identified eight metabolite structures, it only identified a responsible enzyme in two of the 57 drug degradation cases due to the low-throughput nature of enzyme characterization. To further assess SIMMER's ability to predict novel enzymes, and to demonstrate the utility of using SIMMER in an experimental context, I investigated the presence of my predictions in the Javdan, et al. study sequencing results. Because shotgun metagenomics sequencing for the pilot donor was deposited, I was able to confirm via tBLASTn searches that SIMMER enzyme predictions were directly found in the pilot donor stool sample for all eight of the reactions with identified metabolites (Figure 2.7—source data). However, the sequencing data from the 20 human donor study was only 16S profiling, so I was unable to look directly for SIMMER enzyme predictions. I was able to ensure that genomes found in metabolizing stool samples contain SIMMER predictions. I found that donors who could metabolize a given drug possessed a significant enrichment of genomes that contain enzymes predicted by SIMMER. This was the case for five out of the six reactions analyzed (Figure 2.7—source data, Figure 2.7A).

Among the 88 novel transformations was the side-chain cleavage of dexamethasone to 17-oxodexamethasone. Dexamethasone was recently shown to be metabolized solely by *Clostridium scindens* (ATCC 35704) out of a collection of 76 isolates representative of the human gut microbiome (Zimmermann et al., 2019b). When dexamethasone metabolism was assessed in 28 human stool samples, metabolite formation varied substantially by individual, but could not be explained by *C. scindens* species abundance. I sought to understand this lack of correlation. To do so, I assessed the abundance of *C. scindens* SIMMER predictions across the 28 samples (i.e., the amount of enzyme in genomes from different strains, not the amount of the species). I found a significant association between metabolite formation and number of SIMMER enzymes, and also a significant association between parent compound consumption and number of SIMMER enzymes (Figure 2.7B).

It came to my attention while preparing this manuscript that recombinant steroid-17,20-desmolase (DesAB) enzymes from *C. scindens* were shown to perform side-chain cleavage on prednisone, but also to a lesser extent on dexamethasone. DesAB's reduced activity for dexamethasone was assumed to be due to the compound's potentially inhibitory 16α-methyl group (Ly et al., 2020). To ensure that SIMMER's enzyme prediction for dexamethasone cleavage was not enriched in metabolizing stool samples due to co-occurrence with already known DesAB, I next assessed the abundance of *desAB* reads across the 28 samples, and found no significant correlation between number of reads and either metabolite formation or dexamethasone consumption slopes (Figure 2.7—figure supplement 1).

These results indicate that species level information alone is not enough to predict chemical transformations in a microbiome sample, but with SIMMER, knowledge of responsible enzymes can recapitulate a sample's potential for therapeutic degradation.

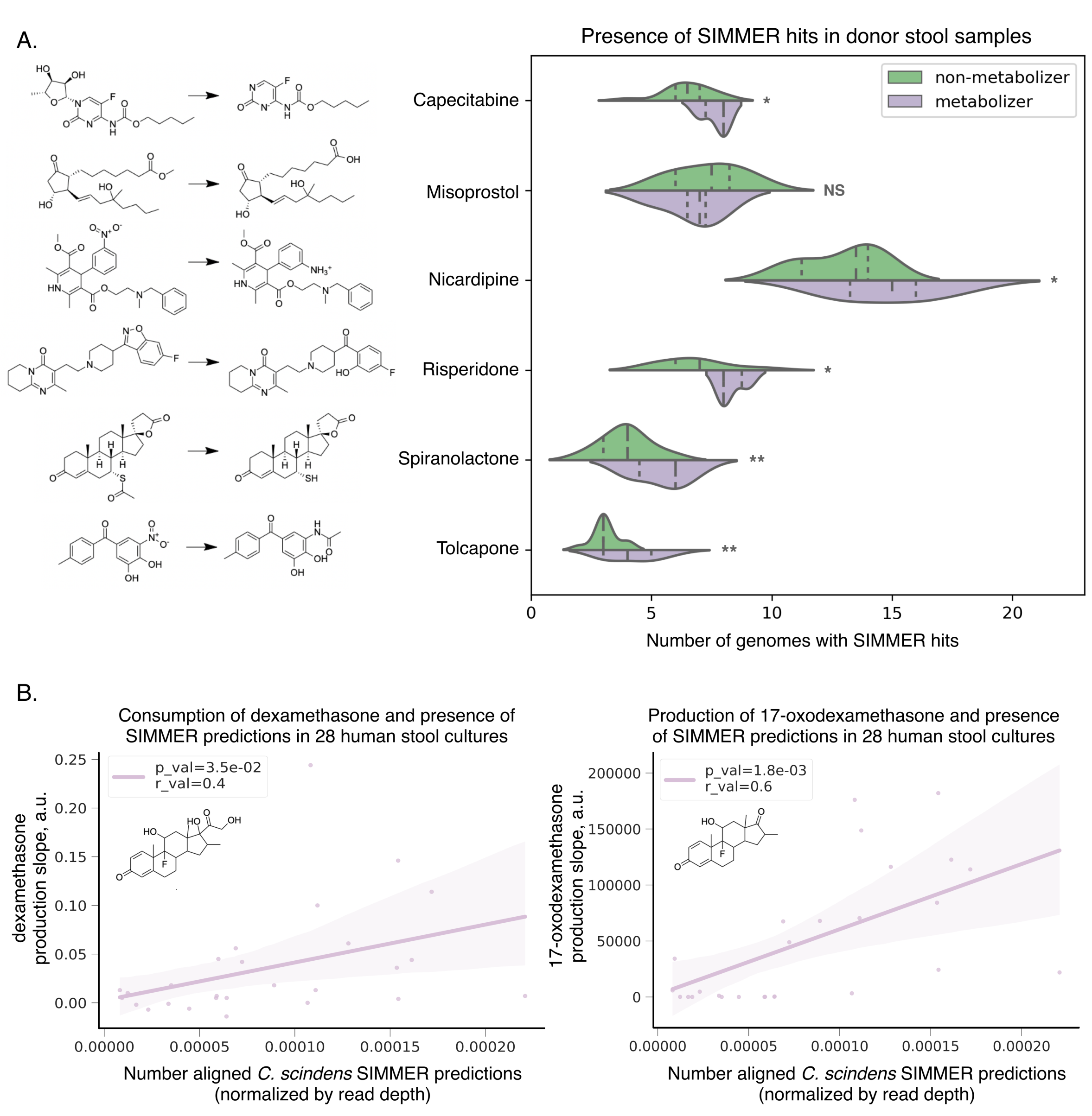
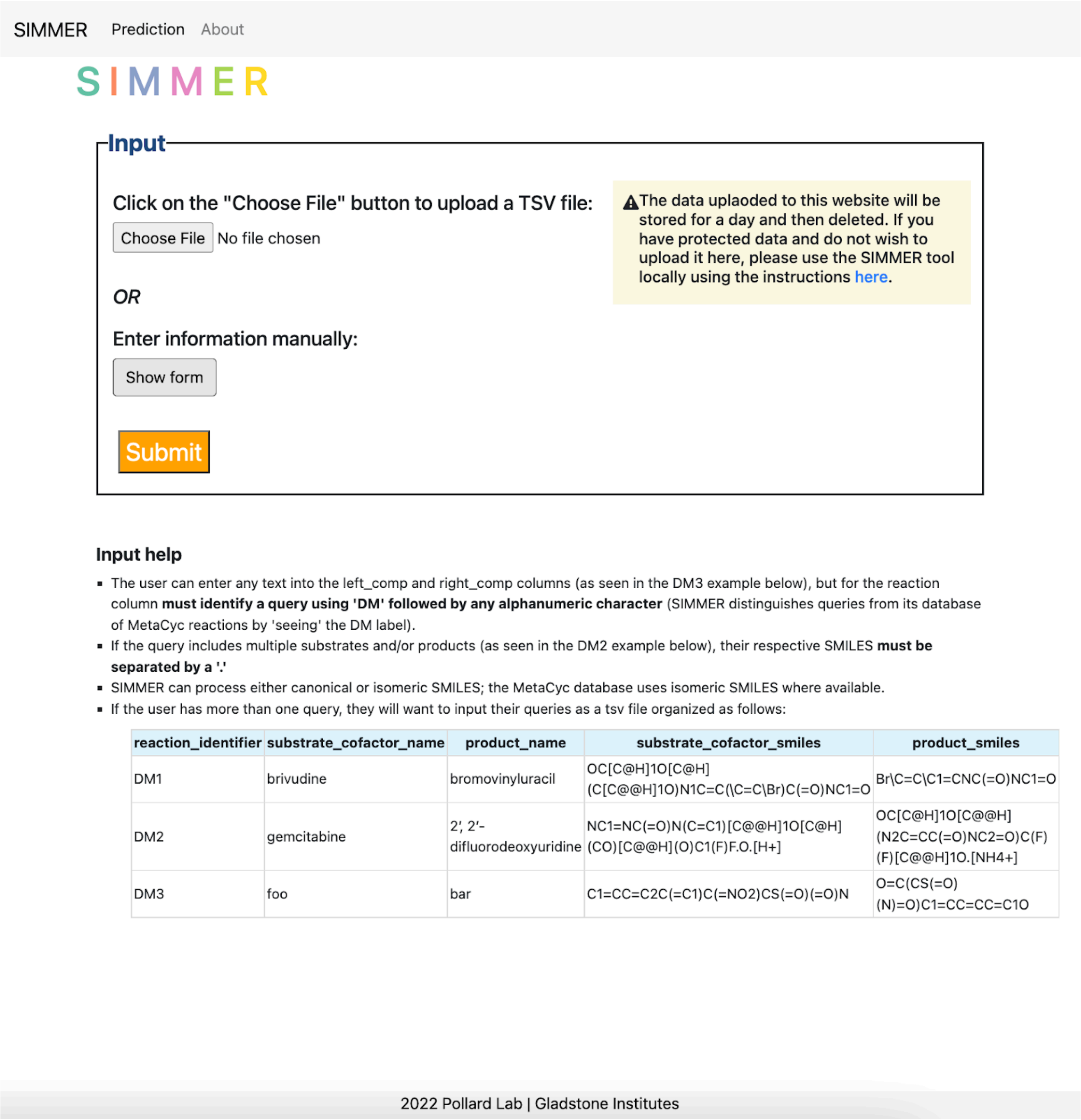


Figure 2.7 SIMMER predicted enzymes explain inter-individual variations in drug metabolism.
(A) Donors (N=20) from the Javdan, et al. 16S rRNA gene sequencing study (Javdan et al., 2020) possessed an enrichment of genomes harboring SIMMER enzyme predictions when metabolism of a given drug was observed. Violin plot curves were made using a seaborn package that performs a kernel density estimation of the underlying datapoint distribution. Chemical

transformations were drawn using ChemDraw software. Single asterisks denote p-values ≤ 0.05, and double denote p-values ≤ 0.01. **(B)** There was a significant correlation between a human stool sample's ability to consume dexamethasone (consumption slope, a.u.), to produce 17-oxodexamethasone (production slope, a.u.), and the number of aligned SIMMER predicted sequences for side chain cleavage of dexamethasone. Patient (N=28) conversion slopes and metagenomics data were accessed from the original study (Zimmermann et al., 2019). Chemical structures were drawn using ChemDraw software.

2.2.6 SIMMER software

In addition to providing SIMMER (https://github.com/aebustion/SIMMER) as a command line tool that quickly generates enzyme sequence predictions (fasta and tab-separated-value files), EC predictions (tab-separated-value file), and MetaCyc reactions ranked by similarity (tab-separated-value file) based on a user's input reaction, SIMMER is also available as a user-friendly website (https://simmer.pollard.gladstone.org/). The user can either input one query reaction at a time, or upload multiple reactions in tsv file format (Figure 2.8). All output types available with the SIMMER command line tool are likewise retrievable via the SIMMER website.



Input

Click on the "Choose File" button to upload a TSV file:

Choose File No file chosen

OR

Enter information manually:

⚠The data uplaoded to this website will be stored for a day and then deleted. If you have protected data and do not wish to upload it here, please use the SIMMER tool locally using the instructions here.

Input help

- The user can enter any text into the left_comp and right_comp columns (as seen in the DM3 example below), but for the reaction column **must identify a query using 'DM' followed by any alphanumeric character** (SIMMER distinguishes queries from its database of MetaCyc reactions by 'seeing' the DM label).
- If the query includes multiple substrates and/or products (as seen in the DM2 example below), their respective SMILES **must be separated by a '.'**
- SIMMER can process either canonical or isomeric SMILES; the MetaCyc database uses isomeric SMILES where available.
- If the user has more than one query, they will want to input their queries as a tsv file organized as follows:

reaction_identifier	substrate_cofactor_name	product_name	substrate_cofactor_smiles	product_smiles
DM1	brivudine	bromovinyluracil	OC[C@H]1O[C@H](C[C@@H]1O)N1C=C(\C=C\Br)C(=O)NC1=O	Br\C=C\C1=CNC(=O)NC1=O
DM2	gemcitabine	2', 2'-difluorodeoxyuridine	NC1=NC(=O)N(C=C1)[C@@H]1O[C@H](CO)[C@@H](O)C1(F)F.O.[H+]	OC[C@H]1O[C@@H](N2C=CC(=O)NC2=O)C(F)(F)[C@@H]1O.[NH4+]
DM3	foo	bar	C1=CC=C2C(=C1)C(=NO2)CS(=O)(=O)N	O=C(CS(=O)(N)=O)C1=CC=CC=C1O

2022 Pollard Lab | Gladstone Institutes

Figure 2.8 SIMMER webtool.
The landing page for the SIMMER website (https://simmer.pollard.gladstone.org/) allows the user to upload a TSV file of queries or add a single query manually to run SIMMER on. It is recommended to use the command-line tool (https://github.com/aebustion/SIMMER) for more than 10 input queries.

2.3 Discussion

In this work, I created a tool that appropriately describes reaction chemistry and harnesses all current information on gut bacterial sequences, both from isolates and metagenome assembled genomes. This advances our ability to discover chemical transformations in the human microbiome, because previous methods for *in silico* metabolism prediction had several key limitations, including low accuracy. Here, I demonstrated SIMMER's ability to recover known drug-metabolizing enzymes in the human gut, to extend previous experimental findings for multiple drug metabolism events by identifying candidate enzymes, and to add clarity to the genetic component of dexamethasone metabolism by *C. scindens*.

To describe chemical reactions, I was initially influenced by recent research that employed substrate and product chemistry to compare bacterial-drug metabolism events to primary reactions in the MetaCyc database, but without the end-goal of EC and enzyme identity prediction (Mallory et al., 2018). From a reaction description standpoint, the published method was still limited in that it only included a description of one substrate and one product per reaction, precluding it from utilizing cofactors and from accurately describing anything other than intramolecular rearrangements (EC class 5, Figure 2.2—figure supplement 3). For this reason, I employed a chemical representation technique that can describe multiple inputs and outputs for a single reaction (Schneider et al., 2015).

To connect these chemical descriptions to bacterial proteins in the human gut, I knew it was important not to rely on EC codes (as previous methods have done) to find relevant sequences. While EC codes are helpful for describing reaction types, from an enzyme perspective they contain no information about substrate specificity for a particular compound. I instead chose to create sequence searches of large genome databases directly from enzymes known to conduct chemically similar reactions, whether or not they have been fully annotated with an EC code. For example, enzyme BT_4096 responsible for diltiazem deacetylation (Zimmermann et al., 2019b) is not yet annotated by the EC, yet SIMMER was able to accurately predict the deacetylase enzyme responsible for diltiazem metabolism because it does not require EC annotation for its enzyme predictions. Indeed, instead of relying on EC codes for sequence searches, I harnessed EC annotations in the MetaCyc database to create a novel EC predictor. Using this tool, I accurately predicted diltiazem deacetylation as an EC3 hydrolysis reaction. While EC prediction methods based on sequence exist, to my knowledge, this is the first instance of an EC prediction method based solely on the chemical description of a reaction.

SIMMER achieved high accuracy when applied to known drug-metabolism events in the gut microbiome. Correct EC designations and enzyme sequences were recovered for 31 drug metabolism events previously characterized in microbiome literature. These reactions span multiple EC classes, and were described by multiple publications, demonstrating the wide application and accuracy of SIMMER. While SIMMER provides high accuracy (i.e. true positive) enzyme predictions for chemical transformations in the human gut, the potential for false positives may be high, as its enzyme lists are not filtered by biologically relevant metrics like substrate affinity or flux consistency in a microbial community. To the former point, users may wish to employ tools like Similarity Ensemble Approach to narrow in on hits most likely to interact with compounds of interest (Keiser et al., 2007). To the latter, a user could choose to further analyze their SIMMER output for flux-balance if the predicted SIMMER bacterial species are described in current metabolic reconstruction models (Heinken et al., 2020; Magnúsdóttir et al., 2017).

