How Do Microbial Metabolites Interact with Their Protein Targets?

Mario Astigarraga, Andrés Sánchez-Ruiz, Aminata Diop-Aw, Raquel Quintero, and Gonzalo Colmenarejo*

Biostatistics and Bioinformatics Unit

IMDEA Food

CEI UAM+CSIC

E28049 Madrid, Spain

*Corresponding Author

e-mail: gonzalo.colmenarejo@imdea.org

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

ABSTRACT

The design of drugs and nutraceutics that mimic microbial metabolites is an emerging drug modality in medicinal chemistry that attempts to modulate the myriad of interactions that these molecules establish with host and microbial proteins. Understanding how microbial metabolites interact with their target proteins is key to perform a rational design of metabolite mimetic molecules for therapeutic usage. In the present work we answer that question by analyzing the functional groups of these molecules, and the interactions they display in a set of more than 71 K protein-metabolite interactions from the PDB. Significant differences in the functional group distributions, their chemical features, and their co-occurrences, are observed for distinct subsets of these molecules. The same is true for the distributions of interaction types. By correlating both datasets, we are able to explain the observed interaction patterns in terms of observed functional group patterns. These results will shed light on the rational design of novel metabolite mimetic molecules for therapeutic purposes.

KEYWORDS:

Microbial metabolite mimetic, microbiome, gut metabolome, new drug modalities, drug design, nutraceutic design, drug target, functional group, cheminformatics, protein-ligand interactions.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

INTRODUCTION

The importance of the microbiome in human health is becoming increasingly relevant in recent years.[1–4] The human body hosts trillions of microbial cells, especially in the gut, and many biological processes, including metabolism, infection, immunity, and even the nervous system, are influenced by the signaling systems established between the host cells and the cells in the microbiota.[3,5–12] This bidirectional communication is exerted through small molecules (metabolites) that bind to proteins, thus triggering a biological response.[13–29] This fact is being exploited to develop new drugs based on microbial metabolites (gut microbial "metabolite-mimetic" drugs), in widely different areas like cardiovascular diseases, Parkinson, Alzheimer disease, infectious diseases, and cancer, now with some compounds in clinical phases.[30–37] These molecules are able to modulate altered metabolite-target interactions in disease states, in order to rebalance them. The use of microbial metabolite mimetic molecules, besides allowing to identify novel drug chemotypes, would expand the target space in about two orders of magnitude, the size of the metabolome.[31–34,38]

On the other hand, rational drug design is based on a deep knowledge of the interactions of small molecules with protein targets. In this regard, the concept of functional groups (FGs),[39] that is, sets of connected atoms in the ligand with specific physicochemical and reactivity properties, is of great utility as it simplifies the analysis of protein-ligand interactions. Small molecules use their FGs to establish non-covalent interactions with proteins of different types: hydrogen bonds, hydrophobic, pi-pi,

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

cation-pi, etc. Data analysis of large quantities of experimental FG-protein interactions at atomic resolution is immensely useful for the rational design of drugs. Therefore, to rationally develop a new metabolite-mimetic drug, we first must know what interactions these molecules establish with host and bacterial proteins, and through what FGs. These interactions can putatively depend on constraints such as the localization, source, and chemical class of the compound. In turn, these interactions, together with molecular shape, define their target specificity.

Our group is currently investigating microbial metabolites as a novel source of drugs and nutraceutics using Cheminformatics and Data Science approaches. In previous works, we have analyzed and modeled their physicochemical, structural, and biodistribution properties.[40] We have also thoroughly studied the full set of published interactions of these molecules with human and bacterial proteins and provided thousands of validated predictions of new interactions using ligand-based virtual screening.[41]

In this work, we aim to expand this research to analyze the FGs present in gut microbial metabolites, and the interactions these FGs display with human and gut microbial proteins in experimentally determined protein-metabolite complexes at atomic resolution. The patterns that we find here will shed light on the rational design of novel drugs and nutraceutics based on gut metabolites, which will be useful in medicinal chemistry efforts oriented to mimetizing microbial metabolites.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

METHODS

All data analysis was performed with Python 3.12.5, with RDKit 2024.03.5 as cheminformatic toolkit.[42] The Bio.PDB module[43] of Biopython 1.84 was used to manipulate protein-ligand structures programmatically. We used the same dataset of gut metabolites as in our previous works,[40,41] that is, the subset of "detected and quantified" and "detected but not quantified" gut molecules in the Human Metabolome Database,[44] plus some recent additions from the literature.[13,14] This set of molecules, that comprises a total of 6663 molecules, span three rather different compound sets: the "GutFL" set (abbreviated "GFL"), that includes all the metabolites (5451) of fatty acid-derived chemical classes, namely Glycerolipids, Glycerophospholipids, Fatty Acyls, and Sphingolipids; the "Gut/Serum" set (abbreviated "G/S"), that includes 548 molecules detected in both gut and serum; and the "GutnoFL" set (abbreviated "GnFL"), of 664 compounds, that correspond to molecules only detected in gut but not fatty acid-derived. For comparison purposes, a set of 1421 drugs was used, corresponding to all small molecules in approved, not-withdrawn, not illicit status, and orally administered, and present in the DrugBank;[45] most of them act systemically (1411), although a few are not absorbed and remain in the intestine where they act locally (11 molecules). All these molecules are assigned one chemical class derived from the ClassyFire algorithm.[46] See our previous papers for a full description of the preparation of the molecules.[40,41,47]

In order to generate the functional groups (FGs) we used a complete and accurate implementation of the Ertl algorithm[48] in RDKit developed by our group. To analyze the

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

protein-ligand interactions we used the BINANA 2 Python module.[49] BINANA 2 is a Python module that is designed to identify programmatically protein-ligand interactions of different types: close contacts (with protein atoms less than 4 Å distant), closest contacts (less than 2.5 Å), hydrophobics, hydrogen bonds, halogen bonds, pi-pi, cation-pi, salt bridges, and metal coordination. It requires that both the protein and the ligand be in PDBQT format, so for that aim we developed a pipeline to prepare these files. First, for each compound, all the protein-ligand complex structures present in the Protein Data Bank (PDB) were retrieved, based on its InChi string and using the PDB API.[50,51] NMR structures were disregarded, as well as those where the ligand was covalently bound to the protein, and with proteins more than 2500 amino acids long. Only proteins from *Homo sapiens* and from a set of prokaryotic genera typical of metagenomic analyses were used, as in our previous work.[40] In addition, only ligands with a fraction of buried solvent accessible surface > 0.2 compared to the free ligand were included in the analysis. Protein structures were fixed with PDBFixer,[52,53] protonated with Moleculekit at pH = 7.4,[54] and finally Gasteiger charges and atom types were added with the *prepare_receptor* script of the ADFR suite[55] in order to generate a protein PDBQT file. Similarly, the ligand structures were processed with the *prepare_ligand* script in the same suite.

To map functional groups into interactions, a match was identified every time at least one of the atoms in a functional group was in the set of ligand atoms of the corresponding interaction in BINANA 2. To be able to do that match, it was previously necessary to set up and use a function in RDKit to “fix” the PDB molecule, as in many

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

cases the bonding patterns and valences had assignment errors, and the ligands in PDB have no hydrogens. The function generated a molecule with the same connectivity, bond types, and explicit / implicit hydrogens as the original molecule used in the analysis of FGs, but with the 3D structure of the PDB molecule in its carbons and heteroatoms. In addition, the function generated a mapping between the index of the different atoms with the two versions of the molecule (one obtained from an InChi string after standardization, and another obtained from the PDB file), to be able to identify intersections between FG atoms and BINANA 2 interactions.

The following interaction types were considered: closest contacts, hydrophobics, hydrogen bonds, halogen bonds, pi-pi, cation-pi, salt bridges, and metal coordination. The “close interaction” type in BINANA 2 was not considered as nearly all binding atoms had it and therefore lacked any discrimination power among the compound sets. Interactions were aggregated by ligand to identify patterns in the interaction types of the different compound sets and functional groups. This aggregation was used to avoid the biasing of the distributions by the very differential presence of some of the ligands in the PDB structures.

