

Analysis of Human Gut Metabolites as Sources for the Design of Gut-Targeted Drugs

Alberto Gil-Pichardo,[†] Andrés Sánchez-Ruiz,[†] and
Gonzalo Colmenarejo^{*}

Biostatistics and Bioinformatics Unit

IMDEA Food

CEI UAM+CSIC

E28049 Madrid, Spain

^{*}Corresponding Author

e-mail: gonzalo.colmenarejo@imdea.org

[†]These two authors contributed equally to this work

ABSTRACT

The recent explosion in gut microbiome research has demonstrated the importance of metabolite-target interactions in the development of different pathologies. This suggests that gut-targeted drugs modulating these interactions would provide a new drug modality besides that of systemically bioavailable small molecules, that could tap from this growing knowledge, and would have little distribution and safety issues. In the present work we analyze a large set of gut metabolites in comparison with serum metabolites and drugs. We find structural and physicochemical similarity between the serum metabolites and the drug sets, and dissimilarity with the gut metabolite set. In addition, we find that the inclusion of chemical class is necessary in order to appropriately understand gut permeance, in contrast to classical oral permeation models (e.g. rule-of-five). To help in gut-targeted drug design, we provide a simple scoring scheme for use in medicinal chemistry, plus a machine learning model to use in cheminformatic applications.

KEYWORDS:

Gut-targeted drugs, gut microbiome, gut metabolome, new drug modalities, drug design, physicochemical properties

INTRODUCTION

The search for new modalities for drug discovery is an area currently receiving a large research focus.¹⁻³ The existence of “undruggable” targets, unmet therapeutic needs in many areas, and the high attrition rates due to safety and distribution issues, calls for alternative approaches to the traditional paradigm of orally, systemically bioavailable small molecules. Among the new modalities, we find oligonucleotide therapeutics,^{4,5} “beyond rule of 5” (bRo5) molecules,^{6,7} protein degraders (e.g. PROTACS⁸), peptides,^{9,10} and biologics.^{11,12} In addition, new knowledge coming from omics technologies is expanding our understanding of the molecular mechanisms and pathways involved in biological processes, that result in new paradigms for drug discovery requiring new modalities. One of the most important of these paradigms stems from the growing knowledge in the last decade about the crucial role of microbiota on human health. The human body hosts trillions of microbial cells, mainly localized in the gut, that carry a genome (the microbiome) about 100 times the size of the human genome.¹³⁻¹⁵ The evidence for the involvement of the gut microbiome in multiple pathologies keeps steadily increasing, in areas like obesity, type 2 diabetes, cardiometabolic diseases, non-alcoholic liver disease, diverticulitis, inflammatory bowel disease, colon cancer, etc.¹⁶⁻²³ From this research, a recurrent picture that emerges is that of host-microbiome interactions mechanistically mediated through metabolites in the gut that bind bacterial or human targets.^{21,22,24-29} In turn, the metabolites would be bacterial, endogenous, or xenobiotics (food, drugs, environmental), or modified versions of any of these produced by putative bacterial and/or host enzymes.

Thus, given all this knowledge, the modulation of all these gut metabolite-target interactions, appears as an interesting new drug modality that would tap from the new targets, pathways, and chemotypes appearing from the human microbiome research, as has been suggested.³⁰⁻³² This would create new opportunities for treating diseases like the ones mentioned above, plus others like intestinal infectious diseases. Moreover, the ability to modulate the bacterial populations in the gut through new chemicals would pave the way for preventive interventions (instead of curative ones) through novel nutraceuticals.

In addition, this new modality would benefit from much reduced distribution and safety issues, as long as the compound is designed to remain in the gut: the administration route would be oral, but with a much more efficient access to the target (it would only require a minimal metabolic stability), and a minimal probability of off-target effects as the compound would not be distributed through the whole body.

Given all this background, in the present work we aim at characterizing the specific features of gut metabolites in order to provide tools for the rational design of gut-targeted drugs and nutraceuticals. It is well known that systemic drugs have a higher resemblance to systemic metabolites than random compounds, which can be rationalized in terms of structural similarity allowing them to compete with endogenous metabolites for their interaction with their targets, or with their transporters.³³⁻³⁷ This is translated into a restrained physicochemical profile that has resulted in different sets of rules and computational methods to predict oral permeability and bioavailability (e.g. the well-known Lipinski rule-of-five (Ro5))³⁸⁻⁴³. In the same way, for the design of gut-targeted drugs the physicochemical characterization of gut metabolites done here

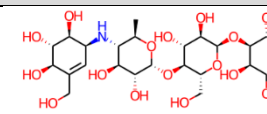
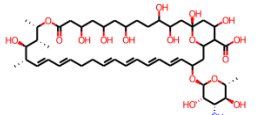
would provide insights about patterns and features that these drugs would require. We analyzed a wide range of physicochemical properties, solubility, and ionic class of gut metabolites in comparison with systemic metabolites and drugs, and found significant differences that strongly depended on the chemical class. In addition, in order to predict gut permanence from molecular structures, we tested the use of reversed versions of oral permeability rules like Ro5³⁸ or Veber's,⁴⁰ finding a low predictive power due to not considering the chemical class in the rules. By including this factor in the prediction, we were able to derive a) a simple scoring system of easy interpretation to guide medicinal chemistry efforts, and b) a machine learning model for reliable in silico prediction of gut permanence in chemical databases for cheminformatics efforts.

RESULTS

In what follows, we will start by describing the few gut-targeted drugs that currently exist, and after that we will perform an extensive analysis of gut metabolites. For that we will use the set of detected (quantified or not) gut compounds from the Human Metabolome Database (HMDB),⁴⁴ corresponding to the feces biospecimen, further processed as described before^{45–47} (see also Materials and Methods), which comprises a total of 5021 molecules. For comparison purposes, two additional compound sets are included in the analysis: the set of detected (quantified or not) serum metabolites from the HMDB as systemic metabolites (16621 molecules), and a set of drug molecules corresponding to the subset of small molecules in approved, not withdrawn, and non-illicit status of the DrugBank (1623 molecules); both additional sets were processed as before.^{45–47} The idea is to identify physicochemical patterns that are specific for gut metabolites, so that they can be used in the design of drugs targeted to the gut, instead of the normal paradigm of systemic drugs. In addition, we develop two predictive tools for gut permanence, useful in medicinal chemistry settings and in cheminformatic settings, respectively.

Existing gut-targeted drugs

There are a few cases of drugs that act in the gut. A collection of them is shown in Table 1.

NAME	CHEMICAL CLASS	INDICATION	MODE OF ACTION	Structure
Acarbose	Organic oxygen compounds	Type 2 diabetes	α -glucosidase and α -amilase inhibitor	
Nystatin	Organic oxygen compounds	Antifungal	Channel-forming ionophore	

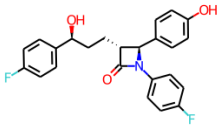
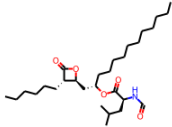
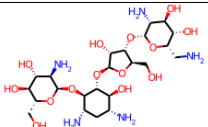
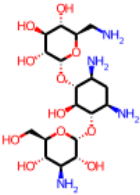
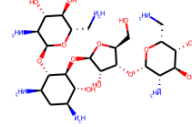
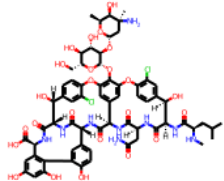
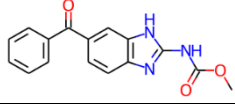
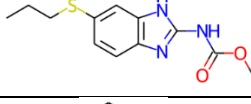
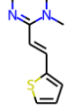
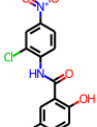
NAME	CHEMICAL CLASS	INDICATION	MODE OF ACTION	Structure
Ezetimibe	Organoheterocyclic compounds	Hypercholesterolemia	NPC1L1 cholesterol transporter inhibitor	
Orlistat	Organic acids and derivatives	Obesity	Lipase inhibitor	
Paromomycin	Organic oxygen compounds	Antibiotic, antiamoebic	Ribosome inhibitor	
Kanamycin	Organic oxygen compound	Antibiotic	Ribosome inhibitor	
Neomycin	Organic oxygen compounds	Antibiotic	Ribosome inhibition	
Vancomycin	Organic acids and derivatives	Antibiotic	Peptidoglycan synthesis inhibitor (transpeptidase)	
Mebendazole	Benzenoids	Anthelmintic	Inhibition of tubulin polymerization	
Albendazole	Organoheterocyclic compounds	Anthelmintic	Inhibition of tubulin polymerization	
Pyrantel	Organoheterocyclic compounds	Anthelmintic	Cholinesterase inhibition	
Nicosamide	Benzenoids	Anthelmintic	Uncoupling of oxydative phosphorylation	

Table 1. Set of gut-acting drugs. Data derived from DrugBank. Drugs were selected if they had a low or null bioavailability, together with a well-defined human or bacterial target (protein or ribonucleoprotein) located in the intestine. Drugs acting through

non-specific physicochemical mechanisms (osmotic laxatives, surfactants, ion exchange resins, etc.), or with high bioavailability, were discarded.

These molecules have different chemotypes and targets, but all of them have low or null systemic bioavailability. On one hand, we have several aminoglycoside antibiotics, that act through inhibition of the bacterial ribosome (paromomycin, kanamycin, and neomycin). Other antibiotic targeting a bacterial target is vancomycin, a glycopeptide, but in this case the bacterial transpeptidase used for the synthesis of peptidoglycan is inhibited. Several molecules, all of them with heterocyclic structures, have anthelmintic activity, like mebendazole and albendazole, which target tubulin polymerization in the worm; pyrantel, which target its cholinesterase; and niclosamide, which uncouple the parasite oxidative phosphorylation. One aminoglycoside compound, nystatin, is an antifungal that acts as pore-forming ionophore. Finally, there are three drugs acting upon human targets: acarbose, an oligosaccharide that inhibits pancreatic amylases and gut α -glucosidases; ezetimibe, an heterocyclic molecule, that inhibits gut NPC1L1 cholesterol transporter; and orlistat, a triglyceride analog that inhibits gastric and pancreatic lipases. These are used in the treatment of type-2 diabetes, hypercholesterolemia, and obesity, respectively.