Due to its high accuracy predictions for previously described drug-metabolism events, I also used SIMMER to predict novel drug metabolism chemistry in the human gut, and expanded what we know of the bacterial enzymes at play in identified drug transformations by gut bacteria. Recent high-throughput experimental research has greatly increased our knowledge of the number of drugs altered by bacteria in the human gut, but has led to a bottleneck in identifying the responsible bacterial enzymes. While direct experimentation is a necessary component to elucidating the bacterial players responsible, *in silico* methods like SIMMER are needed to help prioritize which of the many bacterial species and enzymes to assess. Here I showed that SIMMER both corroborates previous high throughput experimental data, and also adds increased clarity to the findings. While a previous experimental study was able to elucidate the importance of an isolate *C. scindens* in the metabolism of dexamethasone, the abundance of *C. scindens* in human samples did not correlate with metabolism. When assessed with SIMMER, however, a significant correlation between metabolite production and amount of SIMMER predictions was observed. This finding demonstrates that species identity alone is not enough to explain bacterial chemical transformation, and that responsible genetic elements must be interrogated as well.

Two previous computational tools exist for describing non-antibiotic microbial drug metabolism. MicrobeFDT groups thousands of compounds based on their similarity to one another and annotates compound groups based on any known links to EC numbers, and subsequently, microbes known to contain such EC codes (Guthrie et al., 2019). This network approach was an important addition to the exploration of microbiome metabolism, but its use is limited to a fixed database of chemicals and EC annotations which prevents the user from exploring novel chemistry and also from utilizing hypothetical protein data gathered from metagenomic sequencing studies. Furthermore, MicrobeFDT's accuracy within its database of substrates is limited by its exclusive description of substrates rather than full reactions. DrugBug, a tool that employs Random Forests rather than a network approach, also exhibits limited power and accuracy due to its sole reliance on substrate chemistry and relatively small database of only 491 isolated bacterial genomes from the human gut (Sharma et al., 2017). Of note, my comparison of SIMMER's performance to existing methods necessitated downloading and analyzing my positive control list against the other tools, as none of the previous publications provided any computational validation or accuracy metrics.

One user pitfall of SIMMER in comparison to previous methods, is that a reaction's product(s) and cofactor(s) identity is required to achieve the high accuracy enzyme prediction described here. This is a limitation, as a growing amount of LC-MS/MS data in microbiome research only reports whether or not a compound is depleted in the microbiome and the mass/charge ratio of the product formed, not the product identity. While it is technically possible for the user to submit a SIMMER query that only consists of a substrate, or uses a compound identity as both substrate and product, I do not recommend this due to the previously discussed lack of accuracy when only considering substrates (Figure 2.2—figure supplement 3, Table 2.1—source data). For users wishing to utilize SIMMER with a compound of interest and its either unknown or uncharacterized products, additional tools such as BioTransformer could be used in tandem to create product template predictions before querying (Djoumbou-Feunang et al., 2019). Lastly, if the user does not hold a certain level of knowledge in chemistry, appropriate cofactors (such as water employed in a hydrolysis event) might be omitted from a query, leading to lower accuracy predictions. If a user is unsure which cofactors may be at play in their reaction of interest, reaction rules tools such as RetroRules could be employed (Duigou et al., 2019).

Another limitation of SIMMER is that its underlying protein data is solely metagenomics data from the human gastrointestinal tract, but some compounds, such as the vaginal gel tenofovir, are known to be altered by bacteria in non-GI tract settings (Klatt et al., 2017). That being said, for transformations in the human gut, SIMMER employs the largest available database of relevant bacterial sequences, and the tool could easily be expanded in the future to include other human body sites as well as non-host associated environments. Further related to database constraints, while SIMMER is novel in its ability to query reactions not previously described in chemistry databases, its search space is still limited to reactions that broadly relate to those captured in MetaCyc. As MetaCyc expands, or additional databases get employed, SIMMER will likely be able to make increasingly fine-tuned predictions.

SIMMER enters microbiome biotransformation research at an important point: while there are hundreds of microbiome altered compounds which are in need of enzyme identification, there are also a sufficient number with characterized enzymes to enable us to test the tool’s accuracy. Its ability to predict these known enzymes accurately builds confidence for its predictions of yet unknown enzymes. With this tool in hand, microbiome researchers can make informed hypotheses before embarking on the lengthy laboratory experiments required to characterize novel bacterial enzymes that can alter human ingested compounds. Continued refinement of SIMMER and other computational tools will accelerate microbiome research, providing data-driven hypotheses for experimental testing and a first step towards understanding the full scope of metabolism by the human microbiome.

2.4 Materials and methods

Preparation of SIMMER's underlying chemical data

13,387 gene annotated bacterial reactions were downloaded from MetaCyc (Caspi et al., 2008). Each reaction from the database contained Simplified Molecular Input Line Entry System strings (SMILES) of reactant(s) and product(s), EC code if available, and UniProt or Entrez identifiers for sequences that catalyze the reaction. All MetaCyc compounds were then protonated based on the pH environment of 7.4 in the human small intestine, where most oral drug absorption occurs. Protonation states were calculated using ChemAxon's cxcalc majorms software ("cxcalc calculator functions," n.d.).

RDKit's rdChemReactions module was employed to create chemical fingerprints representing each MetaCyc reaction. Chemical reaction objects were constructed from reaction SMILES arbitrary target specification (SMARTS) strings. Fingerprints for these reactions were then created using the resulting difference of product(s) and reactant(s) Atom-Pair fingerprints (Schneider et al., 2015). SIMMER users can also opt to use Topological Torsion, Pattern, or RDKit fingerprints, but unless otherwise stated, all analyses in this manuscript use Atom-Pair difference fingerprints. Of the 13,387 MetaCyc reactions, 8,914 were able to be fingerprinted using this method. Failed fingerprints were due to ambiguous SMILES identifiers or presence of non-small molecule compounds in a reaction, such as peptides.

After creating fingerprint vectors for all MetaCyc reactions, an 8,914 by 8,914 pairwise similarity matrix of Tanimoto coefficients was created. These Tanimoto vectors make up SIMMER's underlying chemical data.

Preparation of SIMMER's underlying protein data

For each of the 8,914 fingerprinted MetaCyc reactions, all relevant gene sequences were retrieved from the MetaCyc reaction's UniProt and Entrez database linkouts. If at least two genes, with a median pairwise sequence similarity greater than or equal to 27%, were linked to a given MetaCyc reaction, the sequences were used to create a multiple sequence alignment and subsequent profile hidden Markov model (pHMM) using Clustal Omega and HMMER3 (version 3.2.1) software respectively (Eddy, 2009; Sievers and Higgins, 2014). This similarity cutoff was chosen based on previous protein family literature (Mi et al., 2021). If fewer than two genes, or genes with less than 27% global similarity, were associated with a given MetaCyc reaction, a pHMM of the MetaCyc gene(s) PANTHER subfamily was retrieved via InterPro linkouts (Mi et al., 2021). MetaCyc derived and PANTHER subfamily pHMMS were then queried against a Unified Human Gastrointestinal Genome (UHGG) collection of 286,997 isolate genomes and metagenome assembled genomes from the human gut environment using the HMMER3 hmmsearch module (Almeida et al., 2020; Eddy, 2009). In the case of MetaCyc reactions with too few sequences, too low a median pairwise sequence identity, *and* a missing PANTHER database subfamily pHMM, single sequence protein queries were conducted against the UHGG databse using HMMER3's phmmer module, which internally created protein profiles for the single query sequences based on a position-independent scoring system. Resulting enzyme hit lists were filtered to only include high significance hits (e-value < 1E−5, and hit length >= half of the input pHMM alignment or single sequence length). In sum, for each MetaCyc reaction, a profile representing the diversity of the enzyme family for that chemical transformation was used to find sequence similar hits in the human gut microbiome that can mediate chemically similar reactions.

Each human gut microbiome hit was further described by the identity, prevalence, and abundance of the bacterial strain in which it resides. To establish prevalence and abundance of UHGG strains, metagenomic analysis was performed on the Predict (Personalised Responses to Dietary Composition Trial) cohort due to its high number of samples and favorable sequencing depth (Asnicar et al., 2021). Shotgun metagenomic reads were analyzed with MIDAS2 an implementation of Metagenomic Intra-Species Diversity Analysis Subcommands (MIDAS) (Nayfach et al., 2016; Zhao et al., 2022) designed for the UHGG database. Presence of a SIMMER predicted species in a given sample was established when reads mapped (HS-BLASTN) to 15 single-copy universal genes for that species (Chen et al., 2015), with at least 75% alignment coverage. To assess the gene content of a sample, shotgun metagenomic reads were aligned to a MIDAS2 created pangenome of the SIMMER species' genes clustered at 99% nucleotide identity. Copy number of a SIMMER gene prediction was established by dividing aligned prediction reads by the full length of the prediction. This number was then normalized by the read coverage of 15 single-copy universal genes in the same sample to estimate copy number per cell. Presence of a SIMMER enzyme was established if at least 0.35 gene copies per cell were present in a sample.

Phylogenetic trees were also constructed for each hmmsearch and phmmer result. For each set of MetaCyc reaction human gut microbiome enzyme hits, CD-HIT was used to cluster results at 95% identity (Fu et al., 2012). Then MUSCLE was used to create a multiple sequence alignment for input to FastTree (Edgar, 2004; Price et al., 2009). Compact tree visualizations were made in R using ggtree and ggtreeExtra (Xu et al., 2021; Yu et al., 2017). All tree tips were colored by phylum, and surrounded by circle annotators describing a given hit's Prokka predicted function, genome type (i.e. from an isolate or metagenome assembled genome), and prevalence/abundance in the Predict cohort (Seemann, 2014).

Query functionality of SIMMER

The query functionality of SIMMER was designed similarly to the precomputed chemistry data. After receiving an input chemical transformation (or tsv describing multiple input reactions) in the form of SMILES, SIMMER fingerprints the reaction(s) and compares it to the precomputed chemical space by computing the Tanimoto coefficients between the input(s) and all precomputed reactions. The 8,914 precomputed MetaCyc reaction Tanimoto vectors are then sorted by ascending euclidean distance to the query Tanimoto vector. SIMMER by default ranks reactions' euclidean distances based directly on the Tanimoto vectors, but if a user's inputs require a decrease in computational burden, PCA can be employed after similarity matrix creation and before euclidean distance rankings. The number of PCs to be used depends on the fingerprint style employed, and was determined by the Kaiser criterion. Unless otherwise stated, all analyses in this manuscript employed the full Tanimoto similarity matrix with no PCA reduction. Human gut microbiome enzymes that may conduct the input reaction are reported from the precomputed UHGG hmmsearch or phmmer results of the closest euclidean distance MetaCyc reaction. Significantly enriched EC identities (i.e. reaction types) are also reported.

Reaction type predictions

SIMMER predicts an EC code (i.e. reaction type) for a query reaction if there is an enrichment of a particular EC at the top of the reaction list. Enrichment was determined in a manner similar to gene set enrichment analysis (GSEA).(Subramanian et al., 2005) For each EC code associated with MetaCyc reactions, an enrichment score (ES) was calculated by walking down the ranked list of reactions. Starting with a score of zero, each time the given EC is encountered the score increases by one, and each time a different EC is encountered the score decreases by one. At the end of this process, each EC receives an ES that is the score's maximum distance from zero after walking through the list (Figure 2.2—figure supplement 1A). Because the MetaCyc database of reactions is unbalanced in its EC code representation, ES scores for a given EC type are divided by the number of times the EC in question occurs in the database. This yields a normalized ES (NES) for SIMMER reporting. Significance is established by comparing the true NES to the NES achieved from 1000 permutations of a shuffled reaction list (Figure 2.2—figure supplements 1B-C). When multiple EC codes are predicted as significant, they are ranked in ascending order of where in the list of 8,914 reactions the NES occurs. This method was verified by subsampling the database of MetaCyc reactions to equal numbers (N=96) of reactions for each EC class, the broadest resolution level of an EC code. Each of these subsampled reactions was then queried with SIMMER against the entire MetaCyc reaction database to create sorted reaction lists for each query. SIMMER predicted an EC code(s) for each reaction based on the most highly enriched EC. SIMMER's recall, precision, and accuracy are high for EC class, sub-class, and sub-sub-class level resolution (Figure 2.2B, Figure 2.2—source data). For the serial designation of an EC code (the most granular description of an EC code), however, SIMMER's performance diminished, potentially because enrichment calculations suffer from increased uniqueness in the ranked list and therefore reduced power to determine a match (Figure 2.2—source data). This indeed appears to be the case; when the database is subsampled to ensure at least three of each unique serial designation, F1-scores (the harmonic mean of precision and recall) and accuracy remain high despite the increased EC resolution (Figure 2.2—source data, Figure 2.2—figure supplement 1D).

Euclidean distance silhouette scores

To analyze SIMMER's resilience to different reaction chemistry representations, I created a silhouette-like euclidean distance score. For the precomputed MetaCyc chemical dataset of 8,914 reactions (i.e. the Tanimoto pairwise similarity matrix), I split all reactions into their top-level EC codes (i.e. EC class) and determined for each reaction its euclidean distance to all reactions within its EC class versus outside its EC class. From the two distributions (within EC and without EC distances) created, I computed a Kolmogorov-Statistic to determine if the distributions significantly ($p<0.05$) differed. I repeated this process for finer resolution EC classifications (sub-class, sub-sub-class, and serial designation). Euclidean distance silhouette scores were used to compare different chemical representations, such as fingerprint style, inclusion of products, and inclusion of cofactors.

Relationship between SIMMER's underlying chemical and protein data

For MetaCyc enzymes (N=34,279) associated with multiple reactions, one reaction was used as a SIMMER query, and the other reaction(s) searched for in the ordered reaction list output. As a negative control, these reaction similarity results were then compared to all pairwise combinations of MetaCyc enzymes (subsampled to N=34,279) that do not conduct the same reaction.

I also assessed the relationship between chemistry and protein similarity for all pairwise combinations of a subset of MetaCyc reactions annotated with only one protein sequence (N=604 reactions). Chemical similarity was based on the Euclidean distance between two reaction fingerprint vectors in SIMMER's precomputed chemical space (Figure 2.1A). Global protein similarity was determined via the Needleman–Wunsch algorithm. The relationship between chemical similarity and protein similarity was assessed with a Pearson's correlation coefficient and *P* value calculated using a Wald Test with t-distribution of the test statistic.

Creating a compendium of drug-metabolism use cases from the human gut

To analyze SIMMER under the use-case of drug metabolism, I created a compendium of drug degradations that occur in the human gut microbiome. The compendium of reactions is based on a literature curation of hundreds of papers, and is organized by reactions producing known/unknown metabolites and driven by known/unknown bacterial enzymes. The drug-metabolism positive controls used to assess SIMMER's accuracy were drawn from the list of reactions possessing a structurally elucidated metabolite and driven by a characterized bacterial enzyme.

I further expanded the positive control list to include sequence-similar enzymes that likely perform the same function. For this expansion, I performed pHMM searches (when a positive control reaction had been characterized with multiple sequence-similar enzymes) and phmmer searches (when a positive control reaction had been characterized with only one sequence) of the UHGG database using HMMER3 software (Almeida et al., 2020; Eddy, 2009). High significance (e-value < 1E−5) hits were kept when the resulting alignment was at least 50% of the input pHMM or sequence length. This list of significant hits was filtered by presence in human ileum or jejunum (the site of human drug absorption) via DIAMOND searches against metagenomic reads from a published study that employed jejunum and ileum endoscopy (Buchfink et al., 2021; Zmora et al., 2018). The hits were also filtered for presence in RNAsequencing data via DIAMOND searches of rnaSPAdes assembled reads from HMP2 metatranscriptomics control patient samples (Bushmanova et al., 2019; Integrative HMP (iHMP) Research Network Consortium, 2019). The hits were lastly filtered by predicted affinity for their substrates using the Similarity Ensemble Approach (Keiser et al., 2007).