Statistical analysis of numeric distributions used Kruskal-Walli’s test to find significantly different distributions. When that was the case, a Conover-type post-hoc analysis (with Holm multiple test correction of p-values) was run to find significant pairs of distributions that could account for the significance of the omnibus test. In addition, the Common Language Effect Size (CLES) statistic[56] between significant pairs

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

was computed to find which of the two distributions was shifted to higher values. The CLES estimates the probability that an instance randomly selected from one of the two distributions will be greater than an instance randomly selected from the other. Statistical tests were applied with a significance level of 0.05.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

RESULTS

In this work we aim to characterize, in a first phase, the FG distributions typical of gut microbial metabolites, as well as the commonalities and differences of these with those of oral drugs. As described in Methods, we use three different compound sets for microbial metabolites: G/S (molecules detected both in gut and in serum), GFL (molecules derived from fatty lipids, not able to cross the gut wall), and GnFL (molecules only detected in gut, but not derived from fatty lipids). The set of oral drugs (DB) is used for comparison purposes. Thus, commonalities and differences among these sets is also studied. To obtain the FGs, Ertl's algorithm is applied, as it does not use a predefined set of FGs but instead is able to extract FGs from any arbitrary molecular structure. In the second phase, we pursue a comprehensive characterization of the interactions between these molecules and proteins through a systematic analysis of the experimental structures available in the PDB. Again, the identification of specific patterns in these molecules is done comparatively with oral drugs, and among the three metabolite chemical sets. Finally, the two sources of data, FGs and interactions, are correlated in order to explain the observed interaction patterns in the compound sets in terms of the interactions the different FG establish. These three aspects will be described in sequence in what follows.

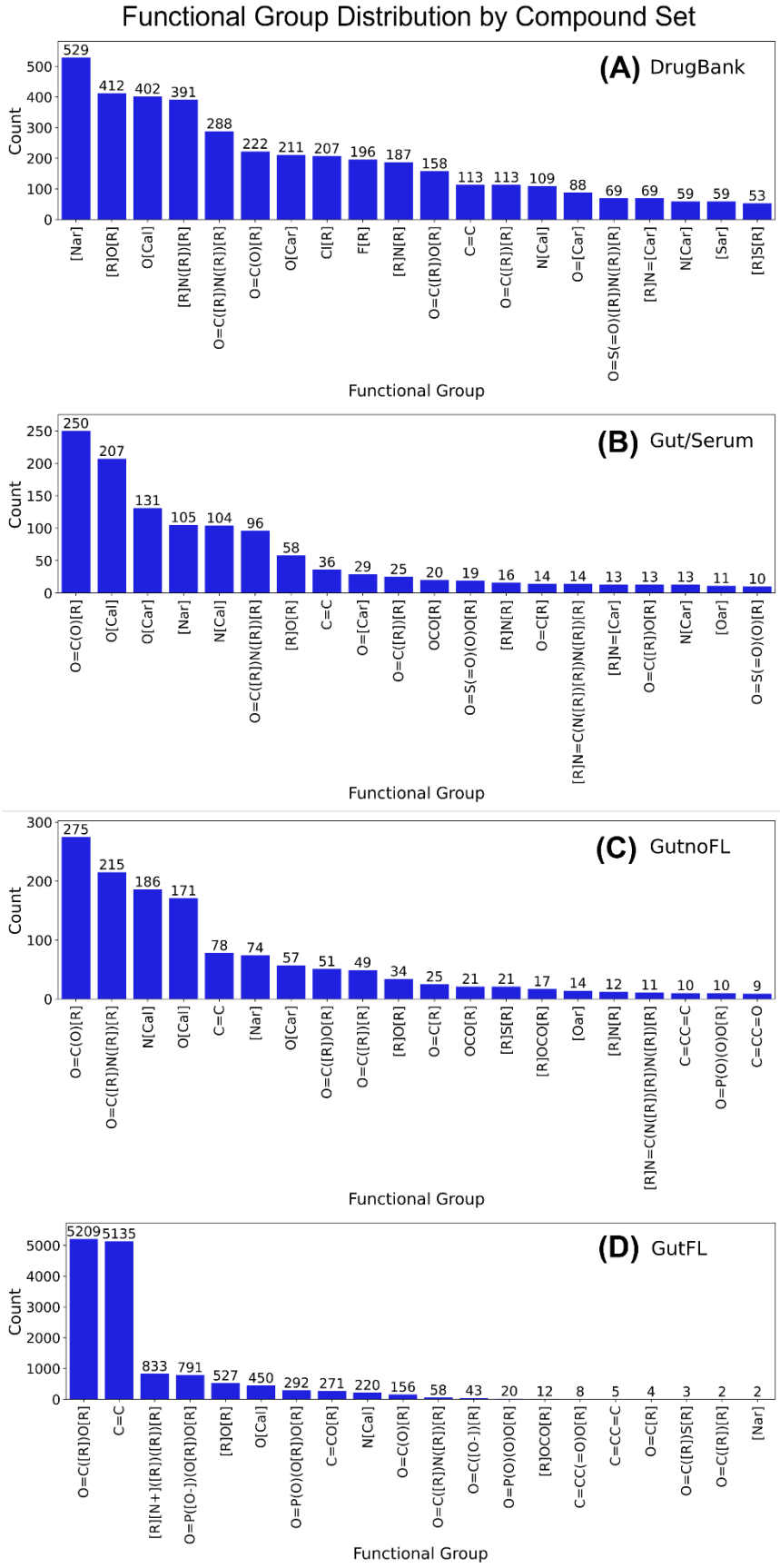
FG analysis

Among all the four compound sets, a total of 294 unique FGs are identified. Figure 1 displays the distribution of the top 20 FGs in decreasing order of all the compound sets studied: DB,

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

G/S, GnFL, GFL. In addition, Table 1 shows different statistics obtained from these distributions.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Figure 1. *Distribution (in decreasing order) of the top 20 FGs in the different compound sets analyzed in this work.*

	n	% mols w/o FG	# FGs (total)	FG (total) / mol	# FG (bin)	FG (bin) / mol	# FG (un)	FG (un) / mol
DB	1421	0	7185	5.056	4913	3.457	253	0.178
G/S	548	1.095	1885	3.44	1310	2.391	69	0.126
GnFL	664	5.12	2058	3.099	1473	2.218	78	0.117
GFL	5451	0	45291	8.309	14061	2.58	37	0.007

Table 1. *Count statistics of FGs in the different compound sets. n = number of molecules; % mols w/o FG = percentage of molecules lacking any FG; # FG (total) = total count of FG in all the molecules of the set; FG (total) / mol = the previous count normalized by the number of molecules in the set; # FG (bin) = binary count of FG in all the molecules in the set, i.e. only considering presence / absence of each FG; FG (bin) / mol = the previous count normalized by the number of molecules in the set; # FG (un) = count of unique FGs in the set; FG (un) / mol = the previous count normalized by the number of molecules in the set.*

It is possible to observe clear differences among the four compound sets. DB has the largest diversity of FGs, with the biggest number of unique FGs, 253, resulting in an average of 0.18 unique FGs per molecule, the largest one. It has the second largest number of FGs per molecule (5.06 total, 3.46 if only binary presence / absence of each FG is counted within a molecule), and no molecule in this compound set lacks FGs. The distribution (Figure 1A) shows a slow and continuous decay, and the four most frequent FGs are the unsubstituted aromatic nitrogen ([Nar]), followed by ether ([R]O[R]), aliphatic hydroxyl (O[Cal]), and tertiary amine ([R]N([R])[R]). We also see some halogen-containing FGs among the top 20 FGs.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

On the other extreme of FG diversity, the GFL set has the lowest one: only 37 unique FGs. As a result, this set has only 0.01 unique FGs per molecule, the lowest average among all the sets. In this case, however, there is again no molecule without FGs, and interestingly this set displays the highest total number of FGs per molecule, 8.31, although it goes down to 2.58 if we use binary counts (now the second largest, after DB). In this case, the distribution shows two outstanding FGs, ester (O=C[R]O[R]) and alkene (C=C), and after these two, the FG counts drop dramatically. All these values reflect the structural homogeneity of this compound set, that results from the repetition of both the alkene and ester FGs in many of these molecules. In other words, these molecules display many although repeated FGs in their structure.