From these examples we see that the concept of gut-directed drugs has already some exemplars that pave the way for more systematic and extensive drug design efforts, especially after novel metabolite-target interactions relevant for diseases are identified through the gut microbiome research.

Chemical classification of metabolites

Figure 1 displays the distribution gut metabolites in 18 chemical classes, based on the ClassyFire chemical taxonomy.⁴⁸ For comparison purposes, the distributions for serum metabolites and DrugBank molecules are also provided.

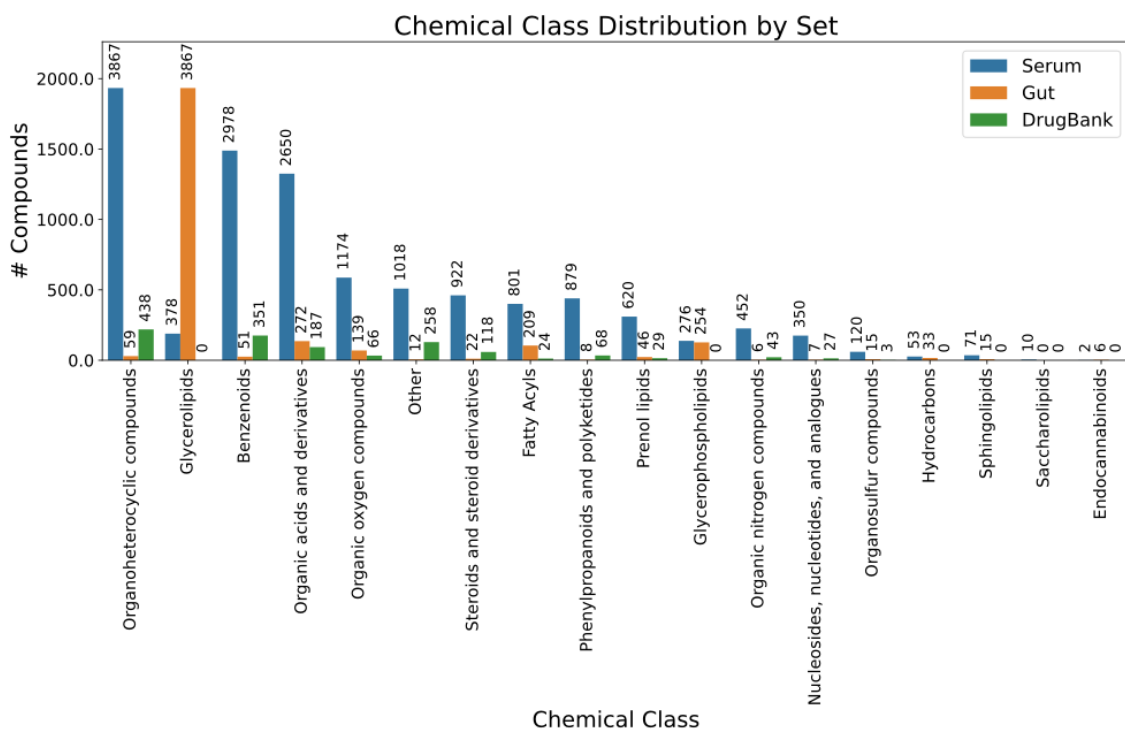


Figure 1. Distribution of chemical classes (based on the ClassyFire taxonomy) for gut metabolites, serum metabolites, and DrugBank molecules.

These classes are quite diverse from the structural point of view, and include some that are not present in the DrugBank set, like “glycerolipids”, “glycerophospholipids”, “sphingolipids”, “hydrocarbons”, “saccharolipids”, and “endocannabinoids”. The other chemical classes are shared with both the DrugBank set and the serum metabolites sets. They are, in decreasing order of abundance in the gut metabolites set (in parenthesis some example chemotypes in the gut are shown): “organic acids and derivatives” (oligopeptides, short carboxylic acids and derivatives, amino acids and derivatives, etc.); “fatty acyls” (fatty acids and derivatives); “organic oxygen compounds” (sugars,

oligosaccharides, alcohols, ketones, etc.); “organoheterocyclic compounds” (e.g. indoles, pyrroles, lactones, etc. and their derivatives); “benzenoids” (e.g. derivatives from benzene, benzoic acid, and phenol mainly); “prenol lipids” (terpenoids, quinones, hydroquinones, etc.); “steroids and steroid derivatives” (bile acid derivatives, cholesterol derivatives, etc.); “organosulfur compounds”; “other”; “phenylpropanoids and polyketides” (mainly flavonoids); “nucleosides, nucleotides, and analogues”; and “organic nitrogen compounds” (amines and nitriles).

The distribution of chemical classes in the gut set is quite different from the other sets. For instance, we see that gut metabolites are highly enriched in “glycerolipids”, followed (in decreasing order) by “organic acids and derivatives”, “glycerophospholipids”, “fatty acyls”, and “organic oxygen compounds”. Chemical classes like “organoheterocyclic compounds” and “benzenoids” are the 6th and 7th most abundant ones. These gut top chemical classes are in contrast with those of the DrugBank and the serum sets, that have more similar distributions. For example, both of them have “organoheterocyclic compounds” and “benzenoids” as the first and second most frequent chemical classes, respectively. In the case of the serum set, these are followed (in decreasing order) by “organic acids and derivatives” (3rd), “organic oxygen compounds” (4th), “other” (5th), and “steroid and steroid derivatives” (6th), while in the case of the DrugBank set these are followed by “other” (3rd), “organic acids and derivatives” (4th), “steroids and steroid derivatives” (5th), and “phenylpropanoids and polyketides” (6th). In turn, the gut set has the latter chemical class as the 14th one.

Thus, it seems that the serum metabolites set have chemotypes (as represented by their chemical classes) resembling more closely that of systemic drugs, that are the

overwhelming majority of the DrugBank set. This supports previous observations regarding the higher structural similarity observed in systemic drugs to systemic metabolites over those of random compounds,³³⁻³⁵ and would point towards the optional use of additional alternative chemotypes in the case of putative gut-targeted drugs.

For example, the large count of “glycerolipids” (mainly triglycerides) in the gut metabolites set is due to the high variety of these molecules in food, and the fact that they are unable to cross the gut wall. Their triglyceride forms have to be hydrolyzed by lipases in the gut in order to be absorbed by the intestine epithelium, where they are again resynthesized and released to the circulation in the form of chylomicrons.⁴⁹ This fact is used by some lipase inhibitors like orlistat (see above), an anti-obesity drug with minimal absorption in the intestine, that act locally through the inhibition of triglyceride hydrolysis and therefore their intestinal absorption. This drug and other lipase inhibitors act through irreversible competitive inhibition of the lipase catalytic center,⁵⁰ as they are substrate analogs. In a similar vein is acarbose, a substrate analog of the highly abundant oligosaccharides in the gut, that is used to inhibit α -glucosidases and α -amylases in the intestinal lumen. These are examples of alternative chemotypes not typical in systemic drugs that have been used to design successful gut-targeted drugs.

Ionic class analysis

Another interesting aspect to analyze is the comparative ionization behavior of these molecules. Figure 2 shows the distribution of ionization classes (acid, basic, neutral, and zwitterion) among the three compound sets.

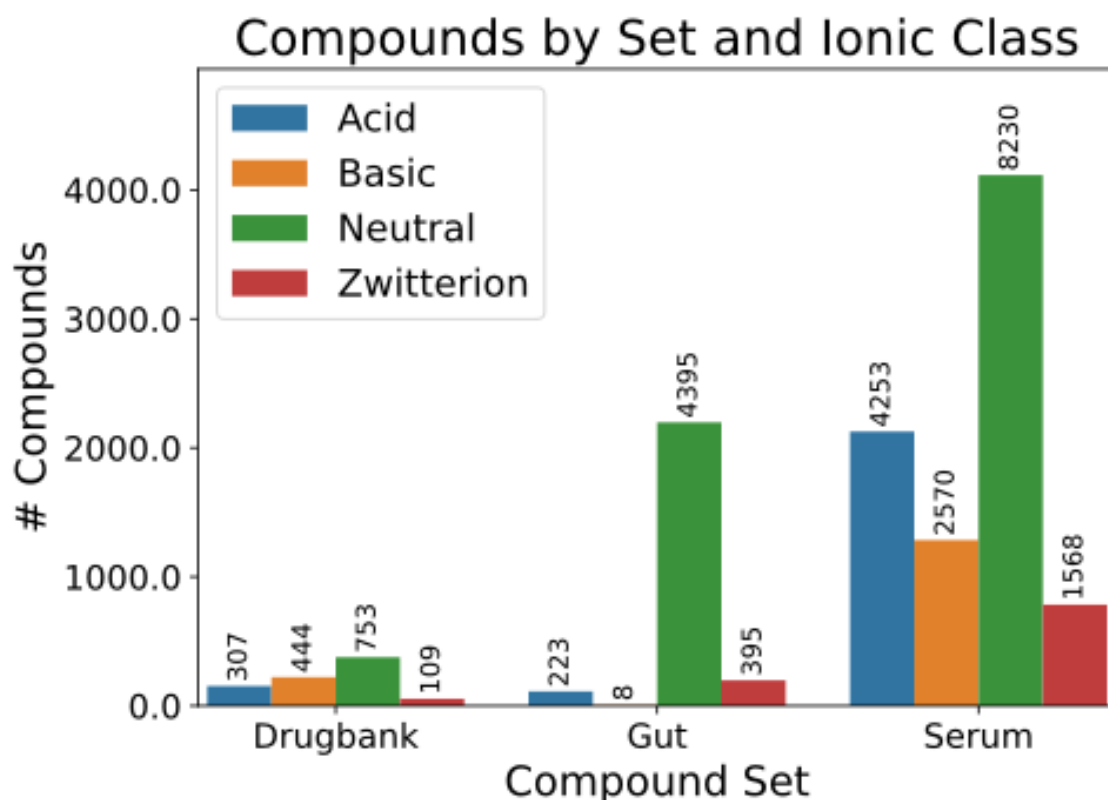


Figure 2. Distribution of ionization states across the three compound sets: DrugBank, gut metabolites, and serum metabolites.