Corroboration of previous high-throughput experimental findings

For the first experimental validation of SIMMER, I analyzed results from sequencing studies (NCBI BioProject: PRJNA593062) described in a previously published high-throughput investigation of bacterial drug metabolism in human stool samples (Javdan et al., 2020). The first sequencing set in this publication was a deep metagenomic sequencing of one pilot individual's *ex vivo* stool originally evaluated for its ability to degrade hundreds of therapeutics. I used MetaSPAdes with default settings to assemble the metagenomics reads into scaffolds (Nurk et al., 2017). I then queried SIMMER with eight reactions that were structurally elucidated (via nuclear magnetic resonance) by the previous publication, and ensured via TBLASTN searches that SIMMER predicted hits were found in the assembled metagenomic reads. The second sequencing set was a 16S rRNA sequencing experiment of twenty human donor stool samples originally evaluated for their inter-individual variation in bacterial drug degradation. I queried SIMMER with five of these reactions possessing structurally elucidated metabolites, and evaluated enrichment of SIMMER predicted bacterial species in metabolizing versus non-metabolizing donors. Species matches between SIMMER predictions and the 16S study were made using the SequenceMatcher class from the difflib python module set to an 80% ratio cutoff. Enrichment of SIMMER predicted bacterial genomes was then assessed by computing a t-test for number of SIMMER genomes in metabolizers versus number of SIMMER genomes in non-metabolizers for a given reaction.

For experimental corroboration of dexamethasone metabolism, I accessed shotgun sequencing data (PRJEB31790) from a cohort of 28 human stool samples shown to metabolize dexamethasone to varying degrees (Zimmermann et al., 2019b). Shotgun reads were assembled using MetaSpades with default settings. Presence of SIMMER enzyme predictions was established via search with DIAMOND and normalized by sample read depth. Significance was established with a Pearson's correlation coefficient and *P* value calculated using a student's *t*-distribution.

Web tool creation

We used the python web framework Flask (https://flask.palletsprojects.com/en/2.1.x/) to make SIMMER available as a user-friendly website. The website accepts either a single query reaction or multiple query reactions via a file upload and provides the same outputs as the SIMMER command-line tool. The website also allows the user to download the outputs of interest. Keeping in mind the privacy and security of the data that a user might upload to the website, the website is designed to delete all uploaded data within 24 hours from the server. This will ensure security of the uploaded data.

Data availability

Data generated and analyzed during this study are provided in Figures 2.2-2.7 source data files, Table 2.1 source data file, supplemental files, and at https://github.com/aebustion/SIMMER. Accession numbers of previously published datasets are provided in the Materials and Methods section. SIMMER code can either be run at the SIMMER website (https://simmer.pollard.gladstone.org/) or downloaded directly from the above-linked GitHub.

2.5 Supplemental data

Supplementary File 1 (attached). Literature curated list of drug-metabolism events in the human gut microbiome.

Table 2.1 source data (attached)

Figures 2.2, 2.4–2.7 source data (attached).

2.6 Supplemental figures

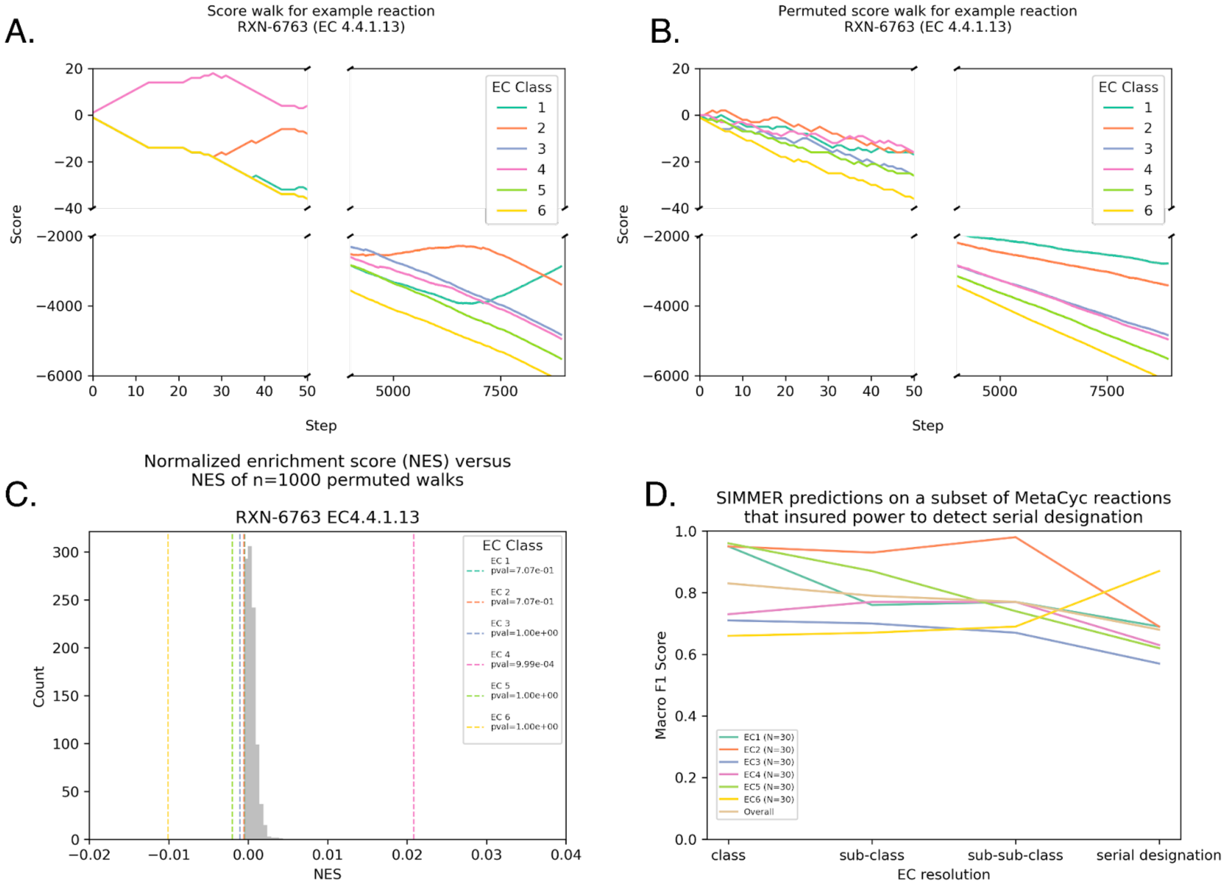


Figure 2.2—figure supplement 1. SIMMER predicts an EC code (i.e. reaction type) for a query reaction if there is an enrichment of a particular EC at the top of the reaction list. (A) For each EC code associated with MetaCyc reactions, an enrichment score (ES) was calculated by walking down the ranked list of reactions. Starting with a score of zero, each time the given EC is encountered the score increases by one, and each time a different EC is encountered the score decreases by one. Panel A is an example of such a walk with MetaCyc reaction RXN-6763. **(B)** Significance is established by comparing the true ES to the ES achieved from 1000 permutations of a shuffled reaction list. Panel B is an example of how RXN-6763 loses its EC enrichment structure after shuffling. **(C)** Because the MetaCyc database of reactions is unbalanced in its EC code representation, ES scores for a given EC type are divided by the number of times the EC in question occurs in the database. This yields a normalized ES (NES) for SIMMER reporting. This is also performed for the shuffled distributions. **(D)** For the serial designation of an EC code (the most granular description of an EC code), SIMMER's performance diminished (Figure 2.2B), because enrichment calculations suffer from increased uniqueness in the ranked list and therefore reduced power to determine a match. This is proven here, as when the database is subsampled to ensure at least three of each unique serial designation, F1-scores (the harmonic mean of precision and recall) and accuracy remain high despite the increased EC resolution.

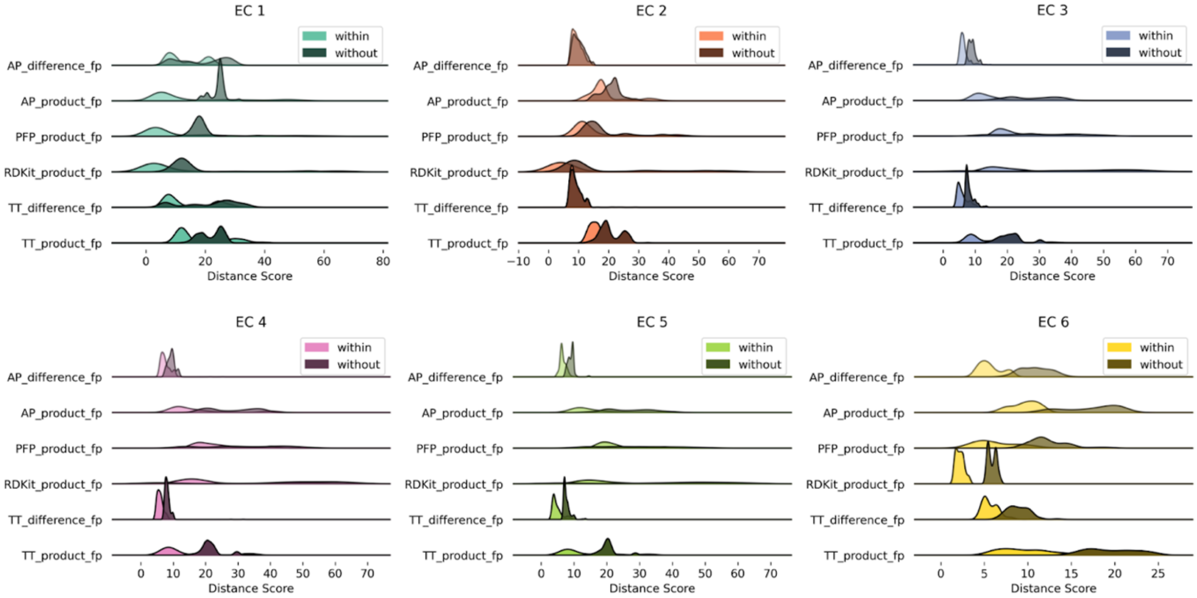


Figure 2.2—figure supplement 2. Euclidean distance distributions and silhouette scores for top-level EC codes are resilient to fingerprint type. Distributions are computed as described in Figure 2A.

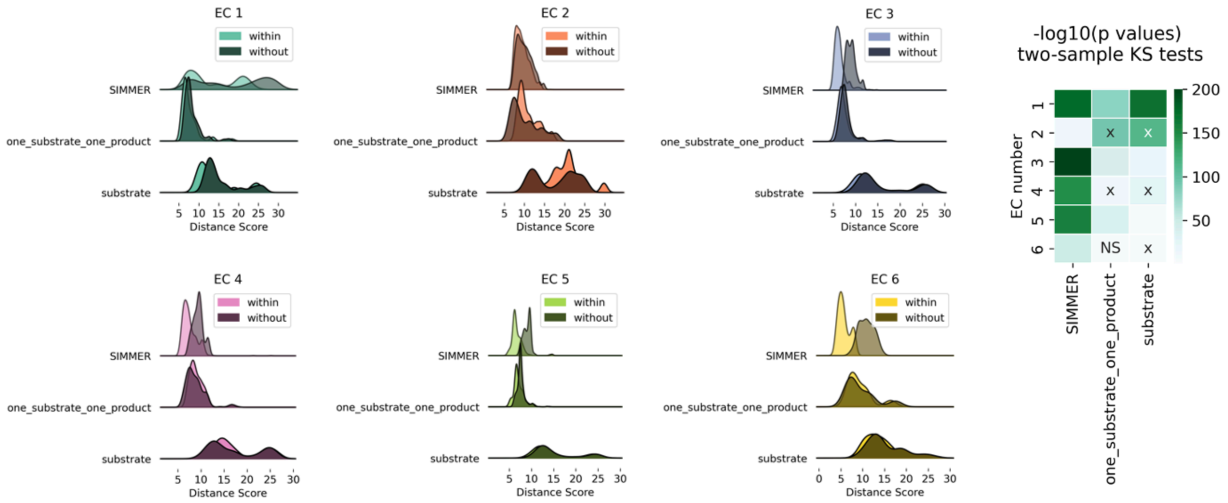


Figure 2.2—figure supplement 3. Euclidean distance distributions and silhouette scores for top-level EC codes are sensitive to chemical representation type. Distributions are computed as described in Figure 2.2A. The resulting score distributions were used to compare SIMMER's resilience to different chemical representations, such as removal of products and cofactors (the representation methods employed by DrugBug (Sharma et al., 2017) and MicrobeFDT (Guthrie et al., 2019) or the inclusion of exactly one substrate and exactly one product (the Mallory, et al. method (Mallory et al., 2018). For all EC classes, scores were smaller within versus between EC classes using SIMMER's chemical representation, indicating that SIMMER can detect reaction similarity within EC classes. However, this was not consistently true when using only the input compound without cofactors or products (substrate), or when using only one substrate and product (one_substrate_one_product). From the pairs of distributions, I computed a Kolmogorov-Statistic to determine if the distributions significantly ($p<0.05$) differed. This showed that the differences were statistically significant for SIMMER chemical representations and reduced, or in the wrong direction, when using a reduced representation. X's in the KS test heatmap indicate an incorrect direction of difference (without grouping closer than within).

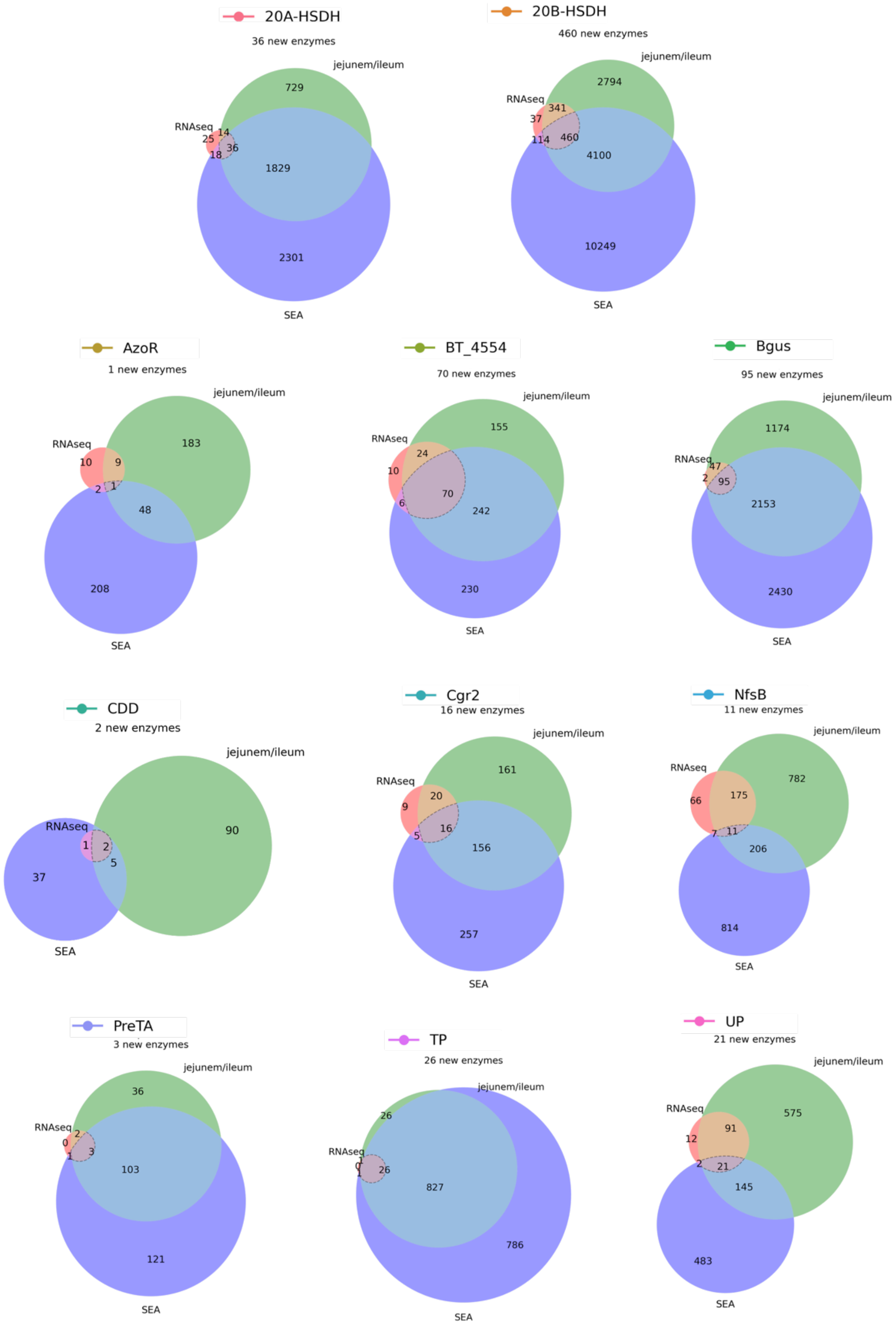


Figure 2.4—figure supplement 1. An expanded list of gut bacterial enzymes relevant to known cases of drug metabolism. 52,849 total candidate homologs (a median of 1,087 candidates per enzyme) were obtained via pHMM searches with sequences from the literature. These putative homologs were filtered for presence in the human jejunum/ileum, presence in

RNA-sequencing studies, and predicted affinity for the substrate in question using SEA. This resulted in 741 additional enzyme sequences for 31 positive control reactions.