In between these two extremes of diversity we have the FG counts of G/S and GnFL. Both sets have similar numbers of unique FGs, 69 and 78, respectively, resulting in nearly the same number of unique FGs per molecule, namely 0.13 and 0.12. Although the diversity is intermediate between those of DB and GFL, the decoration of these molecules with FGs is lower: 3.44 and 3.10 FGs per molecule using total counts, and 2.39 and 2.22 FGs per molecule using binary counts. In addition, we can see a small but not null percentage of molecules lacking FGs, 1.10 % and 5.12 %, respectively for G/S and GnFL. Both distributions show an intermediate decay between those of DB and GFL.

In terms of the most frequent FGs, we see both commonalities and differences. The two compound sets have the carboxylic acid (O=C(O)[R]) as the most frequent FG. This is the sixth in DB. However, carboxylic acid is followed by aliphatic hydroxyl (O[Cal]) in the case of

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

G/S, but by the substituted amide (O=C([R])N([R])[R]) in the case of GnFL, which is the sixth in G/S. Moreover, the two FGs are followed by two aromatic ones (O[Car] and [Nar]) in G/S, but by the aliphatic ones N[Cal] and O[Cal] in GnFL. These similarities and differences are in agreement with the similarities and differences previously observed by us regarding physicochemical properties.[40]

It is possible to look at additional features of these FGs and find clear patterns among the different sets. Table 2 lists the count statistics of FGs with aromatic atoms and with heteroatoms.

	Ar FGs (bin) / mol	# Ar FGs (un)	Ar FGs (un) / mol	Het FGs (bin) / mol	# Het FGs (un)	Het FGs (un) / mol
DB	0.766	21	0.015	3.341	243	0.171
G/S	0.551	6	0.011	2.321	66	0.12
GnFL	0.255	7	0.011	2.081	74	0.111
GFL	0.001	5	0.001	1.637	35	0.006

Table 2. *Count statistics of aromatic atom- and heteroatom-containing FGs in the different compound sets. Aromatic atom- and heteroatom-containing FGs generated as described in Materials and Methods. Ar FGs (bin) / mol = number of aromatic FGs (binary counts) by molecule; # Ar FGs (un) = number of unique aromatic FGs; Ar FGs (un) / mol = the previous count normalized by the number of molecules in the set; Het FGs (bin) / mol = number of heteroatom FGs (binary counts) by molecule; # Het FGs (un) = number of unique heteroatom FG; Het FGs (un) / mol = the previous count normalized by the number of molecules in the set.*

From there, it is possible to find that, both in terms of binary counts of FG per molecule, and unique FG per molecule, DB shows the most aromatic FGs and with the highest heteroatom content, while GFL has the least aromatic FGs and with the lowest heteroatom

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

content. In between are again G/S and GnFL, displaying similar but intermediate values in these two quantities.

We can also look at the element composition of these FGs. Table 3 collects the number of functional groups per molecule (binary counts), plus the fraction of unique functional groups, for the following heteroatoms: oxygen, nitrogen, sulfur, phosphorus, and halogens, respectively.

	O FGs (bin) / mol	Frac O FGs (un)	N FGs (bin) / mol	Frac N FGs (un)	S FGs (bin) / mol	Frac S FGs (un)	P FGs (bin) / mol	Frac P FGs (un)	X FGs (bin) / mol	Frac X FGs (un)
DB	1.894	0.783	1.575	0.589	0.236	0.17	0.026	0.079	0.31	0.04
G/S	1.776	0.725	0.73	0.406	0.082	0.159	0.022	0.043	0.013	0.029
GnFL	1.56	0.654	0.827	0.359	0.083	0.218	0.021	0.038	0	0
GFL	1.442	0.73	0.205	0.216	0.001	0.135	0.203	0.108	0	0

Table 3. *Count statistics of FGs containing O, N, S, P, or halogen atoms (X) by compound set. O FGs (bin) / mol = O-containing FGs (binary counts) per molecule; Frac O FGs (un) = fraction of unique FGs containing O; N FGs (bin) / mol = N-containing FGs (binary counts) per molecule; Frac N FGs (un) = fraction of unique FGs containing N; S FGs (bin) / mol = S-containing FGs (binary counts) per molecule; Frac S FGs (un) = fraction of unique FGs containing S; P FGs (bin) / mol = P-containing FGs (binary counts) per molecule; Frac P FGs (un) = fraction of unique FGs containing P; X FGs (bin) / mol = Halogen-containing FGs (binary counts) per molecule; Frac X FGs (un) = fraction of unique FGs containing Halogen atoms. The counts are not excluding, i.e. it is possible to have FGs that belong to more than one of the FG type (having more than one different heteroatom, and with or without aromatic atoms).*

Some patterns emerge from these data that are worth mentioning. Regarding O- and N-containing FGs, the DB set has both the largest binary counts per molecule, and fractions of

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

unique FGs among all the compound sets, while G/S and GnFL show lower values. In the case of S-containing FGs, DB has again by far the largest binary counts per molecule, but GnFL presents the largest fraction of unique FGs of this type. GFL, while displaying the lowest binary counts of O-, N-, and S-containing FGs per molecule among the four compound sets, has by far the largest counts per molecule and fractions of P-containing FGs, reflecting the presence of glycerophospholipids, sphingolipids, etc., in this compound set. Finally, as far as X-containing FGs are concerned, it is interesting to see their absence in both GnFL and GFL, intermediate values in G/S, and the highest values in DB. In general, we observe a rough trend where the binary counts per molecule of O-, N-, S-, and X-containing FGs show the decreasing order DB > G/S ≈ GnFL > GFL, but GFL has the highest values of P-containing FGs.

Finally, it is possible to analyze the co-occurrence distributions of these FGs. Figure 2 shows the co-occurrence matrices of the top 20 FGs for the four compound sets studied in this work.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