From Figure 2 we see that, treated as a whole group, gut metabolites have a much higher proportion of neutral molecules. The large excess in neutral molecules is mainly the result of the large number of “glycerolipids” in this set, as all of them are neutral (see below). This exceeding large number of neutral molecules is in contrast to what is observed in drug molecules and serum metabolites. In the latter two sets, neutral molecules are the most abundant ionization class too, but the other ionization classes show relatively higher proportions, with the following decreasing order of abundance: basic > acid > zwitterion for the drugs, and acid > basic > zwitterions in the case of serum metabolites. In the distribution of ionization classes in gut metabolites it is also quite

remarkable the different the distribution of ionized forms, where zwitterions is the most abundant one, followed by acid molecules; basic molecules are quite infrequent.

If we focus in the comparison of serum vs gut metabolites, we get a more convoluted picture than this simple approximation when we take into account the chemical classes.

Figure 3a shows the distribution of ionization classes for the whole set of gut metabolites (“all” column), as well as across compound classes, while Figure 3b the results of the statistical analysis of the adjusted residuals of ionization classes for the comparison of gut vs serum molecules, for both the complete two sets (“all” column), or for the different compound classes.

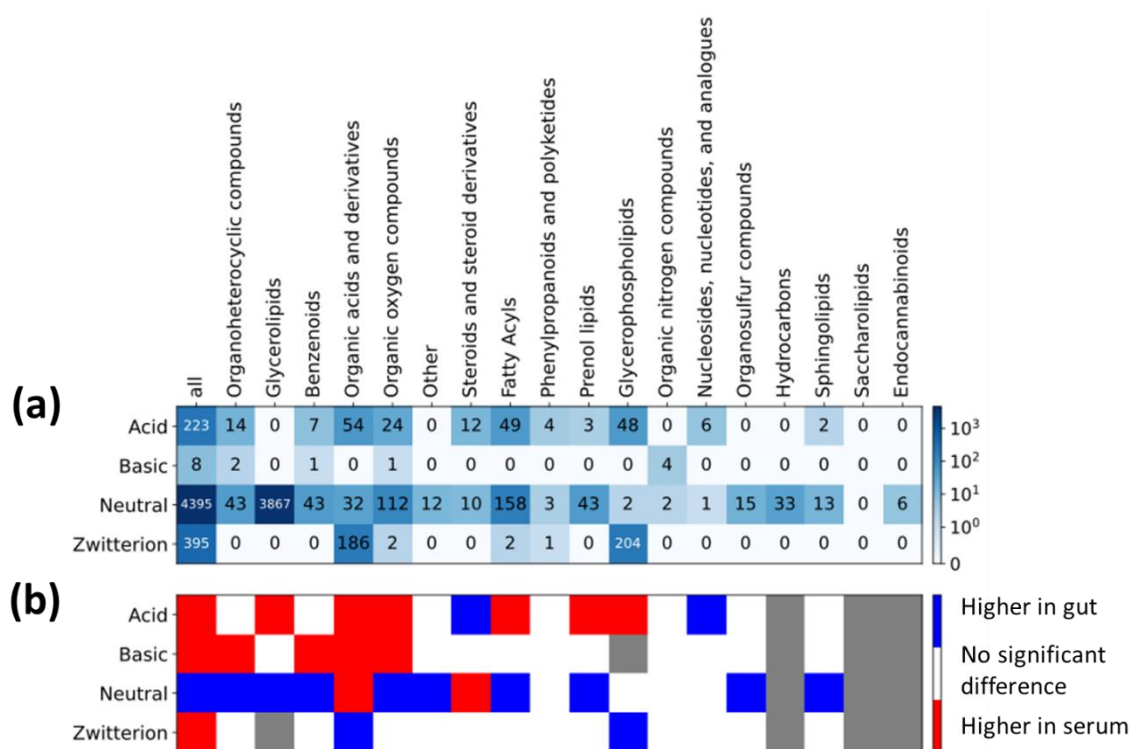


Figure 3. (a) Ionization class distribution across compound classes for gut compounds.

(b) Ionization state enrichment for gut vs serum across compound classes. For all the compounds or particular compound classes, adjusted residuals were calculated for the contingency table of ionic class vs set, followed by a Fisher exact post hoc analysis.

Red cells correspond to significant (p -value < 0.05) enrichment of the ionic class in the serum set, while blue cells correspond to significant enrichment in the gut set. Grey cells correspond to cells for which no test was feasible due to missing compounds in one or both of gut and serum metabolites.

We can see that the enrichment of gut compounds in neutral molecules is, as expected, statistically significant (blue cell in the “all” column), as it is the underrepresentation of acid, basic, and zwitterionic molecules (red cells). By focusing on particular chemical classes this pattern seems to hold in several cases: in “organoheterocyclic compounds”, “glycerolipids”, “benzenoids”, “organic oxygen compounds”, “other”, “fatty acyls”, “prenol lipids”, “organosulfur compounds” and “sphingolipids” the neutral class is significantly overrepresented over that in serum metabolites, while the basic class, acid class, or both are significantly underrepresented. However, we can also see several reversals for the rule of enriched neutral class: for example, the zwitterion ionic class is enriched in the gut set for “organic acids and derivatives” and “glycerophospholipids” (they would be the main responsible for the proportionally higher number of zwitterions in gut metabolites), while the acid ionic class is enriched in the “steroids and steroid derivatives” and “nucleosides, nucleotides, and analogues”. This mixed pattern of preferred ionic class for the gut depending on the compound class offers useful patterns for the design of gut-targeted drugs and nutraceuticals for the different compound classes, as far as the ionic class to choose is concerned.

Analysis of solubility

An important property in drug design is water solubility. We used the logS predicted property available in HMDB as proxy for the water solubility values of the molecules.

Figure 4 compares through a nonparametric test the logs values of gut vs serum metabolites, for both the whole compound sets (“all” column) and across the different chemical classes.

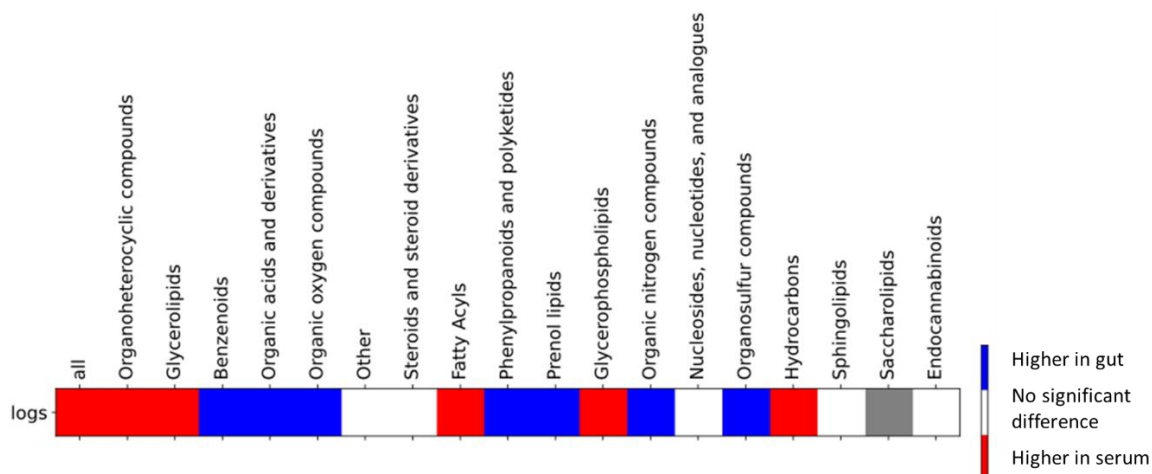


Figure 4. Water solubility (logs) enrichment for gut vs serum compound classes. For all the compounds (“all” column) or particular chemical classes, a nonparametric Mann-Whitney test was conducted, and the Common-Language Effect Size statistic (CLES) was computed. Red cells correspond to significantly (p -value < 0.05) higher solubility in the serum set (negative CLES), while blue cells correspond to significantly higher solubility in the gut set (positive CLES). Grey cells correspond to cells for which no test was feasible due to missing compounds in one or both of gut and serum metabolites.

It is possible to see that, considered as complete compound sets, gut metabolites have lower solubility than serum ones. However, this behavior is not always maintained across the different chemical classes, and as a matter of fact, gut metabolites display significantly higher solubilities in gut for seven chemical classes (“benzenoids”, “organic acids and derivatives”, “organic oxygen compounds”, “phenylpropanoids and

polyketides”, “prenol lipids”, “organic nitrogen compounds”, and “organosulfur compounds”) vs four classes where they display significantly lower solubilities (“glycerolipids”, “fatty acyls”, “glycerophospholipids”, and “hydrocarbons”).