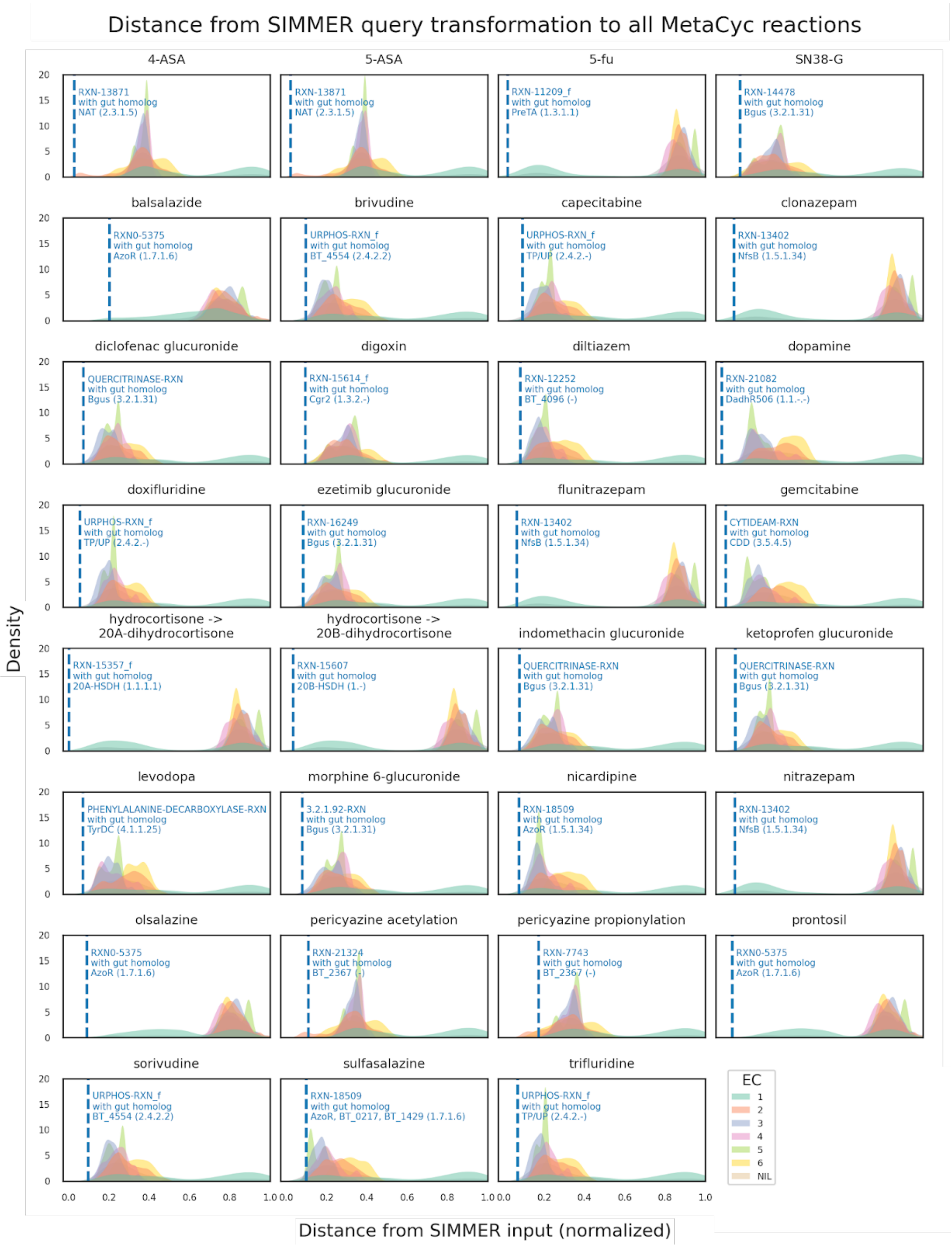


Figure 2.5—figure supplement 1. Distributions of all MetaCyc reactions' euclidean distances to the positive control list queries. For this analysis, each of the 31 positive control reactions was queried to SIMMER. Distributions were created based on all (N=8,914) MetaCyc reactions' distance to a given input metabolism event, and colored by their EC class annotation in MetaCyc (oxidoreductases, transferases, etc.). The dashed blue line depicts the MetaCyc

reaction that yielded a gut microbiome homolog matching the known positive control sequence. For example, when queried with 5-ASA acetylation, SIMMER outputs a most-similar MetaCyc reaction (RXN-13871) that SIMMER linked (via sequence similarity) to an N-acetyltransferase known from previous literature to drive 5-ASA metabolism in the human gut.

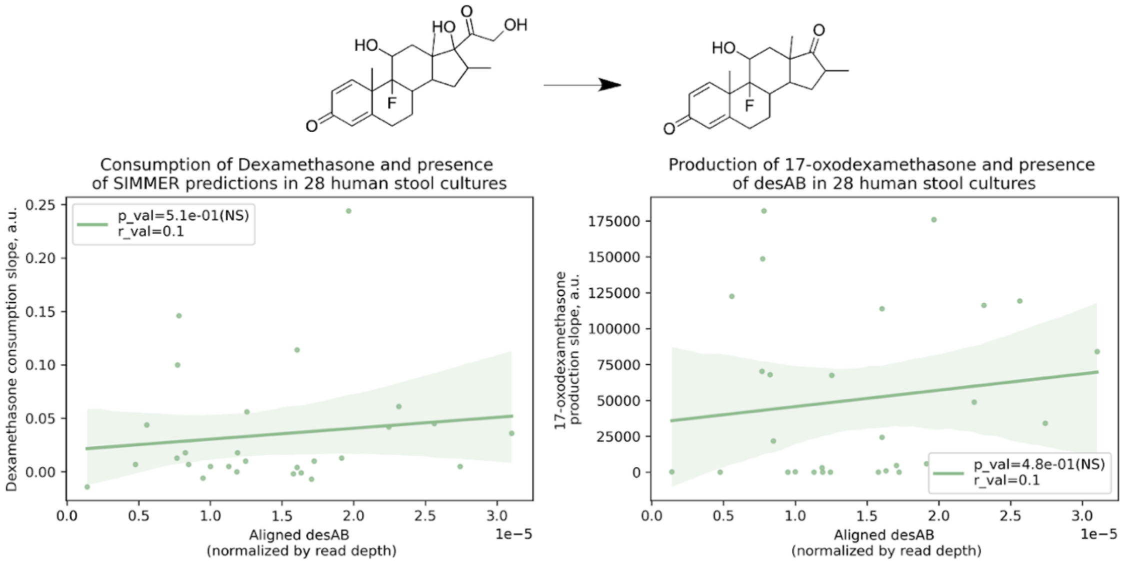


Figure 2.7—figure supplement 1. There was not a significant correlation between a human stool sample's ability to consume dexamethasone (consumption slope, a.u.) or to produce 17-oxodexamethasone (production slope, a.u.), and the number of aligned *desAB* sequences. Patient (N=28) conversion slopes and metagenomics data were accessed from the original study (Zimmermann et al., 2019). This adds confidence to the finding described in Figure 2.7B, as it means SIMMER predictions were not correlated with dexamethasone metabolism due to co-occurrence in *C. scindens* with a gene previously reported to underlie bacterial side-chain cleavage of steroids (Ly et al., 2020).

Chapter 3

Experimental validation of computationally predicted bacterial species and enzymes capable of methotrexate metabolism.

3.1 Introduction

Oral methotrexate (MTX) is an anti-folate immunosuppressant, and the first-line therapy for individuals with Rheumatoid Arthritis (RA). Despite this designation, over half of the RA population receiving MTX exhibit suboptimal improvement (Emery et al., 2008; Scher et al., 2020). This lack of efficacy may be due to altered pharmacokinetics, as oral bioavailability of MTX is subject to large interindividual variation with values ranging from 32-70% (Roon and Laar, 2006). Additionally, MTX is estimated to practice extensive enterohepatic circulation which results in increased exposure to gut bacteria in the small intestine (Roon and Laar, 2006). In line with this assumption, MTX degradation to inactive metabolites 2,4-diamino-N^{10}-methylpteroic acid (DAMPA) and glutamate has been observed in MTX dosed mice, but not in antibiotics pretreated mice (Valerino et al., 1972).

In an early effort to characterize MTX host metabolites, an enrichment culture experiment isolated a soil *Pseudomonas* capable of rapid hydrolysis of MTX into DAMPA and glutamate (Levy and Goldman, 1967). Later purification and characterization of soil-based Carboxypeptidase G1 from *Pseudomonas stutzeri* and Carboxypeptidase G2 (CPG2) from *Pseudomonas* strain RS-16 resulted in the emergency therapy Glucardipase (recombinant CPG2) for patients experiencing MTX toxicity (Buchen et al., 2005; Chabner et al., 1972). Glucardipase's efficient hydrolysis of MTX and knowledge of microbiome MTX degradation suggest that CPG2 homologs in the human gut microbiome may exist, but no enzymes from gut bacteria have yet been characterized (Letertre et al., 2020; Scher et al., 2020).

Increasingly, scientists use metagenomic sequencing to probe associations between the microbiome and RA and/or MTX (Artacho et al., 2020; Kishikawa et al., 2020; Zhang et al., 2015). One such study was able to utilize metagenomics and clinical efficacy data to create a computational predictor of therapeutic outcome before beginning a patient on MTX (Artacho et al., 2020). When comparing the metagenomic data between MTX-responders versus MTX-non-responders, the researchers found consistent differences in the metabolic features of patient metagenomes. Additionally, *ex vivo* stool samples from MTX-non-responders degraded MTX to a higher degree, though the relevant enzymes were not determined (Artacho et al., 2020).

While enzymatic degradation is unlikely to be the only determinant of patient response to MTX, characterization of bacteria capable of MTX hydrolysis will improve RA precision medicine efforts. Using my *in silico* enzyme prediction method, SIMMER (described in Chapter 2), I retrieved bacterial sequences potentially responsible for MTX degradation in the human gut (Bustion et al., 2022). I demonstrated that these enzymes are enriched in MTX-non-responders and in isolates capable of degrading MTX. We also assessed metabolism of MTX in a strain-collection and in a heterologous expression system, with both whole cells and purified proteins tested for activity. This work results in the first species-level characterization of MTX metabolism by bacteria in the human gut microbiome.

3.2 Results

3.2.1 SIMMER predicts methotrexate metabolizing enzymes similar to known environmental methotrexate metabolizers.

When queried with MTX and its gut bacteria associated metabolites DAMPA and glutamate, SIMMER calculated a most similar MetaCyc reaction (3.4.17.11-RXN) and a significant EC prediction (3.4.17.11, p-value<0.001). MetaCyc reaction 3.4.17.11-RXN describes the hydrolysis of folate into pteroate and glutamate, driven by a glutamate carboxypeptidase (Cpg2) found in environmental *Pseudomonas aeruginosa*. Hydrolysis of MTX is chemically similar to hydrolysis of folate (Figure 3.1) with a Tanimoto coefficient=0.6, and normalized euclidean distance=0.05 in SIMMER's precomputed chemical space. SIMMER predicted 2,286 human gut microbiome enzyme predictions that are most frequently Prokka annotated as Carboxypeptidase G2s (Figure 3.2) due to their sequence similarity to MetaCyc's environmental Cpg2. As a result of this similarity, SIMMER's predictions may conduct hydrolysis of methotrexate, chemically similar to *P. aeruginosa's* Cpg2 reaction.

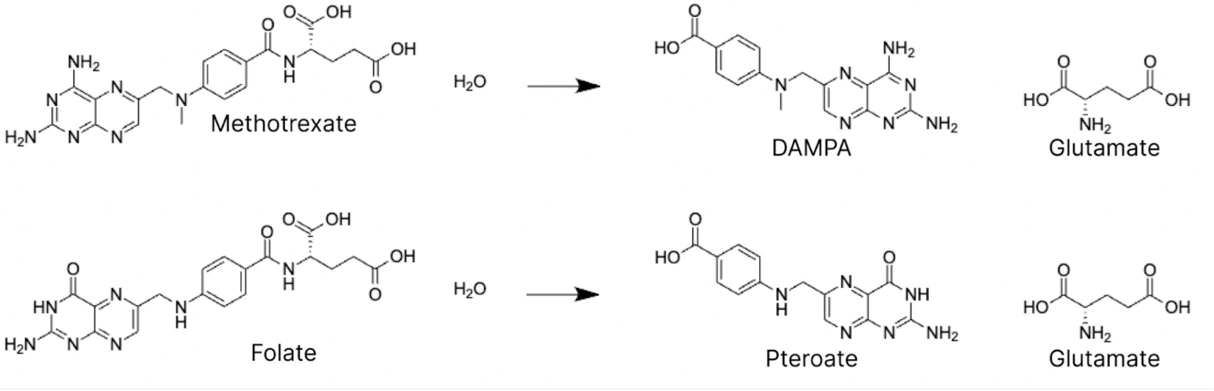


Figure 3.1 Methotrexate hydrolysis and its most similar MetaCyc reaction.
When queried with MTX hydrolysis to DAMPA and glutamate, SIMMER found that folate hydrolysis was the most chemically similar MetaCyc reaction.

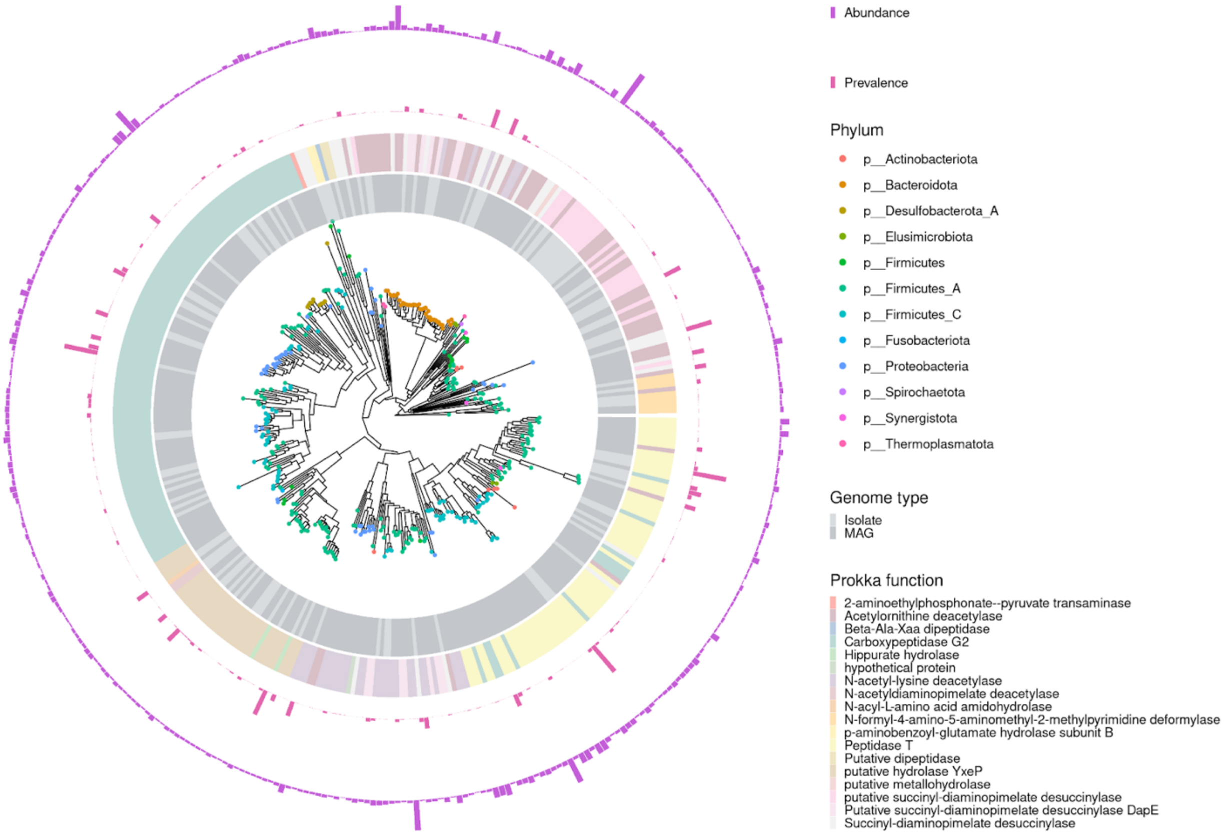


Figure 3.2 Human gut microbiome bacterial sequence predictions.
SIMMER predicted 2,286 unique bacterial sequences putatively capable of MTX hydrolysis to DAMPA and glutamate. There was great variability in the prevalence and abundance of these sequences in healthy human metagenomic data. Among the predictions, Firmicutes members were most common. The most frequent Prokka annotation was Carboxy peptidase G2.