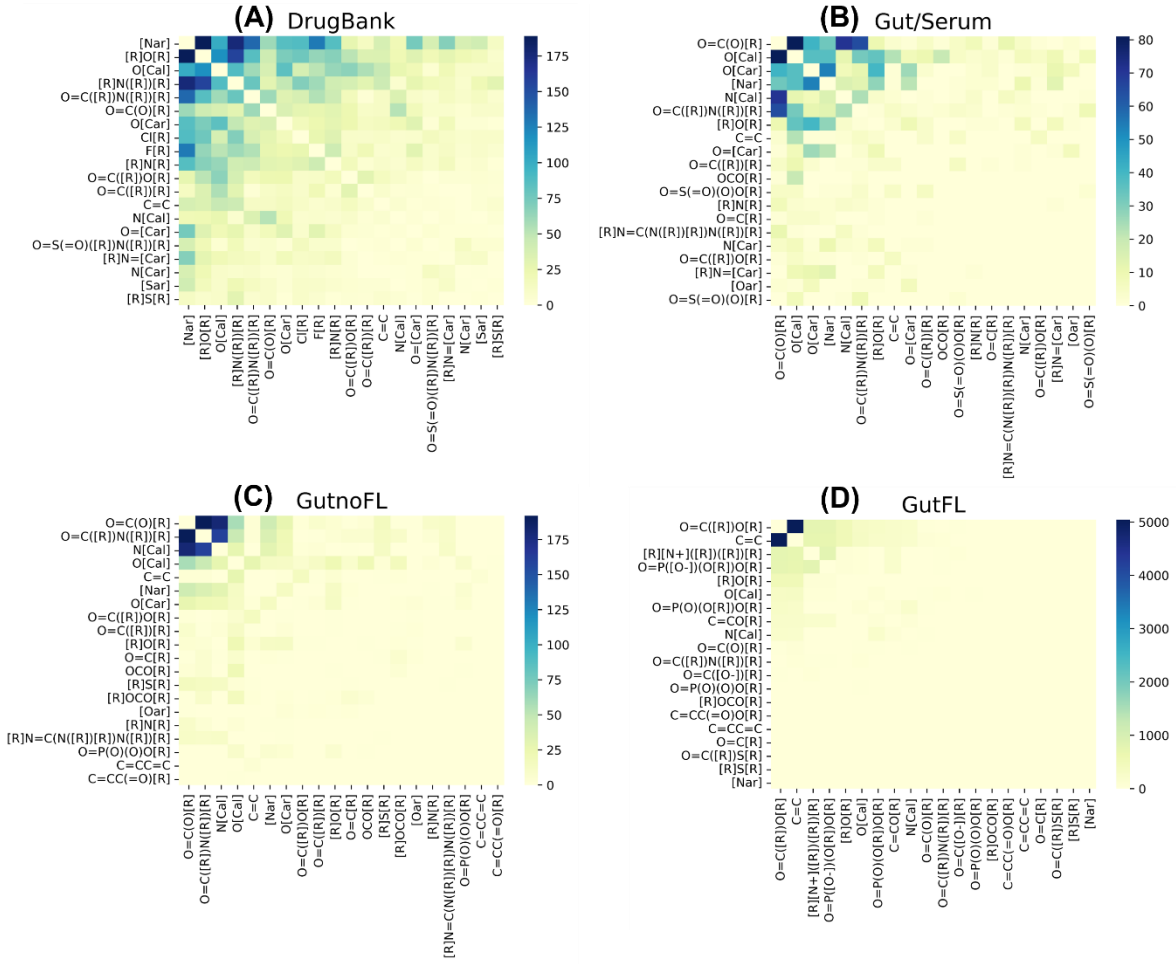


Figure 2. *Co-occurrence matrices of FGs for the four compound sets: (a) DB; (b) G/S; (c) GnFL; (d) GFL. A pair of FGs is said to co-occur in a molecule if one or more copies of each FG is present in the molecule.*

It is interesting to see cases where the same FG is correlated with different FGs depending on the compound set. For example, in both G/S and GnFL we observe the carboxylic acid FG to be highly correlated with both the primary alkyl amine and amide. However, in G/S, the carboxylic acid is mostly correlated with the aliphatic hydroxyl group, but this high correlation is not observed in GnFL. While the aromatic nitrogen [Nar] is mostly correlated in DB with the ether [R]O[R] and tertiary amine [R]N([R])[R], it is mostly correlated with the

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

phenolic hydroxyl O[Car] in G/S and the carboxylic acid FG in GnFL, etc. We can also see that the two FGs dominating the GFL set (the ester and alkene FGs) are highly correlated.

It is possible to have an aggregated view of all these features by calculating the cosine similarity between the normalized distributions of FGs (using binarized counts) between each pair of compound sets. Table 3 shows the similarities, from which it can be concluded that G/S and GnFL display the largest similarity (0.91), followed by DB and G/S (0.71), and DB and GnFL (0.63). GFL is highly dissimilar to the other three compound sets (similarities of 0.24, 0.15, and 0.25 with DB, G/S, and GnFL, respectively).

	Gut/Serum	**GutnoFL**	**GutFL**
DrugBank	0.707	0.626	0.237
Gut/Serum		0.912	0.151
GutnoFL			0.252

Table 4. *Cosine similarities between pairs of compound sets, from the normalized distributions of binary counts of the 607 distinct FGs across all the sets.*

In summary, from these analyses a picture emerges that is in agreement with our previous analysis, that considered physicochemical and scaffold properties.[40] On one extreme we see that the DB set comprises highly decorated drug-like molecules, with the largest diversity and counts per molecule of FGs (~5), especially those with aromatic, N, and halogen atoms. These are molecules capable of crossing the gut wall, in agreement with the Lipinski's[57] and Veber's[58] rules that they follow (with a few exceptions of gut-acting drugs). On the opposite side are the GFL molecules, with the lowest diversity in FGs (although with very high FG counts per molecule, ~8), lowest counts of aromatic FGs, and highest counts of P-containing FGs. These molecules cannot cross the gut wall, as expected as they do not adhere to

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Lipinski and Veber's rules. Finally, between these extremes lie the G/S and GnFL sets, the former with a distribution of FGs closer to DB than the latter. These two sets have intermediate diversity and counts per molecule of FGs (~3), intermediate content of aromatic FGs, as well as intermediate N, S and X content. Combining this intermediate "FG-complexity" with their lower size and lipophilicity, and their higher fsp3 carbon content, we conclude that their structure is more akin to fragments. In these molecules abound carboxylic acids, aliphatic primary amines and alcohols, and amides, but they also display other diverse FGs. While the G/S set is assumed to be "gut-traverser", like the DB that it resembles, the GnFL set is "gut-lingerer", and in addition to the factors identified in our previous work that make them different (the former set has less rotatable bonds, more hydrogen-bond donors/acceptors, and more rings), here we find additional differences based on their FGs: G/S has almost twice the diversity of aromatic FGs per molecule, FGs with halogens, much lower amounts of amide FGs, and low carboxylic acid vs aliphatic hydroxyl correlation. In turn, GnFL has much higher fractions of amides, together with high carboxylic acid vs aliphatic hydroxyl correlations, among others. These additional patterns should be taken into account in order to design molecules to remain in the gut or alternatively to cross the gut wall.

Analysis of interactions with proteins

A total of 22116 PDB structures with one or more of these compounds were retrieved and used for the analysis, which corresponded to a total of 691 non-covalently bound ("free") ligands of our compounds, and a total of 71525 protein-ligand instance complexes. In

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

addition to not being covalently bound to the protein, we also imposed the ligand to be complete, and to have a fraction of solvent accessible surface buried in the protein > 0.2, so that we only considered metabolites significantly bound to the corresponding target. Table 5 shows the distribution of compounds present in PDB complexes, as well as the fraction of the total compounds, for each of the four compound sets.

	# compounds in PDB	% In PDB
DB	352	24.56
G/S	194	34.40
GnFL	76	11.31
GFL	69	1.27

Table 5. *Number of compounds in PDB complexes (n in PDB) and its percentage over all compounds (% in PDB), for each compound set.*

From here we see that GFL is very underrepresented in the PDB, compared to the other compound sets. However, taking into consideration the structural homogeneity of these molecules, it is expected that the molecules here analyzed would provide representatively enough information for that compound set. In fact, the distributions by chemical classes of molecules in PDB complexes remained reasonably similar to the corresponding distributions of the full set of molecules in the compound set, in each of the four compound sets (data not shown). Therefore, we do not expect significant biases in our analyses of interactions in terms of chemical class composition of the compound sets.

In our analysis of interactions with BINANA 2 we included the following interaction types: “closest”, “electrostatic energies”, “hydrophobics”, “hydrogen bonds”, “halogen bonds”, “pi-pi”, “cation-pi”, “salt-bridges”, and “metal coordinations”. We did not consider “close”

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

as it was present in all the molecules, irrespective of the compound set. Interactions were averaged by FGs and ligands, in order to avoid the biasing of the distributions by the very differential abundance of ligands in the dataset. Figure 3 displays the distributions of the different types of interactions for the four compound sets. Statistically significant differences in the distributions are observed for all the interaction types, judging from a Kruskal-Walli's test p-value < 0.001 in all the cases. Additional post-hoc analyses, coupled with CLES calculations, allow us to find the differences between pairs of distributions that account for the significance of the omnibus test and, when that is the case, which of the distributions is shifted towards higher values (higher CLES). In Figure 3, the significant pairs of distributions are surrounded by a same-color box, with continuous border and up-pointing arrow the distribution with higher CLES, and discontinuous border the distribution with lower CLES. We will only mention significant pairs giving CLES values bigger than 0.6 or less than 0.4.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