Other physicochemical properties

To get a more complete idea of additional physicochemical patterns present in gut metabolites, as opposed to serum ones, we analyzed a large set of much used physicochemical properties, namely: topological polar surface area (tpsa), logarithm of octanol/water partition coefficient (logp), number of rotatable bonds (rb), number of hydrogen bond donors (hbd), number of hydrogen bond acceptors (hba), molecular weight (mw), number of rings (nring), number of aromatic rings (naring), quantitative estimation of drug-likeness⁵¹ (qed), and fraction of sp³-hybridized carbons (fsp3). Figure 5 displays the distributions of these properties across the different compound sets. In addition to the DrugBank set (DB) and the serum metabolites set (S), the gut compounds are shown in two subsets, glycerolipids (G-GL) and non-glycerolipids (G-NoGL), given the huge number of the former in this compound set, that would hide the distribution of the non-glycerolipids, and that their property ranges are widely different to the rest of the molecules.

From this figure we can see that while serum metabolites show distributions quite similar to those of drug molecules, in the case of gut metabolites there are clear separation from those of serum metabolites and drug molecules, especially in the glycerolipids subset, but also in the non-glycerolipids one. This similarity in the serum metabolites vs drugs chemical class distributions could be rationalized by the fact that many of the drugs for systemic use show similarity with endogenous metabolites that would be present in serum,^{34,35} as above described in the description of distributions of

chemical classes. On the contrary, the physicochemical distributions of the gut metabolites, both glycerolipids and non-glycerolipids, point towards a different chemical space with alternative properties.

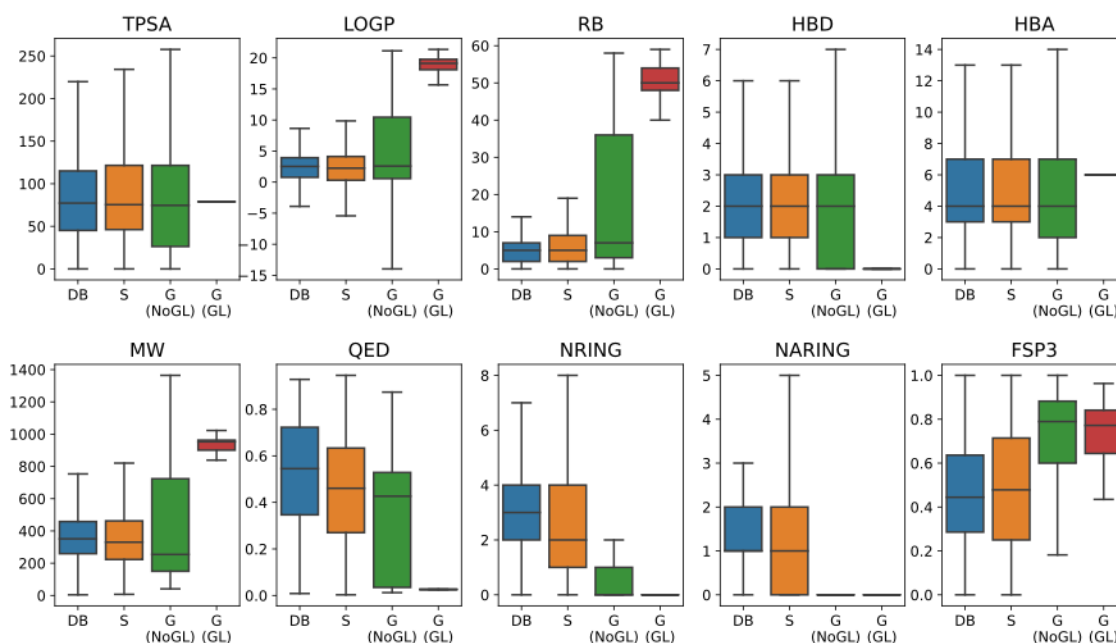


Figure 5. Distribution of different physicochemical properties for the different compound sets: DrugBank (DB); serum metabolites (S); gut metabolite, not glycerolipids subset (G (NoGL)); and gut metabolites, glycerolipids subset (G (GL)). Outliers are not displayed for clarity purposes.

In general, from visual inspection of the distributions it appears that as a general trend the gut compounds tend to be more lipophilic, with more rotatable bonds and higher fsp3, and higher molecular weight; in addition, they would have less hydrogen bond donors and rings, and lower QED. However, if we consider the chemical classes, again we observe more complicated patterns. Figure 6 shows the results of the statistical nonparametric analysis to test the differences in gut vs serum metabolites distribution for the compound sets as a whole (“all” column), as well as for particular chemical

classes. In turn, Figure S1 in Supporting Information collects violin plots and boxplots for the distribution of the different physicochemical properties in both serum and gut metabolites, and for both the whole compound sets as well as for the different chemical classes.

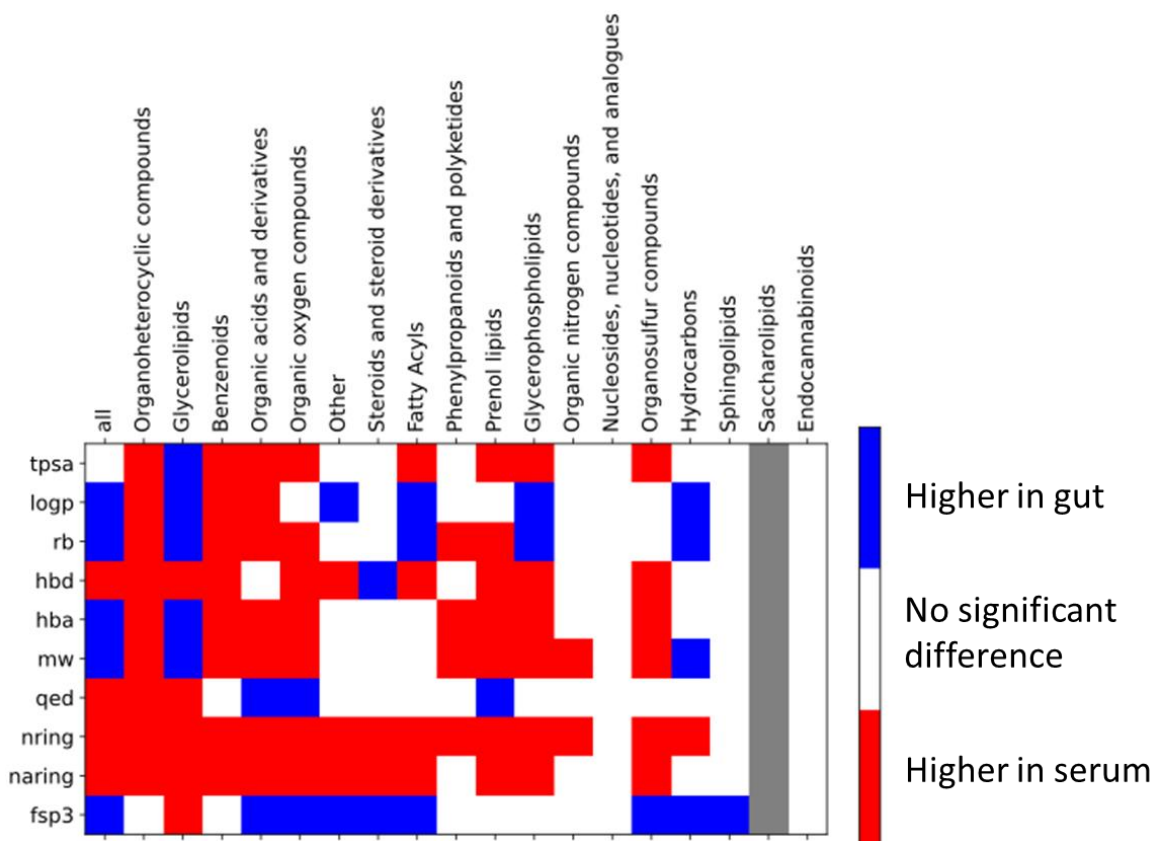


Figure 6. Physicochemical property enrichment for gut vs serum compound classes. For all the compounds (“all” column) or particular compound classes, a nonparametric Mann-Whitney test was conducted, and the Common-Language Effect Size statistic (CLES) calculated. Red cells correspond to significantly (p -value < 0.05) higher physicochemical property in the serum set (negative CLES), while blue cells correspond to significantly lower physicochemical property in the gut set (positive CLES). Grey

cells correspond to cells for which no test was feasible due to missing compounds in one or both of gut and serum metabolites.

By comparing the two complete compound sets ("all" column), it is possible to see as significant trends the higher logp, rb, hba, mw, and fsp3 of gut molecules, as well as their the lower hbd, qed, nring, and naring. No significant difference was observed in tpsa. We must consider, however, that the large majority of molecules are in the "glycerolipids" class in the case of the gut set, that would dominate the comparisons for the whole set. Therefore, if we look at the different chemical classes, we see a more convoluted picture, as in the previous sections, where the global pattern is more or less maintained in some of classes, while others seems to show a partially opposite pattern. Examples of the former are "glycerolipids", as expected, although in this case tpsa is significantly higher in the gut, and fsp3 is significantly lower; in addition, similarity to this global pattern is also observed more or less with "glycerophospholipids", "other", "fatty acyls", and "hydrocarbons". In these chemical classes, a statistically higher logp and rb are observed in gut metabolites, although the trend in hdb, hba, mw, and qed are not always conserved.

A different pattern is observed in "organoheterocyclic compounds", "benzenoids", "organic acids and derivatives", "organic oxygen compounds", "phenyl propanoids and polyketides", "prenol lipids", and "organosulfur compounds". Here the trend is reversed as logp, rb, hba, and mw are concerned (they tend to be significantly higher in serum metabolites), while it is kept as regarding nring, naring, and fsp3. In the case of qed, it is significantly higher in gut or no significant differences are observed.