3.2.2. SIMMER MTX predicted enzymes are enriched in MTX non-responders.

Next, I investigated the presence of the above SIMMER predicted enzymes in metagenomics studies of RA patients on MTX therapy (Artacho et al., 2020; Zhang et al., 2015). Between the two studies, I was able to confirm via DIAMOND blastx searches that N=479 unique SIMMER sequence predictions were found across the patient stool samples. Both studies provided patient metadata such as disease severity (DAS28 scores) or whether individual patients responded to MTX (Table 3.1).

Table 3.1 Rheumatoid arthritis metagenomics studies assessed.

Study	Patient population	Sample size	Medications studied	Measurements reported
Zhang et al. 2015. PMID26214836	RA	N=25	MTX, MTX+HCQ+ etanercept, MTX+T2	Disease activity score (DAS28)
Artacho et al. 2020. PMID33314800	New-onset RA	N=47	MTX	Disease activity score (DAS28) delta after 4 months MTX therapy

For the Zhang et al. cohort therefore, I was able to assess whether or not patients with low, moderate, or high disease differed in their abundance of SIMMER enzyme predictions present (Figure 3.3A-B). Of SIMMER's 2,286 enzyme predictions for MTX hydrolysis, N=289 were found in the Zhang et al. metagenomics data. While there was a slight association between presence of SIMMER predictions and disease severity, no association was statistically significant (Figure 3.3A).

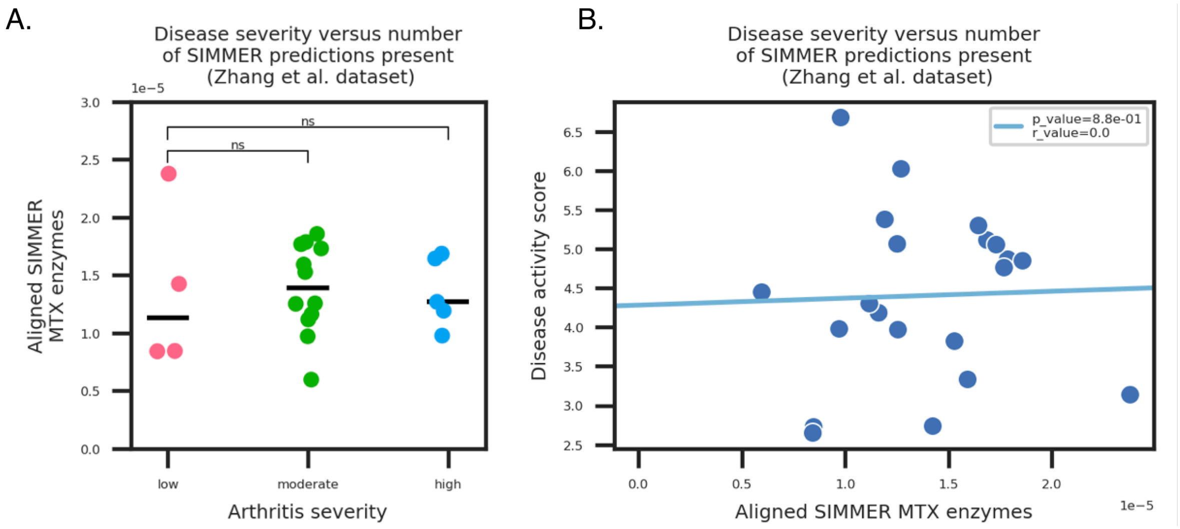


Figure 3.3 SIMMER predictions are found in the Zhang et al. cohort of arthritis patients. N=289 of SIMMER predicted sequences for MTX hydrolysis are found in a cohort of RA patients of varying disease severity. **(A)** No association was found between number of SIMMER predictions and a patient's disease grouping (GLM Poisson), **(B)** or between number of SIMMER predictions and a patient's disease activity score (Pearson correlation, student's t-test).

Next, due to the study design of the Artacho et al. cohort of new-onset RA (NORA) patients, I was able to assess whether or not a relationship existed between abundance of SIMMER predictions within a patient stool sample and that patient's resistance or response to MTX treatment. Response in the study was defined as a DAS28 score improvement by at least 1.8 points within four months of MTX therapy (Artacho et al., 2020). Of SIMMER's 2,286 enzyme predictions for MTX hydrolysis, N=386 were found in the Artacho et al. metagenomics data. A significant negative correlation was seen between patients' disease score improvements and abundance of SIMMER predictions (Figure 3.4B). Similarly, MTX non-responders (defined above) exhibited a significant enrichment of SIMMER MTX predictions in their stool samples (Figure 3.4A). While unlikely to be the complete story of MTX non-response in patients, this analysis points to MTX degradation playing a role in the lack of therapeutic efficacy in these patients.

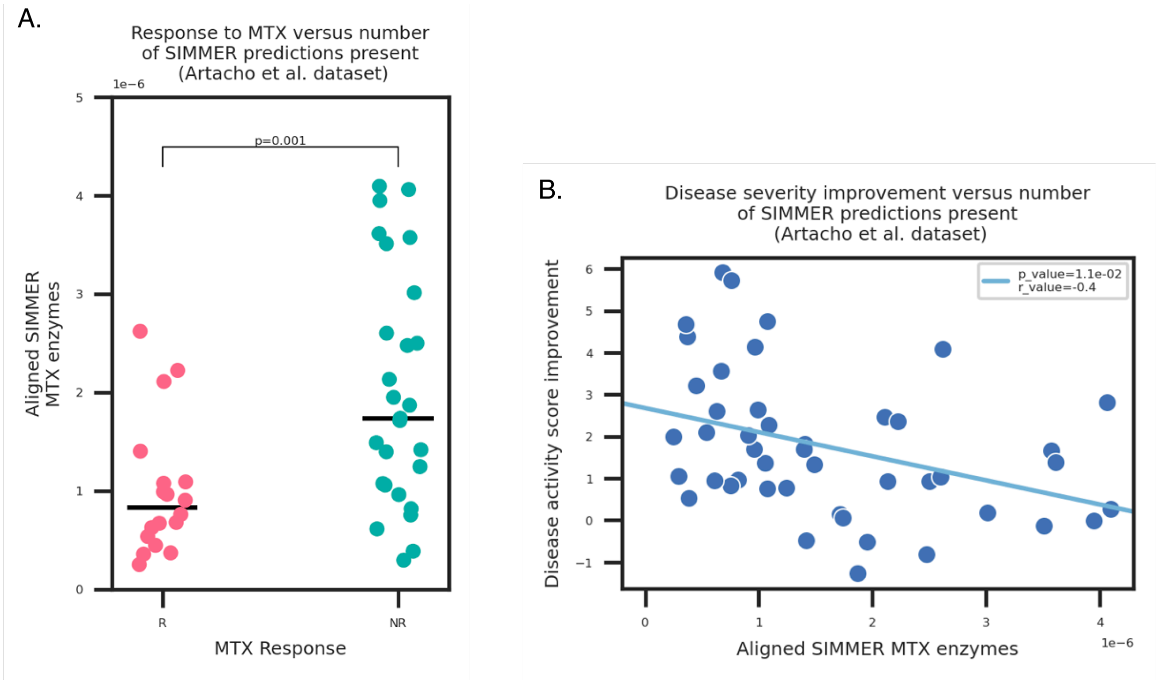


Figure 3.4 SIMMER predictions are enriched in RA patients unresponsive to MTX. N=368 of SIMMER predicted sequences for MTX hydrolysis are found in the Artacho et al. cohort of new-onset RA patients with variable MTX response. **(A)** SIMMER predictions were enriched in patients non-responsive to MTX (GLM Poisson, p-value=0.001), **(B)** and a significant negative association between disease severity improvement and number of SIMMER predictions was observed (Pearson correlation=-0.4, student's t-test p-value=0.01).

3.2.3 Experimental validation of SIMMER predicted species capable of MTX metabolism

We next screened an existing collection of 42 diverse bacterial strains found in the human gut for its ability to degrade MTX to metabolites DAMPA and glutamate. This collection was of interest due to previously determined inter-strain variation in growth inhibition by MTX (Nayak et al., 2021). Each isolate was incubated with MTX (100 µg/mL), and degradation of MTX determined by high-performance liquid chromatography (HPLC) (Figure 3.3). Metabolism varied across the strain collection, with ten isolates capable of metabolism. SIMMER significantly predicted which strains in the library were capable of MTX metabolism (Fishers Exact Test, OR=5.4, p-value<0.05, Figure 3.3). Of note, SIMMER predicted seven isolates in the collection were capable of MTX hydrolysis that did not degrade the drug in the HPLC assay. One of the predicted strains with negative metabolism results, *E. coli* BW25113, possesses AcrAB multidrug resistance efflux pump, an efflux pump that actively exports MTX (Kopytek et al., 2000). Due to the possibility that similar machinery might exist in the other isolates negative for MTX metabolism despite being predicted by SIMMER, I looked for AcrAB sequence in the genomes of all 42 bacterial strains. From this analysis I found two more SIMMER predicted, MTX negative strains containing sequences similar to AcrAB: *Edwardsiella tarda* and *Providencia rettgerri* (Figure 3.5). Thus, it is possible that these SIMMER predicted MTX degraders are functional, but failed HPLC detection because active MTX efflux prevented interaction between substrate and enzyme.

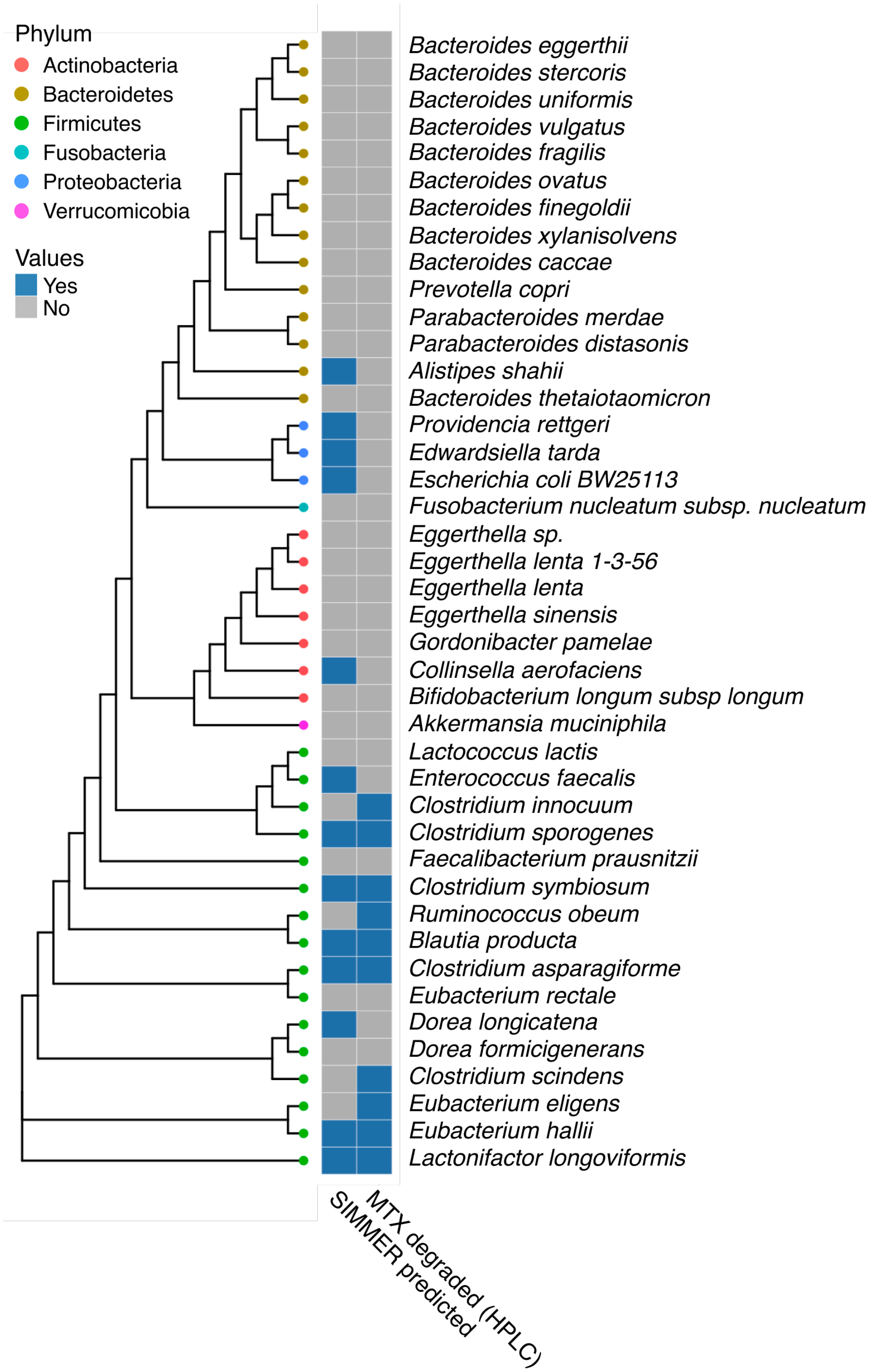


Figure 3.5 SIMMER accurately predicted bacterial strains capable of MTX degradation. A diverse panel of 42 isolates was incubated with MTX, and degradation (yes/no) measured via HPLC. SIMMER predicted (yes/no) that thirteen of the 42 isolates were capable of MTX

metabolism, and HPLC experiments showed that ten isolates were capable of MTX degradation (SIMMER prediction p-value=0.046). The approximately-maximum likelihood phylogenetic tree (FastTree) was created using 16S rRNA gene sequences from the 42 organisms.

3.2.4 Experimental validation of SIMMER predicted enzymes capable of MTX metabolism

Before experimental validation of SIMMER's enzyme candidates, I narrowed in on the subset of hits most likely to be physiologically relevant for MTX metabolism. As discussed in Chapter 2, while SIMMER provides highly accurate enzyme predictions, there is the possibility of false positives in its sequence output that could not be computationally assessed. To mitigate this potential, I filtered SIMMER results before experimentation, as recommended in Chapter 2. SIMMER automatically filters by length; target sequences must be at least 50% the length of the input pHMM. I additionally filtered SIMMER's output by sequences' predicted ligand affinity, localization in the human gastrointestinal tract, and presence in metagenomics data. To predict SIMMER output sequences' affinities for MTX, I used the Similarity Ensemble Approach which filtered SIMMER's output down to 158 sequences (Keiser et al. 2007). Because MTX is absorbed in the proximal intestine (Murakami and Mori, 2012), I also filtered SIMMER's output to sequences (N=289) found in the ileum and jejunum via DIAMOND searches against metagenomic reads from a study that employed jejunum and ileum endoscopy (Buchfink et al., 2021; Zmora et al., 2018). Lastly, SIMMER hits were filtered for their presence (N=479 sequences) in two metagenomics datasets, as described above in section 3.2.2.

SIMMER's predicted enzymes had a median global identity of 33% to a Carboxypeptidase G2 enzyme with solved crystal structures and validated activity assays (Jeyaharan et al., 2016). As such, we were able to adapt the previously developed methods to test MTX metabolism activity of my SIMMER predicted enzymes. From the filtered enzyme prediction list, we chose twenty candidates (Table 3.2) for experimental validation using a heterologous expression system. Ten of the twenty SIMMER predicted enzymes were chosen due to their presence in isolates capable of MTX degradation (Figure 3.5), and the other ten chosen due to their presence in uncultured metagenome-assembled genomes.

Table 3.2 SIMMER MTX metabolism predictions chosen for enzyme activity experiments.