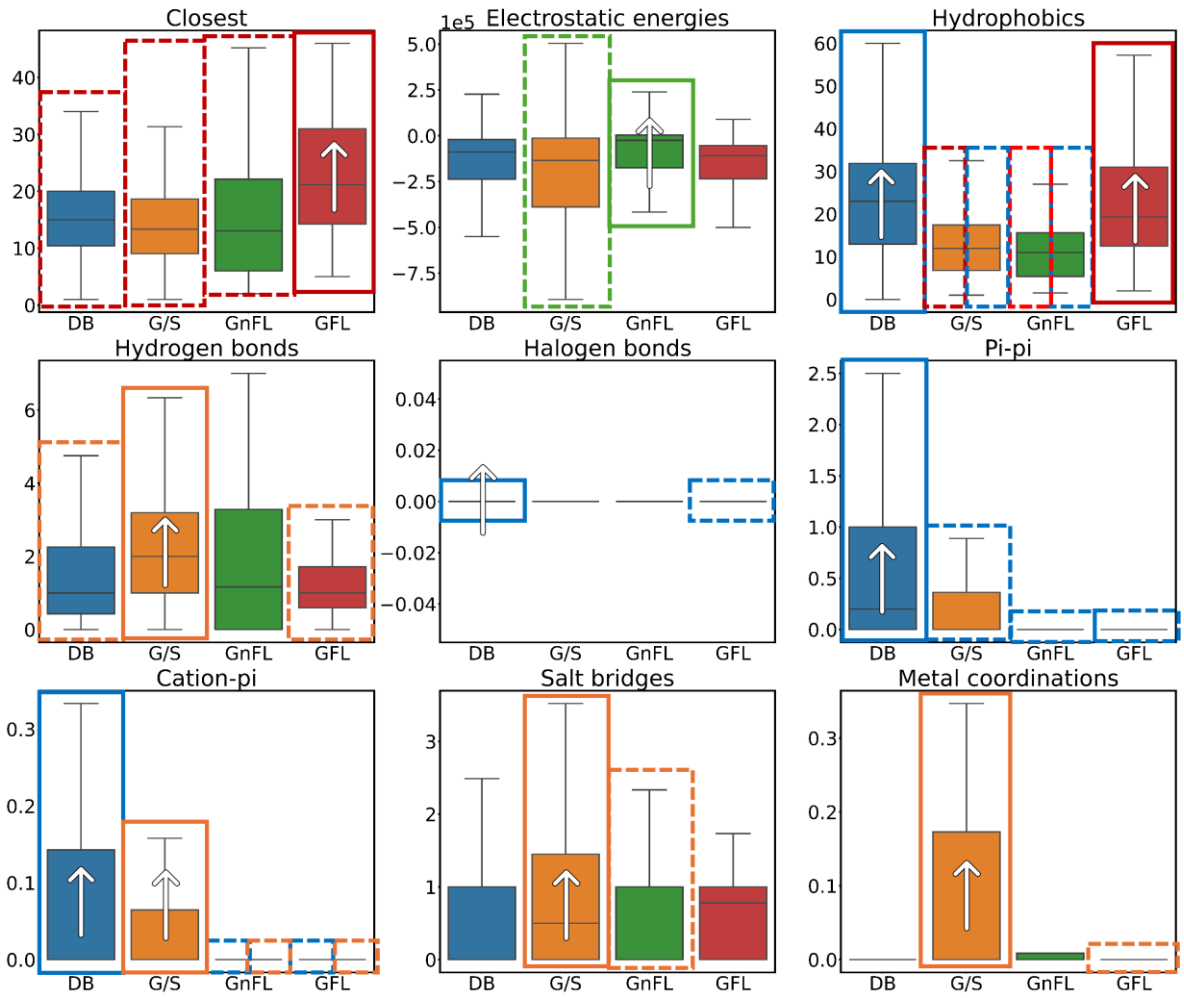


Figure 3. *Distributions of interaction counts for different interaction types and the four compound sets. Outliers have been omitted for clarity purposes. Distributions significantly different in post-hoc tests after multiple test correction, and with CLES > 0.6 or CLES < 0.4, are surrounded by boxes, continuous line with arrow significantly higher, discontinuous line significantly lower.*

From this analysis it is possible to see that GFL has a distribution of the closest type of interaction very significantly shifted to higher values when compared to DB, G/S, and GnFL. This reflects the high density of packing in the binding site of this type of elongated molecules, compared to other more globular compounds. GFL is also characterized by significantly higher CLES in the hydrophobics distribution (compared to G/S and GnFL),

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

reflecting the large hydrophobicity of these molecules. In turn, the DB set has significantly higher distributions of hydrophobic interactions (compared to G/S and GnFL) too, but also of halogen bonds (compared to GFL), pi-pi interactions (compared to the rest of the sets), and cation-pi interactions (compared to GnFL and GFL). The other two metabolite sets, G/S and GnFL, display commonalties and differences, as usual. Both sets have significantly lower hydrophobic and pi-pi interaction distributions. However, the G/S set has significantly higher distributions of hydrogen bonds than DB and GFL, as well as cation-pi interactions (compared to GnFL and GFL, not to DB), salt bridges (compared to GnFL) and metal coordination (compared to GFL), while the GnFL is characterized by significantly higher electrostatic energies (less attractive) compared to G/S.

It is therefore clear that the diverse types of molecules tend to establish significantly different types of interactions with proteins. These patterns can be used in the rational design of new bioactive molecules that mimic the different types of microbial metabolites, and thus their target preferences.[40,41]

Association between interaction types and FGs in the compound sets

Given the analysis of FGs described in the first results section, and the interaction distributions shown in the second section, we correlated both datasets to find which interactions were used by each FG and in each compound set. In this way, we aimed to rationalize the observed differences in interaction types in terms of differences of FG distributions and / or putative differences in the use of interaction types by the same FG in different compound sets. A FG was considered to use a interaction when at least one of the

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

FG atoms itself (not considering the “environment” carbon atoms) were part of the list of ligand atoms identified by BINANA 2 as taking part in the interaction. Electrostatic interactions were not included in this analysis since BINANA 2 does not provide the atom indexes of the atoms involved, only the atom types. For each FG and ligand, the presence / absence of interaction was averaged, to get a value between 0 and 1, i.e. the fraction of the FG having the interaction in that ligand. Then, for each FG and compound set, the fraction values of interaction for each FG were averaged to get a new value between 0 and 1. This process was done in order to avoid the statistics to be biased by the very differential abundance of the different ligands. The resulting fractions are shown as heatmaps in Figure 4, for the four compound sets.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