A third pattern would correspond to compounds with little or no significant properties. These are “organic nitrogen compounds”, “nucleosides, nucleotides, and analogues”, “sphingolipids”, “endocannabinoids”, and can be explained by the small number of compounds in these classes. Finally, as a sort of “outlier” would remain the class of “steroids and steroid derivatives”, that has nring and naring significantly higher in serum metabolites, and fsp3 and hbd significantly higher in gut metabolites.

By looking at the data from the properties point of view, we observe several of them that are highly conserved across chemical classes. These are: nring and naring, that are significantly higher in serum metabolites or not significant, with no exception; fsp3, that is significantly higher in gut or not significant, with the single exception of “glycerolipids”, where it is significantly lower in gut; hbd, that is significantly higher in serum or not significant, with the sole exception of “steroids and steroid derivatives” as already mentioned, where is significantly higher in gut; hba, that is significantly higher in serum or not significant, with the exception of “glycerolipids”, where it is significantly higher in gut; and tpsa, significantly higher in serum or not significant, with the exception of “glycerolipids”, where it is significantly higher in gut.

A simple scoring system to predict gut permanence of small molecules

Lipinski’s rule of five was developed to predict oral bioavailability of molecules, such that those molecules fitting the rules (with the possible exception of at most one) would be more likely to be absorbable and permeable by the intestine.⁵² In principle, by reversing the rule we would expect to achieve a good approximation for the prediction of gut permanence. By doing that, with the full combined set of serum vs gut compounds we obtain an accuracy of 0.82, a precision of 0.60, and a recall of 0.84. Calculating the F_1

statistic we obtain a value of 0.70. However, if we analyze these statistics by chemical class, as shown in Table 2, we observe that while for “glycerolipids” the predictive power of Lipinski rule is excellent, with both precision and recall above 0.95, for the rest of chemical classes it is quite poor or right bad (precisions < 0.1 for 9 out of 13 chemical classes, and recalls < 0.1 for 8 classes). Thus, the global predictive metrics are misleading in that they are biased by the excellent “glycerolipids” prediction, since this group is the largest by far in the set of gut metabolites. If we instead obtain average predictive metrics conditional to chemical classes, we then obtain values 0.18 and 0.31 for the precision and recall, respectively ($F_1 = 0.23$). Similarly poor results are obtained by applying the reversed Veber’s rules,⁴⁰ other set of rules aiming at predicting bioavailability: a total of 8 precisions are < 0.1, while 3 recalls are < 0.1, with average conditional precisions and recalls of 0.18 and 0.4, respectively (Table 2), corresponding to a $F_1 = 0.25$.

	precision (Lipinski)	recall (Lipinski)	precision (Veber)	recall (Veber)	precision (score)	recall (score)	precision (ML)	recall (ML)	precision (score tst)	recall (score tst)	precision (ML tst)	recall (ML tst)
Organoheterocyclic compounds	0.02	0.09	0.02	0.15	0.04	0.69	0.17	0.05	0.05	0.67	0.5	0.11
Glycerolipids	0.92	0.99	0.92	1	0.98	0.98	1	0.99	0.98	0.98	1	1
Benzenoids	0	0	0.01	0.03	0.1	0.62	0.61	0.26	0.05	1	0.5	1
Organic acids and derivatives	0	0	0.03	0.11	0.22	0.34	0.79	0.5	0.21	0.25	0.81	0.61
Organic oxygen compounds	0.03	0.09	0.04	0.21	0.2	0.65	0.74	0.5	0.34	0.7	0.89	0.53
Other	0.01	0.54	0.01	0.54	0.06	0.42	1	0.54	0.22	0.5	1	0.25
Steroids and steroid derivatives	0.01	0.07	0.02	0.26	0.05	0.4	0.24	0.12	0.1	0.5	0	0
Fatty Acyls	0.47	0.43	0.18	0.66	0.31	0.86	0.92	0.8	0.31	0.96	0.88	0.81
Prenol lipids	0	0	0.01	0.05	0.32	0.88	0.81	0.62	0.22	0.67	1	0.67
Glycerophospholipids	0.47	0.9	0.48	0.98	0.88	0.66	0.95	0.92	0.95	0.59	0.97	0.88
Organosulfur compounds	0	0	0	0	0.33	0.44	0.88	0.67	0			
Hydrocarbons		0	0.48	0.23	0.73	0.69	0.8	0.75	0.25	0.5	0.33	0.5
Sphingolipids	0.2	0.89	0.18	1	0.29	0.09	0.7	0.26	1	1		0
Conditioned average	0.18	0.31	0.18	0.4	0.35	0.59	0.74	0.54	0.36	0.69	0.72	0.53

Table 2. Precision and recall of different gut permanence models for the different chemical classes: Lipinski-reversed (Lipinski), Veber-reversed (Veber), score method (score), machine learning method (ML). Here “Endocannabinoids”, “Nucleosides, nucleotides, and analogues”, “Organic nitrogen compounds”, and “Phenylpropanoids and polyketides” are not included due to their small number of compounds each. The metrics for Lipinski- and Veber-reverse methods are based on the whole compound set, while those of the score and machine learning methods are based on the average external prediction in 10-fold cross-validation of 80% of the full compound set. The columns with “tst”, for both score and machine learning models, corresponds to external predictions after the rederivation of the model with the whole training set (80% of molecules) and its application in the test set (20% of molecules); in this case, the test set had no gut molecules of the “Organosulfur compounds” chemical class, and that is the reason some metrics are missing.

In the previous sections we have been able to see that the comparative properties of gut vs serum metabolites are highly dependent on the chemical class. Thus, this factor must be taken into account when trying to predict gut permanence of a molecule in the intestine. In this way, a simple scoring system that includes it has been devised that provides improved predictive power over the whole range of chemical classes. In order to develop and test it, the whole dataset (19922 compounds) was randomly split into a 20% external test (2491 compounds) and 80% training set (17431 compounds), which in turn was used for model development and validation through 10-fold cross-validation. The scoring system is based on the set- and chemical class-specific interquartile ranges (IQR) of the physicochemical properties analyzed before: tpsa, logp, rb, hbd, hba, mw,

ged, nring, naring, and fsp3. Additional introduction of the ionization class in the scoring system did not improve the predictive power. In this scoring system, each molecule to predict is assigned the set (gut vs serum) with the largest number of properties for which the molecule is within its IQR; in the case of ties, the mean squared error (MSE) is calculated between the vector of 10 standardized properties of the molecule, and the vector of medians of the standardized properties of the corresponding chemical class, for both the gut and serum sets, and the set displaying the smaller MSE (the nearest set for that chemical class) is assigned to the molecule. In this way, the predictive metrics improve for the majority of the chemical classes, and indeed, in the 10-fold cross-validation only 3 chemical classes have precisions < 0.1 , and only one has a recall < 0.1 , Table 2; in turn, the average conditional precision and recall rise up to 0.35 and 0.59 (F_1 value of 0.44, nearly twice of the Lipinski-reversed model). The rederivation of the IQRs with the whole training set and its evaluation with the external test set yielded similar predictive performances, as expected (Table 2, “score tst” columns): 0.36 and 0.69 for the precision and recall, respectively ($F_1 = 0.47$).

This simple scoring system has the advantage of its easiness of interpretation, so that the medicinal chemist can perform a guided multiobjective, IQR-based optimization of the different properties of the molecule to make it gut-permanent. By increasing the number of properties within the IQR of the corresponding chemical class in the gut metabolites set, and likewise decreasing the number of properties in the IQR in the serum set, the likelihood of gut permanence will increase. In addition, it is flexible as it assumes the possibility of not all properties to be within the IQR of the chemical class in the gut set, as long as the number of properties in the IQR of the corresponding serum molecules remains smaller. Table S1 in Supporting Information provides the median and

IQRs for the 10 physicochemical properties, plus the two most frequent ionization classes, for all the chemical classes and the two compound sets, after rederivation with the full set of compounds. These values can help in guiding the medicinal chemist to decide chemical modifications appropriate to improve gut permanence.

Machine learning model to predict gut permanence

An additional predictive model was devised for cheminformatic settings, where a reliable ligand-based virtual screening tool would be preferred over interpretation, in order to generate gut permanence predictions for large numbers of molecules. Thus, a random forest was developed, using as input descriptors the 10 physicochemical properties, plus the ionization class and chemical class, both of them one-hot-encoded. A total of 1000 trees were seen in the cross-validation to give a good performance at reasonable computational cost. In Table 2, the predictive metrics are shown for both the 10-fold cross-validation, and the external test prediction ("ML" columns). From here, it is possible to see that the precision largely improves in almost all the chemical classes, while the recall slightly decreases in the majority of the classes, although it improves in the rest. Considering the average conditioned metrics, the precision and recall in cross-validation are of 0.74 and 0.54, respectively, (F_1 value of 0.62), and of 0.72 and 0.53 in the external test set (F_1 value of 0.61). Thus, the random forest provides an additional global improvement in the predictive power, especially as far as precision is concerned; in any case, the precision vs recall balance could be adjusted by using alternative probability thresholds in the model (right now the set assignment is based on a default threshold of 0.5).