UHGG ID	ExperimentID	Genus	species	Prokka function	Length (amino acid)
GUT_GENOME 141054_02582	RAB005	*Clostridium*	*asparagiforme*	Carboxypeptidase G2	386
GUT_GENOME 096178_00385	RAB006	*Clostridium*	*symbiosum*	Carboxypeptidase G2	392
GUT_GENOME 000594_01519	RAB007	*Blautia*	*producta*	N-acetyl-lysine deacetylase	419
GUT_GENOME 096250_01715	RAB008	*Clostridium*	*innocuum*	Putative dipeptidase	463
GUT_GENOME 095973_00847	RAB009	*Clostridium*	*scindens*	putative succinyl-diaminopimelate desuccinylase	437
GUT_GENOME 001261_02056	RAB010	*Blautia*	*obeum*	putative succinyl-diaminopimelate desuccinylase	436
GUT_GENOME 000237_01573	RAB011	*Eubacterium*	*eligens*	Cytosol non-specific dipeptidase	493
GUT_GENOME 001689_01808	RAB012	*Eubacterium*	*halliii*	N-acetyl-lysine deacetylase	467

UHGG ID	**ExperimentID**	**Genus**	**species**	**Prokka function**	**Length (amino acid)**
GUT_GENOME 095982_01544	RAB013	*Clostridium*	*sporogenes*	Putative dipeptidase	463
GUT_GENOME 143497_01051	RAB014	*Lactonifactor*	*longoviformis*	Acetylornithine deacetylase	394
GUT_GENOME 001770_03586	RAB015	*Roseburia*	*inulinivorans*	Carboxypeptidase G2	369
GUT_GENOME 096166_03560	RAB016	*Bilophila*	*wadsworthia*	Carboxypeptidase G2	409
GUT_GENOME 096292_02896	RAB017	*Alistipes*	*senegalensis*	Succinyl-diaminopimelate desuccinylase	358
GUT_GENOME 115272_01369	RAB018	*Akkermansia*	*s__*	Peptidase T	423
GUT_GENOME 159026_01716	RAB019	*Adlercreutzia*	*s__*	Peptidase T	396
GUT_GENOME 103893_02292	RAB020	*Coprococcus_A*	*catus*	putative succinyl-diaminopimelate desuccinylase	387
GUT_GENOME 258044_01425	RAB021	*Dialister*	*invisus*	Peptidase T	399
GUT_GENOME 001208_00104	RAB022	*CAG-81*	*sp900066535*	Carboxypeptidase G2	374
GUT_GENOME 096178_00659	RAB023	*Clostridium_Q*	*symbiosum*	N-carbamoyl-L-amino acid hydrolase	413
GUT_GENOME 143760_01903	RAB024	*Enterobacter*	*himalayensis*	Peptidase T	409

Prior to gene sequencing, N-terminal signal peptides, when present, were removed from all SIMMER candidates and controls (Almagro Armenteros et al., 2019; Jeyaharan et al., 2016). All candidates were then codon optimized and successfully cloned into pET28a vectors with a 6X histidine affinity tag (Figure 3.6). After transformation into BL21 (DE3) cells, high expression was achieved with an Isopropyl β-D-1-thiogalactopyranoside (IPTG) induction system for N=6 of the 20 candidates (Figure 3.7). These six proteins were then purified with immobilized metal affinity enrichment and elution supernatant (Figure 3.8) taken forward for UV/Vis activity assays.

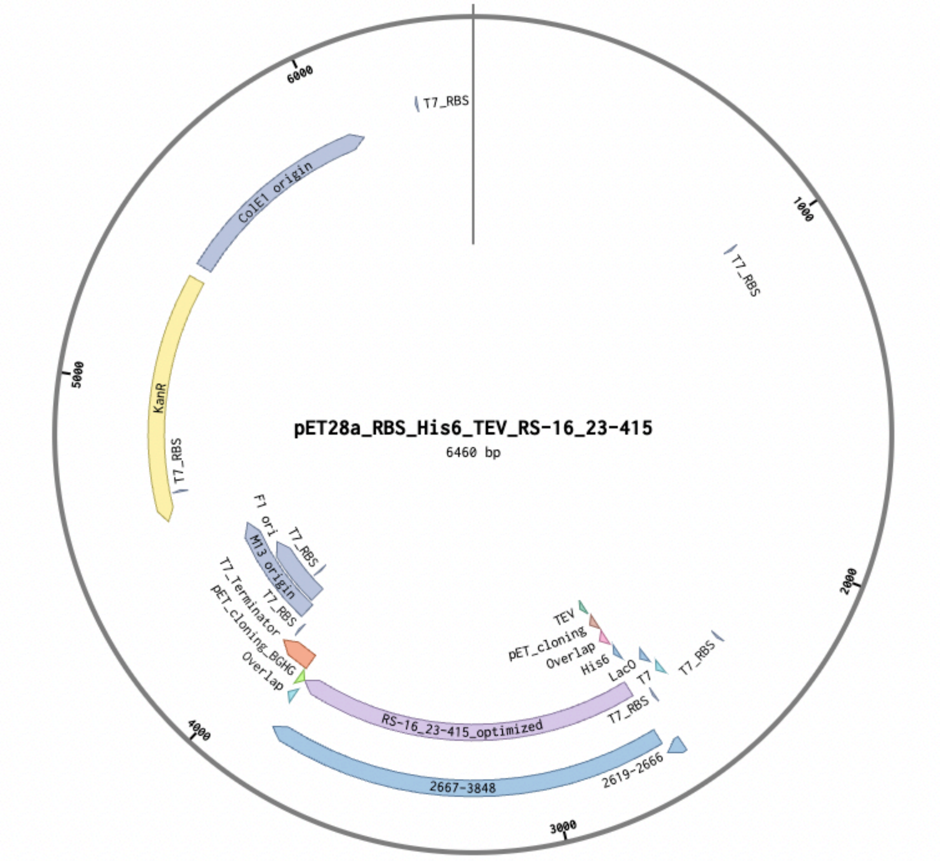


Figure 3.6 Constructed expression plasmid.

pET28a vector design with candidate insertion site shown in purple (RS-16 positive control example).

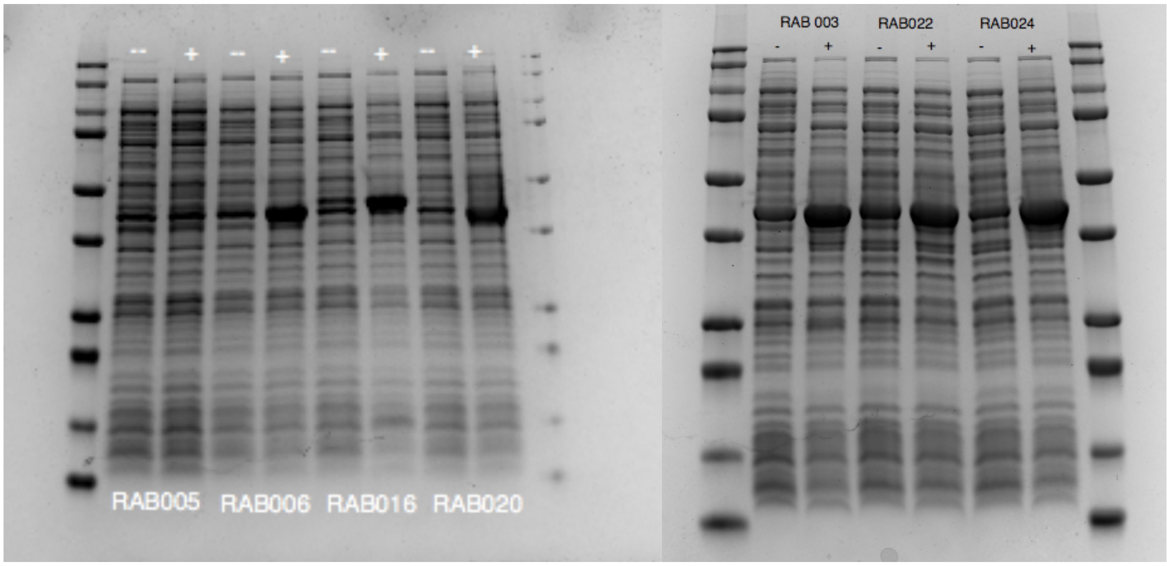


Figure 3.7 Six candidates were expressed via IPTG induction system.
Of the twenty experimental candidates (Table 3.2), six could be overexpressed with IPTG as confirmed by SDS-Page gel. His6-TEV-Cpg2 weight is ~43 kDA. There are two lanes for each experimental sample, one labeled '-' as a pre-induction sample, and a second labeled '+' as a post-induction sample. Overexpression is observable (except for RAB005) based on the darkened post-induction labels at 43 kDa for each sample.

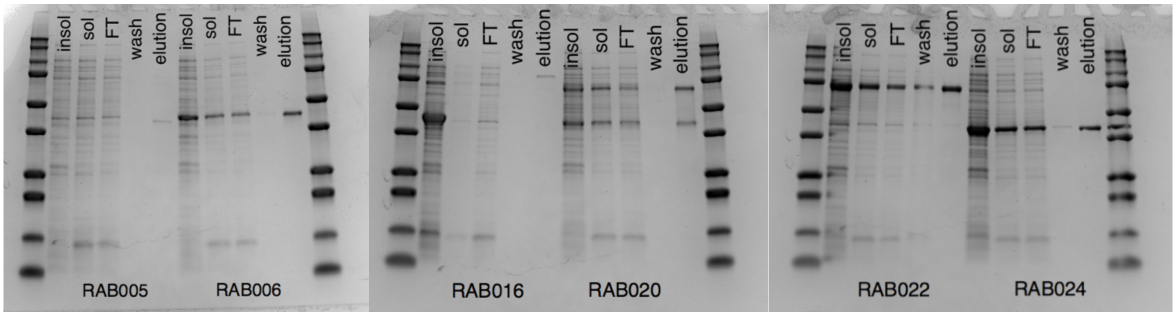

insol sol FT wash elution insol sol FT wash elution
RAB005 RAB006
insol sol FT wash elution insol sol FT wash elution
RAB016 RAB020
insol sol FT wash elution insol sol FT wash elution
RAB022 RAB024

Figure 3.8 His-tagged heterologous candidates were enriched from expression cultures.
The six candidates overexpressed with IPTG were enriched before conducting activity assays with elution supernatant. Of note, most of candidate RAB016 was found in insoluble fraction after the immobilized metal affinity enrichment. Also, candidates RAB020 and RAB022 appear to form dimers.

Activity of SIMMER candidates was assessed with UV/Vis, and defined as a purified candidate's ability to cleave MTX (absorbs at 320nm) into DAMPA (non-absorbing) as previously demonstrated in the literature (Jeyaharan et al., 2016). Positive control protein (WT RS) efficiently broke down MTX, and negative control protein (RS with a scrambled active site design) did not (Figure 3.8). When SIMMER candidates were tested, however, no activity was observed. Metabolism experiments were repeated in whole cells assessed via HPLC to follow up on this lack of activity, but again, no degradation of MTX was observed outside of the positive control assay.

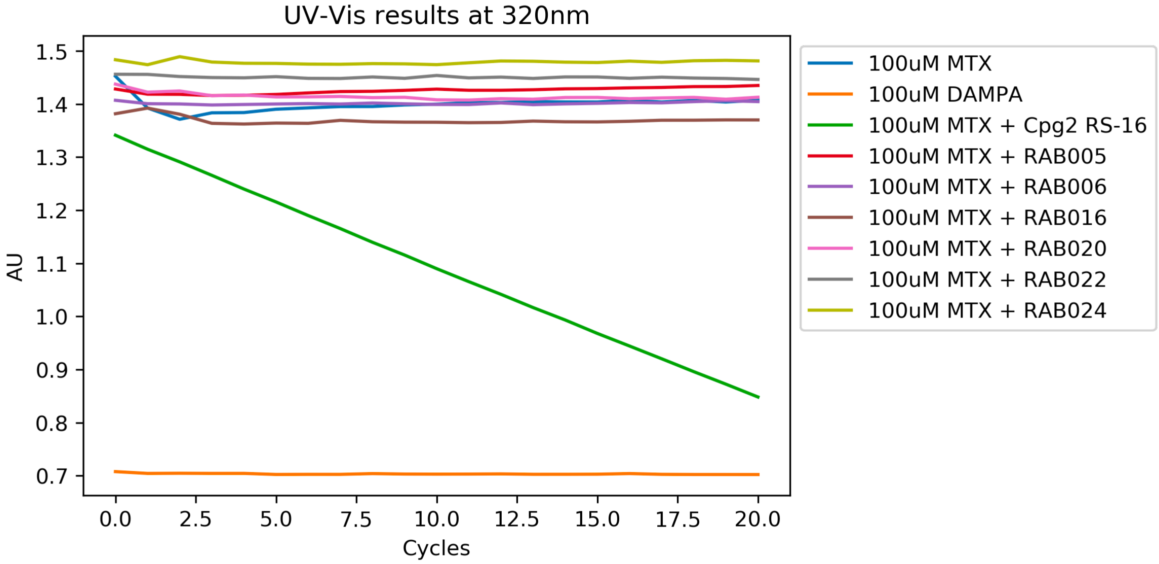


Figure 3.9 MTX degradation is observed only in positive control Cpg2.
Degradation of MTX, defined as a shift from 320nm absorption, was only observed in purified Cpg2 RS-16 positive control (green line). None of the six purified SIMMER candidates exhibited metabolism activity.

From previous literature on CPG2, we were aware that these proteins are often improperly folded and require extensive testing with refolding protocols to restore activity. Following a previously established protocol that determined the optimal assay for RS16 CPG2, we were able to reproduce RS16 CPG2 as a soluble protein with MTX degrading activity as a positive control (Jeyaharan et al., 2016). We also produced soluble fractions of our SIMMER candidates, but the proteins may have been different enough in sequence that the folding protocol for RS16 was not sufficient to yield active proteins.

3.3 Discussion

While it has been suspected for some time now that inter-individual variation in patient response to oral MTX may be due to bacterial hydrolysis of the therapeutic, no characterizations have yet been made of the species of enzymes responsible. This chapter demonstrated the utility of using SIMMER in such a use-case. The most similar MetaCyc reaction to MTX hydrolysis, folate hydrolysis, that SIMMER identified makes much logical sense based on substrate and metabolite structures, and overall reaction chemistry. Similarly, SIMMER's enzyme predictions, largely annotated as Carboxypeptidase G2s, are in line with what would have been expected from the existing glucardipase literature. Further adding confidence to SIMMER's sequence predictions was their enrichment in the shotgun sequenced stool samples of RA patients exhibiting little response to MTX treatment over the course of four months.

This same study of MTX responders versus non-responders also demonstrated for the first time depletion of MTX in human *ex vivo* stool samples (Artacho et al., 2020). Here, we demonstrate for the first time a species level characterization of MTX metabolism in the human gut microbiome. SIMMER significantly predicted which isolates would or would not be able to metabolize MTX in a diverse collection of 42 species representative of the gut environment. When we went on to directly test the SIMMER predicted enzymes harbored in the metabolizing genomes, however, we were unable to demonstrate activity.

While this might signal false-positive enzyme predictions on the part of SIMMER, it may also be the case that despite proper induction and enrichment, candidates were inactive due to protein misfolding. To rule out such a possibility, future experiments will employ a range of treatment conditions while purifying protein candidates. Indeed, the original study that our treatment conditions are based upon required multiple treatment condition iterations before finding a method that consistently yielded soluble, active protein (Jeyaharan et al., 2016).

In sum, this work grows the body of experimental evidence pointing to the microbiome as a source of interindividual variation in response to MTX therapy. It also demonstrates for the first time that human gut microbiome species are directly capable of MTX metabolism, providing a potential mechanism for the aforementioned non-response to the drug. Lastly, this work also provides experimental validation of a SIMMER species-drug metabolism prediction.

3.4 Materials and methods

Methotrexate metabolism predictions

Bacterial enzymes capable of metabolizing MTX were predicted using SIMMER software (Chapter 2, (Bustion et al., 2022)). Input substrates, cofactors, and products were formatted as isomeric SMILES obtained from compounds' PubChem entries. Before input to SIMMER, all SMILES were protonated based on the pH environment of 7.4 in the human small intestine, where most oral drug absorption occurs. Protonation states were calculated using ChemAxon's cxcalc majorms software ("cxcalc calculator functions," n.d.).

Analysis of RA metagenomics data

Presence of SIMMER sequences was assessed in two metagenomics studies of RA patients on MTX therapy. Both sequencing studies' raw reads were assembled using MetaSpades with default settings (Nurk et al., 2017). After assembly, DIAMOND was used to search for SIMMER sequences in reads, with presence defined as at least 50% coverage and at least 97% identity. All abundance measures were normalized by read depth. For the Zhang et al. cohort, correlation between disease severity and number of aligned SIMMER predictions was assessed using

Pearson's correlation coefficient and p-value calculated using a student's *t*-distribution. Enrichment of SIMMER predictions in low, moderate, or high activity disease groups was assessed using a generalized linear model, glm(count~disease, family = poisson). For the Artacho et al. cohort, correlation between MTX response and number of aligned SIMMER predictions was assessed using Pearson's correlation coefficient and p-value calculated using a student's *t*-distribution. Enrichment of SIMMER predictions in MTX responders versus non-responders was assessed using a generalized linear model, glm(count~response, family = poisson).