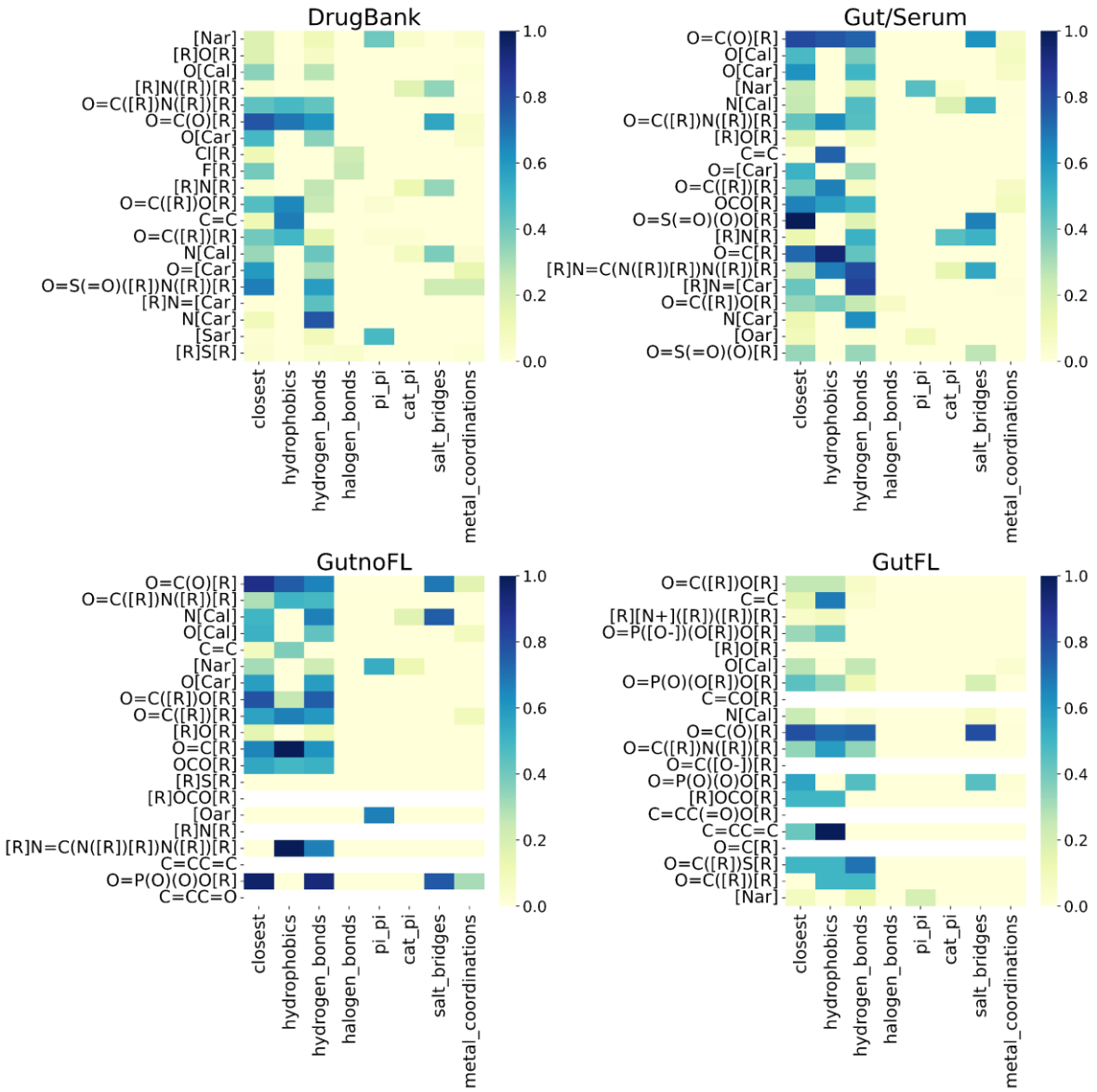


Figure 4. *Fraction of interaction usage for each of the top 20 FG for the four compound sets (aggregated by ligand and compound set). Blank cells correspond to FGs missing in the PDB data.*

From these results, it is possible to assign and understand the way that the different Ertl FGs interact with proteins. For example, the [Nar] one, representing a non-substituted nitrogen in an aromatic ring, interacts preferentially through pi-pi interactions, although also, to a lower extent, through closest (contacts) and hydrogen bonds. In the same way,

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the carboxylic acid O=C(O)[R] interacts through a mixture of closest, salt bridges, hydrogen bonds, and hydrophobic interactions. Both hydroxyl (O[Cal]) and phenol (O[Car]) FGs interact similarly, that is, through closest and hydrogen bond interactions, while unsubstituted aliphatic (N[Cal]) and aromatic (N[Car]) amines have differences: while the former make mainly use of salt bridges, hydrogen bonds, and closest, the latter use mainly hydrogen bonds, and closest. Metal coordination using FGs include phosphate (O=P(O)(O)O[R]), N-substituted sulphonamide (O=S(=O)([R])N([R])[R]), and O=[Car]. Among the top 20 FGs, only Cl[R] and F[R] in the DB compound set show halogen bond interactions.

Focusing on the explanation of the patterns in Figure 3, the interaction-FG fractions here obtained provide a useful means to rationalize them, both in terms of high fractions, and FG abundance. For example, the high hydrophobicity in GFL can be ascribed mainly to the C=C and C=CC=C FGs in this compound set, while the high closest distribution is explained by the large amounts of O=(O)C[R] in these molecules. The G/S high hydrogen bonds are to be attributed to O=(O)C[R], the guanidine [R]N=C(N([R])[R])N([R])[R], and the aryl imine [R]N=[Car]; the high salt bridges to O=(O)C[R], the sulphonic acid O=S(=O)(O)O[R], and N[Cal], among others; and the high metal coordinations is spread across multiple FGs. Finally, the DB high hydrophobics can be understood as a result of C=C, O=C(O)[R], O=C([R])O[R] contributions, while the high pi-pi should be ascribed mainly to the [Nar] and [Sar] FGs in this set, and the high cation-pi interactions would result from the basic aliphatic amines N[Cal], [R]N[R], and [R]N([R])[R].

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

In general, there seem to be no large differences in the interaction patters for the different FGs among the four compound sets. However, the proportions of the different interactions can vary to some extent among the different compound sets, and also in some cases a difference in the global usage of FGs is observed. For example, the aryl imine [R]N=[Car] has in G/S the following non-null fractions: 0.42 for closest, 0.83 for hydrogen bonds, and 0.01 for metal coordinations, summing a total of 1.27. In turn, in DB the non-null fractions are 0.011 for closest, and 0.44 for hydrogen bonds, summing a total of 0.45. Thus, it seems that, although the types of interactions used are similar, in the case of DB the usage of the FG for them is lower. This could reflect the differential co-occurrence of FGs across the compound sets, that would compete for the same protein FGs or create steric constraints in the ligand's nearby FGs.

In summary, we conclude that the distributions of FGs among the four compound sets, as well as their fractions of interaction usage, can explain the observed statistically different distributions of interactions. Thus, the compound set-specific distributions of FGs and FG-FG correlations must be taken into account when designing drugs and nutraceutics mimicking the different microbial metabolites (G/S, GnFL, GFL), as they determine their respective interaction patterns. The full set of mapped FGs and BINANA 2 interactions is collected in Table S1 of Supporting Information.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

DISCUSSION

One of the key concepts used in drug design is that of FGs.[39,59] These are sets of connected atoms in molecules with specific physicochemical and reactivity properties. They determine the types of weak interactions that small molecules establish with proteins, thus explaining, together with molecular shape, the specificity of ligand-protein binding. FGs in the ligand must match complementary FGs in the protein in order to form a stable protein-ligand complex.

In the past, much work has been devoted to analyze FG distributions and their evolution in drugs and natural products.[48,60,61] The emerging field of microbial metabolite mimetic drug design thus requires to extend this work to microbial metabolites, in order to gain insights of the typical FGs in these molecules that could be used for the rational design of mimetic molecules. The present work accomplishes that task and identifies clear differential patterns among the three microbial metabolite sets: G/S, GnFL, GFL, and in addition differences with oral (mainly systemic) drugs. This is reflected in different most frequent FGs, the co-occurrence of these FGs, and the types of FGs in terms of aromaticity and heteroatom content. Briefly, GFL shows the lowest FG diversity, highest counts of FG per molecule, and highest counts of P-containing FGs. G/S and GnFL, both having intermediate diversity, proportion of aromatic atoms and content of N, S, and X in their FGs, and having abundant carboxylic acids, aliphatic amines, alcohols, and amides, display differences such as: higher aromatic and halogen content in G/S, and different co-occurrence rates for several pairs of FGs.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

In order to get further insight on how these molecules bind to their targets, the interactions established between metabolites and proteins are analyzed here using a set of more than 71 K experimental protein-metabolite complexes present in the PDB. From here, we find statistically different distributions for all the interaction types among the metabolite compound sets, and with the oral drug set too. Roughly, G/S stems for high hydrogen bonds, salt bridges, and metal coordination; GnFL shows high electrostatic energies; and GFL has high closest and hydrophobics.

Moreover, the mapping of FG atoms to BINANA 2-detected interactions has been able to explain the interaction patterns in terms of FG distributions and their usage of diverse types of interactions. In general, the same FG establishes similar interaction patterns irrespective of the compound set. This can be expected given that the FG is a relatively isolated chemical entity that in principle displays specific physical properties and interaction propensities with protein FGs. However, there are some cases where a variation in the fractions of interaction usages is identified, as well as global differential usage levels of all the interaction types for the same FG among different compound sets. This could be explained in terms of differential FG co-occurrence, such as if the interaction pattern of one FG is modified by the diverse competition of other FGs for the same site in the protein and / or diverse steric constraints originated by the co-occurring FGs.

In the present work several novelties are worth mentioning: a) it is the first time that Ertl groups are analyzed in microbial metabolites; and b) it is the first time that their interactions with proteins are systematically analyzed. From the point of view of FG analysis, it is the

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

first time where co-occurrences of Ertl FGs are analyzed. This novel co-occurrence analysis has been able to unravel new patterns, such as a differential co-occurrence of the same pairs of highly abundant FGs in two compound sets (e.g. G/S vs GnFL). The application to other chemical sets is ongoing in our group. On the other hand, this is also the first time Ertl FGs are systematically mapped into molecular interactions with proteins, so that they are analyzed not only in terms of their presence in ligands, but also in terms of their interactions with proteins. Interesting patterns are thus detected like the variability in interaction usage in some cases described above, besides the generally conserved usage for the same FG across different compound sets.