Both the score model and the random forest models, together with the dataset, are provided as Python code freely accessible in <https://github.com/bbu-imdea/gutmetabos>

DISCUSSION

Gut-targeted drugs and nutraceuticals appear as a new drug modality that could exploit the new knowledge coming from the human gut microbiome research. The metabolite-target interactions identified through this research could be modulated by these new drugs and nutraceuticals, in order to provide novel curative and preventive approaches for health, in multiple areas such as inflammatory bowel disease, colon cancer, type 2 diabetes, obesity, non-alcoholic liver disease, diverticulitis, etc. In addition, directing the design of these compounds to remain in the gut would largely avoid the distribution, safety, and toxicology problems typical of systemic drugs, the main causes of the high attrition rate in this modality.⁵³

There are some few examples of drugs acting in the gut and with minimal or null bioavailability. Some of them act over host targets, in the metabolic diseases area; others over bacterial targets, being used as antibiotics; one antifungal, acting as a membrane-pore forming ionophore; and the rest of the molecules, acting on parasitic worm targets, as anthelmintic compounds. In terms of gut microbiome research, so far no commercial drug has been developed based on it, but the use of this research in drug discovery has already been pointed out,³⁰⁻³² and in fact some initial successful proof-of-concepts have allowed to find inhibitors of the pregnane X receptor based on gut metabolite mimics.⁵⁴ This has been followed by the development of the aryl hydrocarbon receptor, based on metabolite mimics too.^{55,56} In addition, in other work a combined bioinformatic/cheminformatic analysis based on data from the Human Microbiome Project¹⁵ has allowed to suggest several target-metabolite interactions that could be useful in drug discovery for inflammatory bowel disease.⁵⁷

Given all this background, the current work provides useful analyses that will help in the rational design of gut-targeted drugs based on (host or microbial) gut metabolites. We have compared a set of gut metabolites with a set of serum metabolites and a set of (essentially systemic) drugs in order to find differential patterns for the former that would guide the design of gut-targeted compounds. In this way, we have found that the distribution of chemical classes of gut metabolites is rather different to that of serum metabolites and drugs, while the two later sets have much more similar distributions. This similarity is in agreement with previous analyses that have stressed the structural similarity between systemic drugs and the corresponding metabolites,^{33,34} as well as the above mentioned proof-of-concept examples of inhibitors of gut targets based on intestinal microbial metabolites.⁵⁴⁻⁵⁶ This differential chemical class distribution has indeed been exploited in the case of two commercial drugs, orlistat and acarbose, which are in turn substrate analogs of two abundant sets of compounds in the gut, namely glycerolipids and oligosaccharides, rather unusual as source of systemic drugs.

The chemical class similarity between serum metabolites and drugs, and dissimilarity of these two with the gut metabolites, is confirmed when comparing the distributions of a large set of physicochemical properties (tpsa, logp, rb, hbd, hba, mw, nring, naring, qed, and fsp3). In addition, in terms of ionic class distribution, we observe also some serum metabolites vs drugs similarities, and at the same time dissimilarity with the gut metabolites set: besides the neutral class being the most abundant of the three, in the first two sets, zwitterion is the less frequent ionization class, while acid and basic classes are 2nd and 3rd of the serum metabolites set, and 3rd and 2nd of the DrugBank set; meanwhile, zwitterions is the 2nd most abundant class in the gut metabolites set, acid is the 3rd, and almost no basic compounds are present.

All in all, besides these general patterns, in all the analyses it seems necessary to consider the chemical class in order to account appropriately for the differential patterns specific for gut metabolites. Although gut metabolites seem in general to have higher logp, rb, and fsp3, and lower hbd, hba, qed, nring, and naring, to those of serum metabolites, there are some chemical classes where these trends are reversed: that is the case of e.g. “steroids and steroid derivatives” with hbd, “glycerolipids” with tpsa, hba, and fsp3, etc. Only nring and naring of the gut metabolites set is always significantly lower or not significant for all the chemical classes. Similar differential behavior depending on the chemical class is also observed when analyzing water solubility.

That chemical class is important in the rational design of gut-targeted compounds is confirmed when using the reversed version of rules of widespread use for the rational design of orally permeable, systemic drugs, namely Ro5³⁹ or Veber’s.⁴⁰ When applying these rules to this problem, a poor predictive power is observed for almost all chemical classes but “glycerolipids”. The inclusion of chemical class in our simple scoring system largely improved the predictive power for almost all chemical classes. By providing the IQRs for all chemical classes and physicochemical properties, for both the gut and serum metabolite sets, we hope to guide the rational design efforts of medicinal chemists, so that they can concentrate in optimizing the properties for which a particular compound is outside of the gut set IQR for the corresponding chemical class. We provide the IQR-based scoring system as freely available Python code that can be easily implemented, as well as the set of IQRs and two most frequent ionic classes in Table S1 in Supporting Information. The machine learning model, providing a reliable but black-box type of model, can alternatively be used in cheminformatic settings, like screening set designs,

compound prioritization, chemical space analysis, etc. This model is also freely provided through the corresponding Python source code.

We acknowledge some possible imperfections in our dataset, as the collection of gut and serum metabolites is based on multiple samples that can be obtained with different depths and with different backgrounds, and it is possible that for example, some compound of low but not null bioavailability, that in principle would be with more probability in the gut set, has only been observed in the serum set. Alternatively, it is possible that some highly bioavailable compound has only been observed in the gut set.

We have tried to minimize situations like this by removing the compounds that were both in the gut and serum sets, and this itself could translate in some biases in the chemical class distributions. We think, however, that this would correspond, if present, to a small proportion of compounds that otherwise would not change the qualitative and quantitative conclusions of this work, given the large number of compounds in both the gut and serum metabolite sets, compared to the shared metabolites.

In summary, we expect the current analyses and tools will provide new guides and technologies for the rational design of novel gut-targeted compounds, a new drug modality of high potential for the medicinal chemistry field.

MATERIALS AND METHODS

Data analysis was performed with Python 3.9, and using RDKit 2022.03.2 as cheminformatic toolkit. Metabolite structures and information were retrieved from the Human Metabolome Database (HMDB);⁴⁴ both gut and serum metabolites were retrieved. Only compounds with “detected and quantified” or “detected but not quantified” status were used. Drug structures and information were retrieved from the DrugBank⁵⁸, in particular, the subset of small molecules in approved, not-withdrawn, and non-illicit status. Molecular structures were processed and normalized with the ChEMBL Structure Pipeline⁵⁹ as described previously.^{45–47} A subset of 1735 metabolites was present in both gut and serum, and they were removed from the analysis, as the purpose of the analysis were to identify gut vs serum-specific features that would help in the design of new gut-targeted drugs. Similarly, 363 compounds were shared between DrugBank and serum metabolites, while 13 were shared between DrugBank and gut metabolites; in both cases, these compounds were assigned to the corresponding metabolite set. As a result of this retrieval and processing, the compound sets comprised 5021, 16621, and 1613 molecules, respectively for gut metabolites, serum metabolites, and DrugBank sets.

Ionization class assignment (acid, basic, neutral, and zwitterion) was based on HMDB’ strongest-acidic and strongest-basic pKa’s. Each molecule was assumed to have at least one acidic group if it had a strongest-acidic pKa < 7.4, and at least one basic group if it had a strongest-basic pKa > 7.4. Acid molecules were those with one or more acidic groups and no basic groups; basic molecules were those with one or more basic group and no acid group; neutral molecules were those with neither acidic nor basic groups,

and the rest of the molecules were zwitterions. In silico water solubility (logS) was retrieved from the HMDB.

Post-hoc analysis of contingency tables was based on adjusted residuals, and cell-specific p-values were calculated with an exact Fisher method recently described.⁶⁰ Differences between continuously distributed properties in two groups of molecules were tested through a non-parametric Mann-Whitney test, and direction of the effect was estimated through the Common-Language Effect Size (CLES)⁶¹ statistic, which estimates the probability than a random observation from the first group would be larger than a random observation from the second group.

Model derivation was based on a random-split-based training set comprising 80% of gut + serum metabolite sets, plus a 20% of external test set. In turn, model optimization was based on 10-fold cross-validation of the training set, and external predictive metrics were based on the average of the 10 folds. These were: accuracy, precision, recall, and F_1 . Once the model was optimized, it was rederived with the full training set, and tested in the external test set, not used during the cross-validation, and with the same prediction metrics.

Random forest models were based on the scikit-learn 1.1.2 module of Python. Default parameters were used, except the number of trees that in the cross-validation was selected to be 1000 as a reasonable compromise between performance and computation time.

Models and dataset can be obtained from <https://github.com/bbu-imdea/gutmetabos>

AUTHOR INFORMATION

Corresponding Author:

Gonzalo Colmenarejo - Biostatistics and Bioinformatics Unit. IMDEA Food, CEI UAM+CSIC, E28049 Madrid, Spain. orcid.org/0000-0002-8249-4547.
gonzalo.colmenarejo@imdea.org

Authors:

Alberto Gil-Pichardo, Andrés Sánchez-Ruiz - Biostatistics and Bioinformatics Unit. IMDEA Food, CEI UAM+CSIC, E28049 Madrid, Spain

Notes

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

Cristina González-Guevara, M.D., is thanked for her help in generating Table 1.