Bacterial isolate screen for MTX hydrolysis

42 isolates commonly found in the human gut microbiome were obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) culture collection and subcultured as previously described (Nayak et al., 2021). MTX (100μg/mL) was added to cultures, samples then spun down, and supernatant injected to HPLC. MTX was dosed based on the predicted concentration of the drug in a human gastrointestinal tract, as previously described (Nayak et al., 2021). An approximately-maximum likelihood phylogenetic tree (FastTree) was created using 16S rRNA gene sequences from the 42 organisms. SIMMER predicted enzymes' presence or absence in the 42 isolates was determined by downloading genomes for all 42 isolates and conducting DIAMOND searches. Presence was defined as at least 97% global percent identity. Presence of AcrAB was determined in this manner as well.

Selection of enzyme candidates for heterologous expression

SIMMER candidates for MTX hydrolysis were at adequate percent identity (median 33% global percent identity) to previously expressed proteins for this reaction, but underwent additional filtering for biological relevance. First, SIMMER sequences were filtered by predicted affinity for MTX using the Similarity Ensemble Approach webserver (Keiser et al., 2007). Any SIMMER sequences matched (E-value <= 0.05) to Chembl proteins with predicted affinity (p-value <= 0.05) for MTX were retained.

The hits were next filtered by presence in human ileum or jejunum (the site of methotrexate absorption) via DIAMOND searches against metagenomic reads from a published study that employed jejunum and ileum endoscopy (Buchfink et al., 2021; Zmora et al., 2018). Before running DIAMOND, sequencing runs from the study were decontaminated of any host reads via KneadData with default settings and assembled via MetaSpades with default settings (*kneaddata: Quality control tool on metagenomic and metatranscriptomic sequencing data, especially data from microbiome experiments*, n.d.; Nurk et al., 2017). Finally, twenty candidates were chosen from this list based on their presence in MTX metabolizing organisms described in Section 3.2.3 or based on their presence in metagenome-assembled genomes (i.e. as yet uncultured organisms). In addition to these candidates, we chose Pseudomonas sp. RS-16 CPG2 (Uniprot P06621) as a positive control for MTX metabolism (Jeyaharan et al., 2016), and Pseudomonas sp. RS-16 CPG2 with a mutated active site (D119A) as a negative control for MTX metabolism (Jeyaharan et al., 2018).

Candidate cloning into pET28a vectors

All MTX metabolism candidate and control sequences were assessed for the presence of a signal peptide using the SignalP 5.0 web server for gram-negative bacteria (Almagro Armenteros et al., 2019). Only the positive and negative controls were predicted to possess signal peptides, and these were truncated before designing constructs. All candidates and signal peptide cleaved controls were then constructed with flanking sequences for insertion into a pET28a vector (Figure 3.6) and synthesized as codon optimized gBlocks Gene Fragments from Integrated DNA Technologies. The synthesized gBlocks were assembled into linearized pET28a using NEB Gibson Assembly. Gibson products were transformed into NEB 5-alpha Competent *E. coli* (High Efficiency) cells and their sequences confirmed by Sanger Sequencing.

Protein expression

After PCR and sequence confirmation, constructed plasmids were transformed into *E. coli* expression strain BL21 (DE3) cells (Thermo Scientific). After overnight incubation at 37C in TB broth supplemented with 50 µg/mL kanamycin, cells were subcultured 1:200 into TB + KPO_4 media supplemented with 50 µg/mL kanamycin and grown at 37C. Once overnight cultures reached OD_{600}~0.6, the protein candidates were induced by adding isopropyl β-d-1-thiogalactopyranoside (IPTG) to concentration of 0.5mM and left to shake overnight at 16C. Adequate expression was confirmed by SDS Page Gel for 6 of 20 constructs (Figure 3.7), and cells were harvested by centrifugation at 4000g for 15 minutes at 4C. Before protein purification, samples were flash frozen in liquid nitrogen, and stored at -80C.

Protein purification

Harvested cell pellets from the overnight induction were resuspended in Buffer A (50 mM sodium phosphate, 150 mM NaCl, and 5% glycerol) supplemented with 110 mM benzamidine, 0.4 mM AEBSF, 0.3 mg/mL DNase, Lysozyme, and 5mM β-Mercaptoethanol. Resuspensions were lysed by sonication (1/8" microtip, 2 seconds on, 2 seconds off, for six minutes, at 20% amplitude). Lysis product was then centrifuged at 4C and 30,000g for 30 minutes. 500uL of clean (cleaned 1X in 20% ethanol, and 3X in lysis buffer) Contech TALON Superflow Resin was added to each supernatant and batch bound for thirty minutes at 4C. Beads were then spun down at 500rpm for 1-2 minutes at room temperature, supernatant poured off, and washed 2X at 500rpm with 5mL of Buffer A plus 5mM β-Mercaptoethanol. Buffer B (Buffer A supplemented with 400 mM imidazole) plus 5mM β-Mercaptoethanol was added, and beads spun down a final time at 500rpm for 1-2 minutes. Elution supernatants were snap frozen and stored at -80C until used in activity assays.

Protein activity assays

Purified proteins were assessed for their ability to degrade MTX (absorbs at 320 nm) into DAMPA and glutamate (lower absorption species) via UV/VIS Spectrometer as described previously in CPG2 research literature (Jeyaharan et al., 2016). 100uL of purified protein samples were loaded into quartz cuvettes with 900uL of Minimal Hepes Buffer (25mM Hepes, 25mM NaCl, 0.2mM Zinc acetate) + 100μM MTX solution, and A_{320}measured on UV-VIS for 30 second kinetic cycles for 800-200nm wavelengths, stopping at ten minutes. Absorbance versus time was plotted for each sample.

Whole cell activity assays

In addition to activity assays of purified protein we tested activity of the transformed *E. coli* BL21 (DE3) cells expressing our protein candidates. For this, the protein expression process was identical to above, except that MTX (100µg/mL) was added to overnight cultures two hours after induction with IPTG. Induced samples were then spun down and pulled supernatant injected to HPLC. When activity was not seen with pulled supernatant, we also assessed the saved cell pellets. For this, we thawed pellets from overnight cultures at room temperature, and then resuspended in 1mL methanol before sonicating (1/8” microtip, 2 seconds on, 2 seconds off, for one minute, at 20% amplitude). Sonicated samples were spun down at 15,000g for 15 minutes at 4C, and the resulting supernatant injected to HPLC.

Methotrexate HPLC method

HPLC assays were performed on an Agilent HPLC (1220 Infinity), and data collected with OpenLAB CDS (Agilent Technologies). Solvent A was 0.1% formic acid, and solvent B was 100% methanol. Solvent B concentration was 10–30% from 0–1 minute, 30–100% from 1–7 minutes, and then 100–10% from 7–7.5 minutes. The flow rate was 0.6 mL/min. A C18 column (Kinetex 2.6 µM; 100 Å; 15 cm × 0.46 cm; Phenomenex; 00F-4462-E0) was used with a SecurityGuard ULTRA cartridge guard column (Phenomenex part number AJ0-8768). The injection volume was 30 µl. At 320 nm, MTX retention time was 5.5 minutes. We compared the amount of MTX present in the bacterial supernatant compared to sterile and DMSO controls to assess MTX metabolism.

Chapter 4

Conclusions and Perspectives

While the field of pharmacogenomics is well established, pharmacomicrobiomics lags behind. In part this can be attributed to the added complexity of many, rather than one, genomes that must be evaluated for drug metabolism genes. The number of studies exploring drug metabolism within the human gut microbiome is growing constantly, and recent studies employ higher-throughput analysis of representative strain collections or *ex vivo* human stool samples. Analyzing the genetic elements responsible, however, is still a low-throughput process. In this dissertation, I have made the case for developing computational methods that can rescue current bottlenecks in pharmacomicrobiomics data, specifically identification of the bacterial species and enzymes responsible for single step biotransformations. I laid out the landscape of current tools, commented on needed improvements, and then created an accurate predictive method based on my proposed recommendations.

This project leveraged the wealth of recent experimental data that points to the gut microbiome as a source of drug-metabolizing enzymes and expanded prior work by determining the specific genes (i.e., enzyme sequences) responsible for xenobiotic degradations. From this work, a more complete picture emerged for interindividual variation in the metabolism of drugs such as dexamethasone and methotrexate. Beyond species predictions though, more experimental work is required to verify that the predicted enzymes are sufficient for the metabolism events in question. As discussed in Chapter 2, the possibility of false positive output from SIMMER is high, as this was not computationally assessed. For this reason, despite having a high true positive rate, SIMMER predictions will not be immediately relevant in a clinical setting. To demonstrate clinical utility of SIMMER enzyme predictions, subsequent analysis must be performed against site-specific human gut microbiome metagenomics studies, as shown in Chapter 3 of this dissertation. Future work must also employ pharmacokinetic modeling (PK) to disentangle the effects of host and microbial metabolic machinery (Mendez-Catala et al., 2020; Wu, 2012; Zimmermann et al., 2019b). Other recent xenobiotic metabolism modeling efforts include the use of metabolic reconstructions that leverage community interactions to explain interindividual variation in drug conversion (Heinken et al., 2020).

Related to clinical efficacy, to date there are no systematic analyses that can be performed on the full set of FDA compounds to assess potential for bacterial metabolism. While the tool developed here is currently focused on known drug metabolism events, in the future SIMMER could be applied to all FDA-approved compounds and their theoretical metabolites predicted using tools like Biotransformer (Djoumbou-Feunang et al., 2019). Introducing these theoretical drug metabolism events to the gene retrieval pipeline may enable the discovery of new chemical transformation events.

A benefit of SIMMER's design is that its underlying chemistry and protein spaces can easily be modified to reflect updates to the MetaCyc and UHGG databases. This can also be extended to pharmacomicrobiomics databases once they become more comprehensive. What results is a method whose accuracy and resolution improve as the field's knowledge-base grows.

During the course of conducting this research, staggering technical advances were made in the field of protein structure predictions, and now methods like AlphaFold and the ESMFold (evolutionary scale modeling) allow users to predict 3D structures from any primary protein sequence (Jumper et al., 2021; Lin et al., 2022). ESMFold research presents a resource particularly relevant to microbiome research in the form of the ESM Metagenomic Atlas, a compendium of over 600 million structural predictions from largely unannotated metagenomic sequencing data (Lin et al., 2022). As they immediately relate to this thesis work, use of such structural prediction tools and atlases may be an aid while designing heterologous expression and protein activity assays for validation of enzyme predictions.

One day, structural prediction tools like AlphaFold and ESMFold may additionally enable large-scale and accurate reverse molecular docking (the docking of a small molecule against many potential protein targets), which would be a valuable augmentation to the work presented in this thesis (e.g., by replacing the filtering step of SEA in Chapter 3). As combined protein structure prediction and docking methods currently stand, however, reverse docking with predicted structures has resulted in weak model performance (Wong et al., 2022). Another future application of structure prediction methods as they relate to my thesis work are improved homology searches. In Chapters 1 and 2 I stressed the importance of using sequence search algorithms like pHMMs which retain evolutionarily meaningful information, such as binding and active sites. As protein structure prediction algorithms improve, one could further improve homology searches by directly comparing overall 3D structure (van Kempen et al., 2022).

All in all, what stems from my thesis research is a novel computational tool that can rescue the knowledge bottlenecks that plague pharmacomicrobiomics research. By using it, scientists researching drug metabolism in the human gut can filter their search space down to the species and enzymes most likely relevant to the biotransformation in question. An exciting aspect of this work is that it can easily scale to accommodate additional database knowledge, whether it be bacterial chemistry reflected in MetaCyc, new sequencing data reflected in UHGG, or improved protein structural prediction data.

References

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423. doi:10.1038/s41587-019-0036-z

Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2020. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. doi:10.1038/s41587-020-0603-3

Altman T, Travers M, Kothari A, Caspi R, Karp PD. 2013. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics 14:112. doi:10.1186/1471-2105-14-112

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Artacho A, Isaac S, Nayak R, Flor-Duro A, Alexander M, Koo I, Manasson J, Smith PB, Rosenthal P, Homsi Y, Gulko P, Pons J, Puchades-Carrasco L, Izmirly P, Patterson A, Abramson SB, Pineda-Lucena A, Turnbaugh PJ, Ubeda C, Scher JU. 2020. The Pre-treatment Gut Microbiome is Associated with Lack of Response to Methotrexate in New Onset Rheumatoid Arthritis. Arthritis Rheumatol. doi:10.1002/art.41622

Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, Leeming E, Gibson R, Le Roy C, Khatib HA, Francis L, Mazidi M, Mompeo O, Valles-Colomer M, Tett A, Beghini F, Dubois L, Bazzani D, Thomas AM, Mirzayi C, Khleborodova A, Oh S, Hine R, Bonnett C, Capdevila J, Danzanvilliers S, Giordano F, Geistlinger L, Waldron L, Davies R, Hadjigeorgiou G, Wolf J, Ordovás JM, Gardner C, Franks PW, Chan AT, Huttenhower C, Spector TD, Segata N. 2021. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat Med. doi:10.1038/s41591-020-01183-8

Aziz RK, Hegazy SM, Yasser R, Rizkallah MR, ElRakaiby MT. 2018. Drug pharmacomicrobiomics and toxicomicrobiomics: from scattered reports to systematic studies of drug-microbiome interactions. Expert Opin Drug Metab Toxicol 14:1043–1055. doi:10.1080/17425255.2018.1530216

Bairoch A. 2000. The ENZYME database in 2000. Nucleic Acids Res 28:304–305. doi:10.1093/nar/28.1.304

Barrett AJ, Rawlings ND, O'Brien EA. 2001. The MEROPS database as a protease information system. *J Struct Biol* **134**:95–102. doi:10.1006/jsbi.2000.4332

Bayjanov JR, Molenaar D, Tzeneva V, Siezen RJ, van Hijum SAFT. 2012. PhenoLink--a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. BMC Genomics 13:170. doi:10.1186/1471-2164-13-170

Bisanz JE, Soto-Perez P, Noecker C, Aksenov AA, Lam KN, Kenney GE, Bess EN, Haiser HJ, Kyaw TS, Yu FB, Rekdal VM, Ha CWY, Devkota S, Balskus EP, Dorrestein PC, Allen-Vercoe E, Turnbaugh PJ. 2020. A Genomic Toolkit for the Mechanistic Dissection of Intractable Human Gut Bacteria. Cell Host Microbe 27:1001–1013.e9. doi:10.1016/j.chom.2020.04.006

Bisanz JE, Spanogiannopoulos P, Pieper LM, Bustion AE, Turnbaugh PJ. 2018. How to Determine the Role of the Microbiome in Drug Disposition. Drug Metab Dispos 46:1588–1595. doi:10.1124/dmd.118.083402

Buchen S, Ngampolo D, Melton RG, Hasan C, Zoubek A, Henze G, Bode U, Fleischhack G. 2005. Carboxypeptidase G2 rescue in patients with methotrexate intoxication and renal failure. Br J Cancer 92:480–487. doi:10.1038/sj.bjc.6602337

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. doi:10.1038/s41592-021-01101-x

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience 8. doi:10.1093/gigascience/giz100

Bustion AE, Agrawal A, Turnbaugh PJ, Pollard KS. 2022. A novel in silico method employs chemical and protein similarity algorithms to accurately identify chemical transformations in the human gut microbiome. bioRxiv. doi:10.1101/2022.08.02.502504

Cao L, Shcherbin E, Mohimani H. 2019. A Metabolome- and Metagenome-Wide Association Network Reveals Microbial Natural Products and Microbial Biotransformation Products from the Human Microbiota. mSystems 4. doi:10.1128/mSystems.00387-19

Capecchi A, Probst D, Reymond J-L. 2020. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform 12:43. doi:10.1186/s13321-020-00445-4

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD. 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36:D623–31. doi:10.1093/nar/gkm900

Chabner BA, Johns DG, Bertino JR. 1972. Enzymatic cleavage of methotrexate provides a method for prevention of drug toxicity. Nature 239:395–397. doi:10.1038/239395b0

Chen Y, Ye W, Zhang Y, Xu Y. 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 43:7762–7768. doi:10.1093/nar/gkv784

cxcalc calculator functions. n.d. https://docs.chemaxon.com/display/docs/cxcalc-calculator-functions.md

Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS. 2019. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J Cheminform 11:2. doi:10.1186/s13321-018-0324-5

Duan J, Dixon SL, Lowrie JF, Sherman W. 2010. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. J Mol Graph Model 29:157–170. doi:10.1016/j.jmgm.2010.05.008