We expect that this work will illuminate the rational design of novel microbial metabolite mimetic drugs and nutraceutics. The patterns here found can guide the medicinal chemists in designing novel molecules based on the different metabolites' compound sets. In addition, our group is working on novel generative AI models that tap from these data and methods.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

AUTHOR INFORMATION

Corresponding Author:

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

Authors:

Mario Astigarraga, Andrés Sánchez-Ruiz (https://orcid.org/0000-0001-7393-542X), Aminata Diop-Aw, Raquel Quintero - Biostatistics and Bioinformatics Unit. IMDEA Food, CEI UAM+CSIC, E28049 Madrid, Spain

Notes

The authors declare no competing financial interest.

DATA AND SOFTWARE AVAILABILITY STATEMENT

Metabolite and drugs chemical structures, together with the corresponding metadata (ClassyFire hierarchy, etc.) were retrieved from the Human Metabolome Database (https://hmdb.ca/) and DrugBank (https://go.drugbank.com/), respectively. Protein-ligand structures were retrieved from the Protein Data Bank (https://www.rcsb.org/) through its API. BINANA 2 was obtained from https://github.com/durrantlab/binana. Moleculekit module was installed in Conda through the acellera channel. PDBFixer, openbabel, and biopython modules were installed in Conda through the conda-forge

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

channel. The complete and accurate Ertl algorithm was implemented locally for RDKit and will be published elsewhere.

SUPPORTING INFORMATION

Supporting Information Table S1 is collected in file “Supporting Information.xlsx”. In addition, file “Supporting Information.pdf” contains captions for that table.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

REFERENCES

(1) 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.

(2) Chavira, A.; Belda-Ferre, P.; Kosciolek, 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.

(3) Schmidt, T. S. B.; Raes, J.; Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **2018**, *172* (6), 1198–1215. https://doi.org/10.1016/j.cell.2018.02.044.

(4) 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.

(5) Cammann, D.; Lu, Y.; Cummings, M. J.; Zhang, M. L.; Cue, J. M.; Do, J.; Ebersole, J.; Chen, X.; Oh, E. C.; Cummings, J. L.; Chen, J. Genetic Correlations between Alzheimer's Disease and Gut Microbiome Genera. *Sci Rep* **2023**, *13* (1), 5258. https://doi.org/10.1038/s41598-023-31730-5.

(6) 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.

(7) Ferreiro, A. L.; Choi, J.; Ryou, J.; Newcomer, E. P.; Thompson, R.; Bollinger, R. M.; Hall-Moore, C.; Ndao, I. M.; Sax, L.; Benzinger, T. L. S.; Stark, S. L.; Holtzman, D. M.; Fagan, A. M.; Schindler, S. E.; Cruchaga, C.; Butt, O. H.; Morris, J. C.; Tarr, P. I.; Ances, B. M.; Dantas, G. Gut Microbiome Composition May Be an Indicator of Preclinical Alzheimer's Disease. *Science Translational Medicine* **2023**, *15* (700), eabo2984. https://doi.org/10.1126/scitranslmed.abo2984.

(8) Harris, V. C.; Haak, B. W.; Boele Van Hensbroek, M.; Wiersinga, W. J. The Intestinal Microbiome in Infectious Diseases: The Clinical Relevance of a Rapidly Emerging Field. *Open Forum Infectious Diseases* **2017**, *4* (3), ofx144. https://doi.org/10.1093/ofid/ofx144.

(9) Hawkins, K. G.; Casolaro, C.; Brown, J. A.; Edwards, D. A.; Wikswo, J. P. The Microbiome and the Gut-Liver-Brain Axis for Central Nervous System Clinical Pharmacology: Challenges in Specifying and Integrating In Vitro and In Silico Models. *Clinical Pharmacology & Therapeutics* **2020**, *108* (5), 929–948. https://doi.org/10.1002/cpt.1870.

(10) Liu, J.; Lahousse, L.; Nivard, M. G.; Bot, M.; Chen, L.; van Klinken, J. B.; Thesing, C. S.; Beekman, M.; van den Akker, E. B.; Slieker, R. C.; Waterham, E.; van der Kallen, C. J. H.; de Boer, I.; Li-Gao, R.; Vojinovic, D.; Amin, N.; Radjabzadeh, D.; Kraaij, R.; Alferink, L. J. M.; Murad, S. D.; Uitterlinden, A. G.; Willemsen, G.; Pool, R.; Milaneschi, Y.; van Heemst, D.; Suchiman, H. E. D.; Rutters, F.; Elders, P. J. M.; Beulens, J. W. J.; van der Heijden, A. A. W. A.; van Greevenbroek, M. M. J.; Arts, I. C. W.; Onderwater, G. L. J.; van den Maagdenberg, A. M. J. M.; Mook-Kanamori, D. O.; Hankemeier, T.; Terwindt, G. M.; Stehouwer, C. D. A.; Geleijnse, J. M.; 't Hart, L. M.; Slagboom, P. E.; van Dijk, K. W.;

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Zhernakova, A.; Fu, J.; Penninx, B. W. J. H.; Boomsma, D. I.; Demirkan, A.; Stricker, B. H. C.; van Duijn, C. M. Integration of Epidemiologic, Pharmacologic, Genetic and Gut Microbiome Data in a Drug–Metabolite Atlas. *Nat Med* **2020**, *26* (1), 110–117. https://doi.org/10.1038/s41591-019-0722-x.

(11) Sepich-Poore, G. D.; Zitvogel, L.; Straussman, R.; Hasty, J.; Wargo, J. A.; Knight, R. The Microbiome and Human Cancer. *Science* **2021**, *371* (6536), eabc4552. https://doi.org/10.1126/science.abc4552.

(12) Wang, C.; Bai, J.; Chen, X.; Song, J.; Zhang, Y.; Wang, H.; Suo, H. Gut Microbiome-Based Strategies for Host Health and Disease. *Critical Reviews in Food Science and Nutrition* **2023**.

(13) Zheng, X.; Cai, X.; Hao, H. Emerging Targetome and Signalome Landscape of Gut Microbial Metabolites. *Cell Metabolism* **2022**, *34* (1), 35–58. https://doi.org/10.1016/j.cmet.2021.12.011.

(14) Morozumi, S.; Ueda, M.; Okahashi, N.; Arita, M. Structures and Functions of the Gut Microbial Lipidome. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2022**, *1867* (3), 159110. https://doi.org/10.1016/j.bbalip.2021.159110.

(15) Aleti, G.; Troyer, E. A.; Hong, S. G Protein-Coupled Receptors: A Target for Microbial Metabolites and a Mechanistic Link to Microbiome-Immune-Brain Interactions. *Brain, Behavior, & Immunity - Health* **2023**, *32*, 100671. https://doi.org/10.1016/j.bbih.2023.100671.

(16) 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.

(17) Han, S.; Van Treuren, W.; Fischer, C. R.; Merrill, B. D.; DeFelice, B. C.; Sanchez, J. M.; Higginbottom, S. K.; Guthrie, L.; Fall, L. A.; Dodd, D.; Fischbach, M. A.; Sonnenburg, J. L. A Metabolomics Pipeline for the Mechanistic Interrogation of the Gut Microbiome. *Nature* **2021**, *595* (7867), 415–420. https://doi.org/10.1038/s41586-021-03707-9.

(18) Mager, L. F.; Burkhard, R.; Pett, N.; Cooke, N. C. A.; Brown, K.; Ramay, H.; Paik, S.; Stagg, J.; Groves, R. A.; Gallo, M.; Lewis, I. A.; Geuking, M. B.; McCoy, K. D. Microbiome-Derived Inosine Modulates Response to Checkpoint Inhibitor Immunotherapy. *Science* **2020**, *369* (6510), 1481–1489. https://doi.org/10.1126/science.abc3421.

(19) Agus, A.; Clément, K.; Sokol, H. Gut Microbiota-Derived Metabolites as Central Regulators in Metabolic Disorders. *Gut* **2021**, *70* (6), 1174–1182. https://doi.org/10.1136/gutjnl-2020-323071.