REFERENCES

- (1) Valeur, E.; Guéret, S. M.; Adihou, H.; Gopalakrishnan, R.; Lemurell, M.; Waldmann, H.; Grossmann, T. N.; Plowright, A. T. New Modalities for Challenging Targets in Drug Discovery. *Angewandte Chemie International Edition* **2017**, *56* (35), 10294–10323. <https://doi.org/10.1002/anie.201611914>.
- (2) Blanco, M.-J.; Gardinier, K. M. New Chemical Modalities and Strategic Thinking in Early Drug Discovery. *ACS Med. Chem. Lett.* **2020**, *11* (3), 228–231. <https://doi.org/10.1021/acsmchemlett.9b00582>.
- (3) Blanco, M.-J.; Gardinier, K. M.; Namchuk, M. N. Advancing New Chemical Modalities into Clinical Studies. *ACS Med. Chem. Lett.* **2022**, *13* (11), 1691–1698. <https://doi.org/10.1021/acsmchemlett.2c00375>.
- (4) Kulkarni, J. A.; Witzigmann, D.; Thomson, S. B.; Chen, S.; Leavitt, B. R.; Cullis, P. R.; van der Meel, R. The Current Landscape of Nucleic Acid Therapeutics. *Nat. Nanotechnol.* **2021**, *16* (6), 630–643. <https://doi.org/10.1038/s41565-021-00898-0>.
- (5) Roberts, T. C.; Langer, R.; Wood, M. J. A. Advances in Oligonucleotide Drug Delivery. *Nat Rev Drug Discov* **2020**, *19* (10), 673–694. <https://doi.org/10.1038/s41573-020-0075-7>.
- (6) Caron, G.; Digiesi, V.; Solaro, S.; Ermondi, G. Flexibility in Early Drug Discovery: Focus on the beyond-Rule-of-5 Chemical Space. *Drug Discovery Today* **2020**, *25* (4), 621–627. <https://doi.org/10.1016/j.drudis.2020.01.012>.
- (7) Matsson, P.; Doak, B. C.; Over, B.; Kihlberg, J. Cell Permeability beyond the Rule of 5. *Advanced Drug Delivery Reviews* **2016**, *101*, 42–61. <https://doi.org/10.1016/j.addr.2016.03.013>.
- (8) Békés, M.; Langley, D. R.; Crews, C. M. PROTAC Targeted Protein Degraders: The Past Is Prologue. *Nat Rev Drug Discov* **2022**, *21* (3), 181–200. <https://doi.org/10.1038/s41573-021-00371-6>.
- (9) Muttenthaler, M.; King, G. F.; Adams, D. J.; Alewood, P. F. Trends in Peptide Drug Discovery. *Nat Rev Drug Discov* **2021**, *20* (4), 309–325. <https://doi.org/10.1038/s41573-020-00135-8>.
- (10) Leader, B.; Baca, Q. J.; Golan, D. E. Protein Therapeutics: A Summary and Pharmacological Classification. *Nat Rev Drug Discov* **2008**, *7* (1), 21–39. <https://doi.org/10.1038/nrd2399>.
- (11) Anselmo, A. C.; Gokarn, Y.; Mitragotri, S. Non-Invasive Delivery Strategies for Biologics. *Nat Rev Drug Discov* **2019**, *18* (1), 19–40. <https://doi.org/10.1038/nrd.2018.183>.
- (12) Kinch, M. S. An Overview of FDA-Approved Biologics Medicines. *Drug Discovery Today* **2015**, *20* (4), 393–398. <https://doi.org/10.1016/j.drudis.2014.09.003>.
- (13) Turnbaugh, P. J.; Ley, R. E.; Hamady, M.; Fraser-Liggett, C. M.; Knight, R.; Gordon, J. I. The Human Microbiome Project. *Nature* **2007**, *449* (7164), 804–810. <https://doi.org/10.1038/nature06244>.
- (14) Almeida, A.; Mitchell, A. L.; Boland, M.; Forster, S. C.; Gloor, G. B.; Tarkowska, A.; Lawley, T. D.; Finn, R. D. A New Genomic Blueprint of the Human Gut Microbiota. *Nature* **2019**, *568* (7753), 499–504. <https://doi.org/10.1038/s41586-019-0965-1>.
- (15) Proctor, L. M.; Creasy, H. H.; Fettweis, J. M.; Lloyd-Price, J.; Mahurkar, A.; Zhou, W.; Buck, G. A.; Snyder, M. P.; Strauss, J. F.; Weinstock, G. M.; White, O.; Huttenhower, C.; The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature* **2019**, *569* (7758), 641–648. <https://doi.org/10.1038/s41586-019-1238-8>.
- (16) Gilbert, J. A.; Blaser, M. J.; Caporaso, J. G.; Jansson, J. K.; Lynch, S. V.; Knight, R. Current Understanding of the Human Microbiome. *Nat Med* **2018**, *24* (4), 392–400. <https://doi.org/10.1038/nm.4517>.
- (17) Fan, Y.; Pedersen, O. Gut Microbiota in Human Metabolic Health and Disease. *Nat Rev Microbiol* **2021**, *19* (1), 55–71. <https://doi.org/10.1038/s41579-020-0433-9>.

- (18) Feng, Q.; Liang, S.; Jia, H.; Stadlmayr, A.; Tang, L.; Lan, Z.; Zhang, D.; Xia, H.; Xu, X.; Jie, Z.; Su, L.; Li, X.; Li, X.; Li, J.; Xiao, L.; Huber-Schönauer, U.; Niederseer, D.; Xu, X.; Al-Aama, J. Y.; Yang, H.; Wang, J.; Kristiansen, K.; Arumugam, M.; Tilg, H.; Datz, C.; Wang, J. Gut Microbiome Development along the Colorectal Adenoma–Carcinoma Sequence. *Nature Communications* **2015**, *6* (1). <https://doi.org/10.1038/ncomms7528>.
- (19) Javdan, B.; Lopez, J. G.; Chankhamjon, P.; Lee, Y.-C. J.; Hull, R.; Wu, Q.; Wang, X.; Chatterjee, S.; Donia, M. S. Personalized Mapping of Drug Metabolism by the Human Gut Microbiome. *Cell* **2020**, *181* (7), 1661–1679.e22. <https://doi.org/10.1016/j.cell.2020.05.001>.
- (20) Jeganathan, N. A.; Davenport, E. R.; Yochum, G. S.; Koltun, W. A. The Microbiome of Diverticulitis. *Current Opinion in Physiology* **2021**, *22*, 100452. <https://doi.org/10.1016/j.cophys.2021.06.006>.
- (21) Lavelle, A.; Sokol, H. Gut Microbiota-Derived Metabolites as Key Actors in Inflammatory Bowel Disease. *Nat Rev Gastroenterol Hepatol* **2020**, *17* (4), 223–237. <https://doi.org/10.1038/s41575-019-0258-z>.
- (22) Lee, W.-J.; Hase, K. Gut Microbiota–Generated Metabolites in Animal Health and Disease. *Nat Chem Biol* **2014**, *10* (6), 416–424. <https://doi.org/10.1038/nchembio.1535>.
- (23) Olson, C. A.; Vuong, H. E.; Yano, J. M.; Liang, Q. Y.; Nusbaum, D. J.; Hsiao, E. Y. The Gut Microbiota Mediates the Anti-Seizure Effects of the Ketogenic Diet. *Cell* **2018**, *173* (7), 1728–1741.e13. <https://doi.org/10.1016/j.cell.2018.04.027>.
- (24) Funabashi, M.; Grove, T. L.; Wang, M.; Varma, Y.; McFadden, M. E.; Brown, L. C.; Guo, C.; Higginbottom, S.; Almo, S. C.; Fischbach, M. A. A Metabolic Pathway for Bile Acid Dehydroxylation by the Gut Microbiome. *Nature* **2020**, *582* (7813), 566–570. <https://doi.org/10.1038/s41586-020-2396-4>.
- (25) Donia, M. S.; Fischbach, M. A. Small Molecules from the Human Microbiota. *Science* **2015**, *349* (6246). <https://doi.org/10.1126/science.1254766>.
- (26) Henke, M. T.; Clardy, J. Molecular Messages in Human Microbiota. *Science* **2019**, *366* (6471), 1309–1310. <https://doi.org/10.1126/science.aaz4164>.
- (27) Quinn, R. A.; Melnik, A. V.; Vrbanc, A.; Fu, T.; Patras, K. A.; Christy, M. P.; Bodai, Z.; Belda-Ferre, P.; Tripathi, A.; Chung, L. K.; Downes, M.; Welch, R. D.; Quinn, M.; Humphrey, G.; Panitchpakdi, M.; Weldon, K. C.; Aksenov, A.; da Silva, R.; Avila-Pacheco, J.; Clish, C.; Bae, S.; Mallick, H.; Franzosa, E. A.; Lloyd-Price, J.; Bussell, R.; Thron, T.; Nelson, A. T.; Wang, M.; Leszczynski, E.; Vargas, F.; Gauglitz, J. M.; Meehan, M. J.; Gentry, E.; Arthur, T. D.; Komor, A. C.; Poulsen, O.; Boland, B. S.; Chang, J. T.; Sandborn, W. J.; Lim, M.; Garg, N.; Lumeng, J. C.; Xavier, R. J.; Kazmierczak, B. I.; Jain, R.; Egan, M.; Rhee, K. E.; Ferguson, D.; Raffatellu, M.; Vlamakis, H.; Haddad, G. G.; Siegel, D.; Huttenhower, C.; Mazmanian, S. K.; Evans, R. M.; Nizet, V.; Knight, R.; Dorrestein, P. C. Global Chemical Effects of the Microbiome Include New Bile-Acid Conjugations. *Nature* **2020**, *579* (7797), 123–129. <https://doi.org/10.1038/s41586-020-2047-9>.
- (28) Lavelle, A.; Sokol, H. Gut Microbiota-Derived Metabolites as Key Actors in Inflammatory Bowel Disease. *Nat Rev Gastroenterol Hepatol* **2020**, *17* (4), 223–237. <https://doi.org/10.1038/s41575-019-0258-z>.
- (29) Silpe, J. E.; Balskus, E. P. Deciphering Human Microbiota–Host Chemical Interactions. *ACS Cent. Sci.* **2021**, *7* (1), 20–29. <https://doi.org/10.1021/acscentsci.0c01030>.
- (30) Saha, S.; Rajpal, D. K.; Brown, J. R. Human Microbial Metabolites as a Source of New Drugs. *Drug Discovery Today* **2016**, *21* (4), 692–698. <https://doi.org/10.1016/j.drudis.2016.02.009>.
- (31) Chavira, A.; Belda-Ferre, P.; Kosciolk, T.; Ali, F.; Dorrestein, P. C.; Knight, R. The Microbiome and Its Potential for Pharmacology. In *Concepts and Principles of Pharmacology: 100 Years of the Handbook of Experimental Pharmacology*; Barrett, J. E., Page, C. P., Michel, M. C., Eds.; Handbook of Experimental Pharmacology; Springer