Duigou T, du Lac M, Carbonell P, Faulon J-L. 2019. RetroRules: a database of reaction rules for engineering biology. Nucleic Acids Res 47:D1229–D1235. doi:10.1093/nar/gky940

Eddy SR. 2009. A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCEGenome Informatics 2009. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO. pp. 205–211. doi:10.1142/9781848165632_0019

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763. doi:10.1093/bioinformatics/14.9.755

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi:10.1093/nar/gkh340

Emery P, Breedveld FC, Hall S, Durez P, Chang DJ, Robertson D, Singh A, Pedersen RD, Koenig AS, Freundlich B. 2008. Comparison of methotrexate monotherapy with a combination of methotrexate and etanercept in active, early, moderate to severe rheumatoid arthritis (COMET): a randomised, double-blind, parallel treatment trial. *Lancet* **372**:375–382. doi:10.1016/S0140-6736(08)61000-4

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. doi:10.1093/bioinformatics/bts565

Gerlt JA, Babbitt PC, Jacobson MP, Almo SC. 2012. Divergent evolution in enolase superfamily: strategies for assigning functions. J Biol Chem 287:29–34. doi:10.1074/jbc.R111.240945

Gibson MK, Forsberg KJ, Dantas G. 2015. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 9:207–216. doi:10.1038/ismej.2014.106

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol 8:2224. doi:10.3389/fmicb.2017.02224

Guthrie L, Wolfson S, Kelly L. 2019. The human gut chemical landscape predicts microbe-mediated biotransformation of foods and drugs. Elife 8. doi:10.7554/eLife.42866

Haiser HJ, Seim KL, Balskus EP, Turnbaugh PJ. 2014. Mechanistic insight into digoxin inactivation by Eggerthella lenta augments our understanding of its pharmacokinetics. Gut Microbes 5:233–238. doi:10.4161/gmic.27915

Heinken A, Acharya G, Ravcheev DA, Hertel J, Nyga M, Okpala OE, Hogan M, Magnúsdóttir S, Martinelli F, Preciat G, Edirisinghe JN, Henry CS, Fleming RMT, Thiele I. 2020. AGORA2: Large scale reconstruction of the microbiome highlights wide-spread drug-metabolising capacities. bioRxiv. doi:10.1101/2020.11.09.375451

Integrative HMP (iHMP) Research Network Consortium. 2019. The Integrative Human Microbiome Project. Nature 569:641–648. doi:10.1038/s41586-019-1238-8

Javdan B, Lopez JG, Chankhamjon P, Lee Y-CJ, Hull R, Wu Q, Wang X, Chatterjee S, Donia MS. 2020. Personalized Mapping of Drug Metabolism by the Human Gut Microbiome. Cell 181:1661–1679.e22. doi:10.1016/j.cell.2020.05.001

Jeyaharan D, Aston P, Garcia-Perez A, Schouten J, Davis P, Dixon AM. 2016. Soluble expression, purification and functional characterisation of carboxypeptidase G2 and its individual domains. Protein Expr Purif 127:44–52. doi:10.1016/j.pep.2016.06.015

Jeyaharan D, Brackstone C, Schouten J, Davis P, Dixon AM. 2018. Characterisation of the Carboxypeptidase G2 Catalytic Site and Design of New Inhibitors for Cancer Therapy. Chembiochem 19:1959–1968. doi:10.1002/cbic.201800186

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. doi:10.1038/s41586-021-03819-2

Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. 2007. Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25:197–206. doi:10.1038/nbt1284

Kishikawa T, Maeda Y, Nii T, Motooka D, Matsumoto Y, Matsushita M, Matsuoka H, Yoshimura M, Kawada S, Teshigawara S, Oguro E, Okita Y, Kawamoto K, Higa S, Hirano T, Narazaki M, Ogata A, Saeki Y, Nakamura S, Inohara H, Kumanogoh A, Takeda K, Okada Y. 2020. Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. Ann Rheum Dis 79:103–111. doi:10.1136/annrheumdis-2019-215743

Klatt NR, Cheu R, Birse K, Zevin AS, Perner M, Noël-Romas L, Grobler A, Westmacott G, Xie IY, Butler J, Mansoor L, McKinnon LR, Passmore J-AS, Abdool Karim Q, Abdool Karim SS, Burgener AD. 2017. Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women. Science 356:938–945. doi:10.1126/science.aai9383

kneaddata: Quality control tool on metagenomic and metatranscriptomic sequencing data, especially data from microbiome experiments. n.d. . Github.

Koppel N, Bisanz JE, Pandelia M-E, Turnbaugh PJ, Balskus EP. 2018. Discovery and characterization of a prevalent human gut bacterial enzyme sufficient for the inactivation of a family of plant toxins. *Elife* **7**. doi:10.7554/eLife.33953

Koppel N, Maini Rekdal V, Balskus EP. 2017. Chemical transformation of xenobiotics by the human gut microbiota. Science 356. doi:10.1126/science.aag2770

Kopytek SJ, Dyer JC, Knapp GS, Hu JC. 2000. Resistance to methotrexate due to AcrAB-dependent export from Escherichia coli. Antimicrob Agents Chemother 44:3210–3212. doi:10.1128/AAC.44.11.3210-3212.2000

Letertre MPM, Munjoma N, Wolfer K, Pechlivanis A, McDonald JAK, Hardwick RN, Cherrington NJ, Coen M, Nicholson JK, Hoyles L, Swann JR, Wilson ID. 2020. A Two-Way Interaction between Methotrexate and the Gut Microbiota of Male Sprague-Dawley Rats. J Proteome Res 19:3326–3339. doi:10.1021/acs.jproteome.0c00230

Levy CC, Goldman P. 1967. The enzymatic hydrolysis of methotrexate and folic acid. J Biol Chem 242:2933–2938. doi:10.1016/S0021-9258(18)99594-3

Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, Kristiansen K, Guarner F, Pedersen O, Doré J, Ehrlich SD, MetaHIT Consortium, Bork P, Wang J, MetaHIT Consortium. 2014. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 32:834–841. doi:10.1038/nbt.2942

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. 2022. Evolutionary-scale prediction of atomic level protein structure with a language model. bioRxiv. doi:10.1101/2022.07.20.500902

Ly LK, Rowles JL 3rd, Paul HM, Alves JMP, Yemm C, Wolf PM, Devendran S, Hudson ME, Morris DJ, Erdman JW Jr, Ridlon JM. 2020. Bacterial steroid-17,20-desmolase is a taxonomically rare enzymatic pathway that converts prednisone to 1,4-androstanediene-3,11,17-trione, a metabolite that causes proliferation of prostate cancer cells. J Steroid Biochem Mol Biol 199:105567. doi:10.1016/j.jsbmb.2019.105567

Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jäger C, Baginska J, Wilmes P, Fleming RMT, Thiele I. 2017. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol 35:81–89. doi:10.1038/nbt.3703

Maini Rekdal V, Bess EN, Bisanz JE, Turnbaugh PJ, Balskus EP. 2019. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. Science 364. doi:10.1126/science.aau6323

Maini Rekdal V, Nol Bernadino P, Luescher MU, Kiamehr S, Le C, Bisanz JE, Turnbaugh PJ, Bess EN, Balskus EP. 2020. A widely distributed metalloenzyme class enables gut microbial metabolism of host- and diet-derived catechols. Elife 9. doi:10.7554/eLife.50845

Mallory EK, Acharya A, Rensi SE, Turnbaugh PJ, Bright RA, Altman RB. 2018. Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome. Pac Symp Biocomput 23:56–67. doi:10.1142/9789813235533_0006

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40:D115–22. doi:10.1093/nar/gkr1044

Maurice CF, Haiser HJ, Turnbaugh PJ. 2013. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell 152:39–50. doi:10.1016/j.cell.2012.10.052

McDonald AG, Boyce S, Tipton KF. 2009. ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res 37:D593–7. doi:10.1093/nar/gkn582

McDonald AG, Tipton KF. 2021. Enzyme nomenclature and classification: the state of the art. *FEBS J*. doi:10.1111/febs.16274

McDonald AG, Tipton KF. 2014. Fifty-five years of enzyme classification: advances and difficulties. *FEBS J* **281**:583–592. doi:10.1111/febs.12530

Melnik AV, da Silva RR, Hyde ER, Aksenov AA, Vargas F, Bouslimani A, Protsyuk I, Jarmusch AK, Tripathi A, Alexandrov T, Knight R, Dorrestein PC. 2017. Coupling Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide Association Studies of Human Fecal Samples. Anal Chem 89:7549–7559. doi:10.1021/acs.analchem.7b01381

Mendez-Catala DM, Spenkelink A, Rietjens IMCM, Beekmann K. 2020. An in vitro model to quantify interspecies differences in kinetics for intestinal microbial bioactivation and detoxification of zearalenone. Toxicology Reports. doi:10.1016/j.toxrep.2020.07.010

Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD. 2021. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res 49:D394–D403. doi:10.1093/nar/gkaa1106

Murakami T, Mori N. 2012. Involvement of Multiple Transporters-mediated Transports in Mizoribine and Methotrexate Pharmacokinetics. Pharmaceuticals 5:802–836. doi:10.3390/ph5080802

Nayak RR, Alexander M, Deshpande I, Stapleton-Gray K, Rimal B, Patterson AD, Ubeda C, Scher JU, Turnbaugh PJ. 2021. Methotrexate impacts conserved pathways in diverse human gut bacteria leading to decreased host immune activation. Cell Host Microbe. doi:10.1016/j.chom.2020.12.008

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res 26:1612–1625. doi:10.1101/gr.201863.115

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. doi:10.1101/gr.213959.116

Patel JR, Oh J, Wang S, Crawford JM, Isaacs FJ. 2022. Cross-kingdom expression of synthetic genetic elements promotes discovery of metabolites in the human microbiome. Cell 185:1487–1505.e14. doi:10.1016/j.cell.2022.03.008

Pollet RM, D'Agostino EH, Walton WG, Xu Y, Little MS, Biernat KA, Pellock SJ, Patterson LM, Creekmore BC, Isenberg HN, Bahethi RR, Bhatt AP, Liu J, Gharaibeh RZ, Redinbo MR. 2017. An Atlas of β-Glucuronidases in the Human Intestinal Microbiome. Structure 25:967–977.e5. doi:10.1016/j.str.2017.05.003

Pouliot Y, Karp PD. 2007. A survey of orphan enzyme activities. *BMC Bioinformatics* **8**:244. doi:10.1186/1471-2105-8-244

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650. doi:10.1093/molbev/msp077

Rizkallah MR, Gamal-Eldin S, Saad R, Aziz RK. n.d. The PharmacoMicrobiomics Portal: A Database for Drug-Microbiome Interactions. Curr Pharmacogenomics Person Med 10:195–203.

Rizkallah MR, Saad R, Aziz RK. 2010. The Human Microbiome Project, Personalized Medicine and the Birth of Pharmacomicrobiomics. Curr Pharmacogenomics Person Med 8:182–193. doi:10.2174/187569210792246326

Roon E, Laar MVD. 2006. Methotrexate bioavailability.

Scher JU, Nayak RR, Ubeda C, Turnbaugh PJ, Abramson SB. 2020. Pharmacomicrobiomics in inflammatory arthritis: gut microbiome as modulator of therapeutic response. Nat Rev Rheumatol 16:282–292. doi:10.1038/s41584-020-0395-3

Schneider N, Lowe DM, Sayle RA, Landrum GA. 2015. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. J Chem Inf Model 55:39–53. doi:10.1021/ci5006614

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. doi:10.1093/bioinformatics/btu153

Sharma AK, Jaiswal SK, Chaudhary N, Sharma VK. 2017. A novel approach for the prediction of species-specific biotransformation of xenobiotic/drug molecules by the human gut microbiota. Sci Rep 7:9751. doi:10.1038/s41598-017-10203-6

Sievers F, Higgins DG. 2014. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences In: Russell DJ, editor. Multiple Sequence Alignment Methods. Totowa, NJ: Humana Press. pp. 105–116. doi:10.1007/978-1-62703-646-7_6

Spanogiannopoulos P, Bess EN, Carmody RN, Turnbaugh PJ. 2016. The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. Nat Rev Microbiol 14:273–287. doi:10.1038/nrmicro.2016.17

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550. doi:10.1073/pnas.0506580102

Sun Y-Z, Zhang D-H, Cai S-B, Ming Z, Li J-Q, Chen X. 2018. MDAD: A Special Resource for Microbe-Drug Associations. Front Cell Infect Microbiol 8:424. doi:10.3389/fcimb.2018.00424

Thomas AM, Segata N. 2019. Multiple levels of the unknown in microbiome research. BMC Biol 17:48. doi:10.1186/s12915-019-0667-z

Thorn CF, Klein TE, Altman RB. 2013. PharmGKB: The Pharmacogenomics Knowledge Base In: Innocenti F, van Schaik RHN, editors. Pharmacogenomics: Methods and Protocols. Totowa, NJ: Humana Press. pp. 311–320. doi:10.1007/978-1-62703-435-7_20

Tian W, Skolnick J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333:863–882. doi:10.1016/j.jmb.2003.08.057

Valerino DM, Johns DG, Zaharko DS, Oliverio VT. 1972. Studies of the metabolism of methotrexate by intestinal flora—I: Identification and study of biological properties of the metabolite 4-amino-4-deoxy-N10-methylpteroic acid. Biochem Pharmacol 21:821–831. doi:10.1016/0006-2952(72)90125-6

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Gilchrist CLM, Söding J, Steinegger M. 2022. Foldseek: fast and accurate protein structure search. bioRxiv. doi:10.1101/2022.02.07.479398

Wild DJ, Blankley CJ. 2000. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. J Chem Inf Comput Sci 40:155–162. doi:10.1021/ci990086j

Wong F, Krishnan A, Zheng EJ, Stärk H, Manson AL, Earl AM, Jaakkola T, Collins JJ. 2022. Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. Mol Syst Biol 18:e11081. doi:10.15252/msb.202211081

Wu B. 2012. Use of physiologically based pharmacokinetic models to evaluate the impact of intestinal glucuronide hydrolysis on the pharmacokinetics of aglycone. J Pharm Sci 101:1281–1301. doi:10.1002/jps.22827

Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, Tang W, Feng T, Chen M, Zhan L, Wu T, Hu E, Jiang Y, Bo X, Yu G. 2021. ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. Mol Biol Evol 38:4039–4042. doi:10.1093/molbev/msab166

Yan D, Cao L, Zhou M, Mohimani H. 2022. TransDiscovery: Discovering Biotransformation from Human Microbiota by Integrating Metagenomic and Metabolomic Data. Metabolites 12:119. doi:10.3390/metabo12020119

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. doi:10.1111/2041-210x.12628

Zeng X, Yang X, Fan J, Tan Y, Ju L, Shen W, Wang Y, Wang X, Chen W, Ju D, Chen YZ. 2021. MASI: microbiota-active substance interactions database. Nucleic Acids Res 49:D776–D782. doi:10.1093/nar/gkaa924

Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, Wu X, Li J, Tang L, Li Y, Lan Z, Chen B, Li Y, Zhong H, Xie H, Jie Z, Chen W, Tang S, Xu X, Wang X, Cai X, Liu S, Xia Y, Li J, Qiao X, Al-Aama JY, Chen H, Wang L, Wu Q-J, Zhang F, Zheng W, Li Y, Zhang M, Luo G, Xue W, Xiao L, Li J, Chen W, Xu X, Yin Y, Yang H, Wang J, Kristiansen K, Liu L, Li T, Huang Q, Li Y, Wang J. 2015. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med 21:895–905. doi:10.1038/nm.3914

Zhao C, Dimitrov B, Goldman M, Nayfach S, Pollard KS. 2022. MIDAS2: Metagenomic Intra-species Diversity Analysis System. bioRxiv. doi:10.1101/2022.06.16.496510

Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL. 2019a. Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature. doi:10.1038/s41586-019-1291-3

Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL. 2019b. Separating host and microbiome contributions to drug pharmacokinetics and toxicity. Science 363. doi:10.1126/science.aat9931

Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, Bashiardes S, Kotler E, Zur M, Regev-Lehavi D, Brik RB-Z, Federici S, Cohen Y, Linevsky R, Rothschild D, Moor AE, Ben-Moshe S, Harmelin A, Itzkovitz S, Maharshak N, Shibolet O, Shapiro H, Pevsner-Fischer M, Sharon I, Halpern Z, Segal E, Elinav E. 2018. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. Cell 174:1388–1405.e21. doi:10.1016/j.cell.2018.08.041

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
16E82C01E16B4BE...
Author Signature

11/29/2022
Date