- International Publishing: Cham, 2019; pp 301–326.
https://doi.org/10.1007/164_2019_317.
- (32) Nuzzo, A.; Brown, J. R. Microbiome Metabolite Mimics Accelerate Drug Discovery. *Trends in Molecular Medicine* **2020**, *26* (5), 435–437.
<https://doi.org/10.1016/j.molmed.2020.03.006>.
- (33) Dobson, P. D.; Patel, Y.; Kell, D. B. ‘Metabolite-Likeness’ as a Criterion in the Design and Selection of Pharmaceutical Drug Libraries. *Drug Discovery Today* **2009**, *14* (1–2), 31–40.
<https://doi.org/10.1016/j.drudis.2008.10.011>.
- (34) O’Hagan, S.; Swainston, N.; Handl, J.; Kell, D. B. A ‘Rule of 0.5’ for the Metabolite-Likeness of Approved Pharmaceutical Drugs. *Metabolomics* **2015**, *11* (2), 323–339.
<https://doi.org/10.1007/s11306-014-0733-z>.
- (35) O’Hagan, S.; Kell, D. B. Analysis of Drug–Endogenous Human Metabolite Similarities in Terms of Their Maximum Common Substructures. *J Cheminform* **2017**, *9* (1), 18.
<https://doi.org/10.1186/s13321-017-0198-y>.
- (36) Bofill, A.; Jalencas, X.; Oprea, T. I.; Mestres, J. The Human Endogenous Metabolome as a Pharmacology Baseline for Drug Discovery. *Drug Discovery Today* **2019**, *24* (9), 1806–1820. <https://doi.org/10.1016/j.drudis.2019.06.007>.
- (37) Dobson, P. D.; Kell, D. B. Carrier-Mediated Cellular Uptake of Pharmaceutical Drugs: An Exception or the Rule? *Nat Rev Drug Discov* **2008**, *7* (3), 205–220.
<https://doi.org/10.1038/nrd2438>.
- (38) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. PII of Original Article: S0169-409X(96)00423-1. The Article Was Originally Published in *Advanced Drug Delivery Reviews* *23* (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001**, *46* (1), 3–26. [https://doi.org/10.1016/S0169-409X\(00\)00129-0](https://doi.org/10.1016/S0169-409X(00)00129-0).
- (39) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technologies* **2004**, *1* (4), 337–341.
<https://doi.org/10.1016/j.ddtec.2004.11.007>.
- (40) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623. <https://doi.org/10.1021/jm020017n>.
- (41) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1* (1), 55–68. <https://doi.org/10.1021/cc9800071>.
- (42) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51* (4), 817–834. <https://doi.org/10.1021/jm701122q>.
- (43) Billat, P.-A.; Roger, E.; Faure, S.; Lagarce, F. Models for Drug Absorption from the Small Intestine: Where Are We and Where Are We Going? *Drug Discovery Today* **2017**, *22* (5), 761–775. <https://doi.org/10.1016/j.drudis.2017.01.007>.
- (44) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, H. B.; Greiner, R.; Gautam, V. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Research* **2022**, *50* (D1), D622–D631. <https://doi.org/10.1093/nar/gkab1062>.
- (45) Kaya, I.; Colmenarejo, G. Analysis of Nuisance Substructures and Aggregators in a Comprehensive Database of Food Chemical Compounds. *J. Agric. Food Chem.* **2020**, *68* (33), 8812–8824. <https://doi.org/10.1021/acs.jafc.0c02521>.

- (46) Sánchez-Ruiz, A.; Colmenarejo, G. Updated Prediction of Aggregators and Assay-Interfering Substructures in Food Compounds. *J. Agric. Food Chem.* **2021**, *69* (50), 15184–15194. <https://doi.org/10.1021/acs.jafc.1c05918>.
- (47) Sánchez-Ruiz, A.; Colmenarejo, G. Systematic Analysis and Prediction of the Target Space of Bioactive Food Compounds: Filling the Chemobiological Gaps. *J. Chem. Inf. Model.* **2022**, *62* (16), 3734–3751. <https://doi.org/10.1021/acs.jcim.2c00888>.
- (48) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *Journal of Cheminformatics* **2016**, *8* (1), 61. <https://doi.org/10.1186/s13321-016-0174-Y>.
- (49) Sitrin, M. D. Digestion and Absorption of Dietary Triglycerides. In *The Gastrointestinal System: Gastrointestinal, Nutritional and Hepatobiliary Physiology*; Leung, P. S., Ed.; Springer Netherlands: Dordrecht, 2014; pp 159–178. https://doi.org/10.1007/978-94-017-8771-0_7.
- (50) *The lipase inhibitor tetrahydrolipstatin binds covalently to the putative active site serine of pancreatic lipase.* | Elsevier Enhanced Reader. [https://doi.org/10.1016/S0021-9258\(18\)52203-1](https://doi.org/10.1016/S0021-9258(18)52203-1).
- (51) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nature Chem* **2012**, *4* (2), 90–98. <https://doi.org/10.1038/nchem.1243>.
- (52) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Advanced Drug Delivery Reviews* **1997**, *23* (1), 3–25. [https://doi.org/10.1016/S0169-409X\(96\)00423-1](https://doi.org/10.1016/S0169-409X(96)00423-1).
- (53) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nat Rev Drug Discov* **2015**, *14* (7), 475–486. <https://doi.org/10.1038/nrd4609>.
- (54) Dvořák, Z.; Kopp, F.; Costello, C. M.; Kemp, J. S.; Li, H.; Vrzalová, A.; Štěpánková, M.; Bartoňková, I.; Jískrová, E.; Poulíková, K.; Vyhřídálová, B.; Nordstroem, L. U.; Karunaratne, C. V.; Ranhotra, H. S.; Mun, K. S.; Naren, A. P.; Murray, I. A.; Perdew, G. H.; Brtko, J.; Toporova, L.; Schön, A.; Wallace, B. D.; Walton, W. G.; Redinbo, M. R.; Sun, K.; Beck, A.; Kortagere, S.; Neary, M. C.; Chandran, A.; Vishveshwara, S.; Cavalluzzi, M. M.; Lentini, G.; Cui, J. Y.; Gu, H.; March, J. C.; Chatterjee, S.; Matson, A.; Wright, D.; Flannigan, K. L.; Hirota, S. A.; Sartor, R. B.; Mani, S. Targeting the Pregnane X Receptor Using Microbial Metabolite Mimicry. *EMBO Molecular Medicine* **2020**, *12* (4), e11621. <https://doi.org/10.15252/emmm.201911621>.
- (55) Grycová, A.; Joo, H.; Maier, V.; Illés, P.; Vyhřídálová, B.; Poulíková, K.; Sládeková, L.; Nádvorník, P.; Vrzal, R.; Zemánková, L.; Pečínková, P.; Poruba, M.; Zapletalová, I.; Večeřa, R.; Anzenbacher, P.; Ehrmann, J.; Ondra, P.; Jung, J.-W.; Mani, S.; Dvořák, Z. Targeting the Aryl Hydrocarbon Receptor with Microbial Metabolite Mimics Alleviates Experimental Colitis in Mice. *J. Med. Chem.* **2022**, *65* (9), 6859–6868. <https://doi.org/10.1021/acs.jmedchem.2c00208>.
- (56) Dvořák, Z.; Li, H.; Mani, S. Microbial Metabolites as Ligands to Xenobiotic Receptors: Chemical Mimicry as Potential Drugs of the Future. *Drug Metab Dispos* **2023**, *51* (2), 219–227. <https://doi.org/10.1124/dmd.122.000860>.
- (57) Nuzzo, A.; Saha, S.; Berg, E.; Jayawickreme, C.; Tocker, J.; Brown, J. R. Expanding the Drug Discovery Space with Predicted Metabolite–Target Interactions. *Communications Biology* **2021**, *4* (1), 1–11. <https://doi.org/10.1038/s42003-021-01822-x>.
- (58) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Lynkaran, I.; Liu, Y.; Maciejewski, A.;

- Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Research* **2018**, *46* (D1), D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>.
- (59) Bento, A. P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; De Veij, M.; Leach, A. R. An Open Source Chemical Structure Curation Pipeline Using RDKit. *Journal of Cheminformatics* **2020**, *12* (1), 51. <https://doi.org/10.1186/s13321-020-00456-1>.
- (60) Shan, G.; Gerstenberger, S. Fisher's Exact Approach for Post Hoc Analysis of a Chi-Squared Test. *PLoS One* **2017**, *12* (12), e0188709. <https://doi.org/10.1371/journal.pone.0188709>.
- (61) McGraw, K. O.; Wong, S. P. A Common Language Effect Size Statistic. *Psychological Bulletin* **1992**, *111*, 361–365. <https://doi.org/10.1037/0033-2909.111.2.361>.

FUNDING SOURCES ACKNOWLEDGEMENT

Grant PID2021-127318OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”

AS-R acknowledges the Consejería de Ciencia, Universidades e Innovación de la Comunidad de Madrid, Spain (Ref. PEJ-2020-AI/BIO-17904), for a research assistant contract.

SUPPORTING INFORMATION

TableS1.xlsx:

Table S1: Distribution of physicochemical properties for each chemical class in gut and serum metabolites. For each combination, the median and IQR is displayed for both gut and serum metabolites, plus the CES, and p-value. For ionization classes, the first and second most abundant are displayed for both gut and serum metabolites.

Figures S1-S10.pdf:

Figures S1-S10: Distribution of physicochemical properties for the whole set and for each chemical class of gut and serum metabolites. Both violin plots and boxplots (without outliers, for clarity purposes) are displayed.

TABLE OF CONTENTS GRAPHIC