(20) Ay, Ü.; Leníček, M.; Classen, A.; Olde Damink, S. W. M.; Bolm, C.; Schaap, F. G. New Kids on the Block: Bile Salt Conjugates of Microbial Origin. *Metabolites* **2022**, *12* (2), 176. https://doi.org/10.3390/metabo12020176.

(21) Chen, H.; Nwe, P.-K.; Yang, Y.; Rosen, C. E.; Bielecka, A. A.; Kuchroo, M.; Cline, G. W.; Kruse, A. C.; Ring, A. M.; Crawford, J. M.; Palm, N. W. A Forward Chemical Genetic Screen Reveals Gut Microbiota Metabolites That Modulate Host Physiology. *Cell* **2019**, *177* (5), 1217-1231.e18. https://doi.org/10.1016/j.cell.2019.03.036.

(22) Colosimo, D. A.; Kohn, J. A.; Luo, P. M.; Piscotta, F. J.; Han, S. M.; Pickard, A. J.; Rao, A.; Cross, J. R.; Cohen, L. J.; Brady, S. F. Mapping Interactions of Microbial Metabolites with Human G-Protein-Coupled Receptors. *Cell Host & Microbe* **2019**, *26* (2), 273-282.e7. https://doi.org/10.1016/j.chom.2019.07.002.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(23) Fobofou, S. A.; Savidge, T. Microbial Metabolites: Cause or Consequence in Gastrointestinal Disease? *American Journal of Physiology-Gastrointestinal and Liver Physiology* **2022**, *322* (6), G535–G552. https://doi.org/10.1152/ajpgi.00008.2022.
(24) Garrett, W. S. Immune Recognition of Microbial Metabolites. *Nat Rev Immunol* **2020**, *20* (2), 91–92. https://doi.org/10.1038/s41577-019-0252-2.
(25) Ghosh, S.; Whitley, C. S.; Haribabu, B.; Jala, V. R. Regulation of Intestinal Barrier Function by Microbial Metabolites. *Cellular and Molecular Gastroenterology and Hepatology* **2021**, *11* (5), 1463–1482. https://doi.org/10.1016/j.jcmgh.2021.02.007.
(26) Krautkramer, K. A.; Fan, J.; Bäckhed, F. Gut Microbial Metabolites as Multi-Kingdom Intermediates. *Nat Rev Microbiol* **2021**, *19* (2), 77–94. https://doi.org/10.1038/s41579-020-0438-4.
(27) 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.
(28) Nemet, I.; Saha, P. P.; Gupta, N.; Zhu, W.; Romano, K. A.; Skye, S. M.; Cajka, T.; Mohan, M. L.; Li, L.; Wu, Y.; Funabashi, M.; Ramer-Tait, A. E.; Naga Prasad, S. V.; Fiehn, O.; Rey, F. E.; Tang, W. H. W.; Fischbach, M. A.; DiDonato, J. A.; Hazen, S. L. A Cardiovascular Disease-Linked Gut Microbial Metabolite Acts via Adrenergic Receptors. *Cell* **2020**, *180* (5), 862-877.e22. https://doi.org/10.1016/j.cell.2020.02.016.
(29) Rahman, S.; O'Connor, A. L.; Becker, S. L.; Patel, R. K.; Martindale, R. G.; Tsikitis, V. L. Gut Microbial Metabolites and Its Impact on Human Health. *Ann Gastroenterol* **2023**, *36* (4), 360–368. https://doi.org/10.20524/aog.2023.0809.
(30) Cully, M. Microbiome Therapeutics Go Small Molecule. *Nat Rev Drug Discov* **2019**, *18* (8), 569–572. https://doi.org/10.1038/d41573-019-00122-8.
(31) 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.
(32) 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.
(33) 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.
(34) Descamps, H. C.; Herrmann, B.; Wiredu, D.; Thaiss, C. A. The Path toward Using Microbial Metabolites as Therapies. *eBioMedicine* **2019**, *44*, 747–754. https://doi.org/10.1016/j.ebiom.2019.05.063.
(35) Dvořák, Z.; Kopp, F.; Costello, C. M.; Kemp, J. S.; Li, H.; Vrzalová, A.; Štěpánková, M.; Bartoňková, I.; Jiskrová, E.; Poulíková, K.; Vyhlídalová, 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.
(36) Dvořák, Z.; Sokol, H.; Mani, S. Drug Mimicry: Promiscuous Receptors PXR and AhR, and Microbial Metabolite Interactions in the Intestine. *Trends in Pharmacological Sciences* **2020**, *41* (12), 900–908. https://doi.org/10.1016/j.tips.2020.09.013.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(37) Grycová, A.; Joo, H.; Maier, V.; Illés, P.; Vyhlídalová, B.; Poulíková, K.; Sládeková, L.; Nádvorník, P.; Vrzal, R.; Zemánková, L.; Pečinková, 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.
(38) Guan, X.-J.; Zhang, Y.-Y.; Zheng, X.; Hao, H.-P. Drug Discovery Inspired from Nuclear Receptor Sensing of Microbial Signals. *Trends in Molecular Medicine* **2021**, *27* (7), 624–626. https://doi.org/10.1016/j.molmed.2021.03.007.
(39) Hanson, J. R.; Hanson, J. R. *Functional Group Chemistry*; Tutorial chemistry texts; Royal Society of Chemistry: Cambridge, 2001.
(40) Gil-Pichardo, A.; Sánchez-Ruiz, A.; Colmenarejo, G. Analysis of Metabolites in Human Gut: Illuminating the Design of Gut-Targeted Drugs. *Journal of Cheminformatics* **2023**, *15* (1), 96. https://doi.org/10.1186/s13321-023-00768-y.
(41) Orgaz, C.; Sánchez-Ruiz, A.; Colmenarejo, G. Identifying and Filling the Chemobiological Gaps of Gut Microbial Metabolites. *J. Chem. Inf. Model.* **2024**, *64* (17), 6778–6798. https://doi.org/10.1021/acs.jcim.4c00903.
(42) *RDKit: Open-source cheminformatics*. https://www.rdkit.org/ (accessed 2021-09-03).
(43) Hamelryck, T.; Manderick, B. PDB File Parser and Structure Class Implemented in Python. *Bioinformatics* **2003**, *19* (17), 2308–2310. https://doi.org/10.1093/bioinformatics/btg299.
(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) Knox, C.; Wilson, M.; Klinger, C. M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N. E. (Lucy); Strawbridge, S. A.; Garcia-Patino, M.; Kruger, R.; Sivakumaran, A.; Sanford, S.; Doshi, R.; Khetarpal, N.; Fatokun, O.; Doucet, D.; Zubkowski, A.; Rayat, D. Y.; Jackson, H.; Harford, K.; Anjum, A.; Zakir, M.; Wang, F.; Tian, S.; Lee, B.; Liigand, J.; Peters, H.; Wang, R. Q. (Rachel); Nguyen, T.; So, D.; Sharp, M.; da Silva, R.; Gabriel, C.; Scantlebury, J.; Jasinski, M.; Ackerman, D.; Jewison, T.; Sajed, T.; Gautam, V.; Wishart, D. S. DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Research* **2024**, *52* (D1), D1265–D1275. https://doi.org/10.1093/nar/gkad976.
(46) 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.
(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) Ertl, P. An Algorithm to Identify Functional Groups in Organic Molecules. *Journal of Cheminformatics* **2017**, *9* (1), 36. https://doi.org/10.1186/s13321-017-0225-z.
(49) Young, J.; Garikipati, N.; Durrant, J. D. BINANA 2: Characterizing Receptor/Ligand Interactions in Python and JavaScript. *J. Chem. Inf. Model.* **2022**, *62* (4), 753–760. https://doi.org/10.1021/acs.jcim.1c01461.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(50) Rose, Y.; Duarte, J. M.; Lowe, R.; Segura, J.; Bi, C.; Bhikadiya, C.; Chen, L.; Rose, A. S.; Bittrich, S.; Burley, S. K.; Westbrook, J. D. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. *Journal of Molecular Biology* **2021**, *433* (11), 166704. https://doi.org/10.1016/j.jmb.2020.11.003.

(51) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V.; Dalenberg, K.; Duarte, J. M.; Dutta, S.; Fayazi, M.; Feng, Z.; Flatt, J. W.; Ganesan, S.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Henry, J.; Hudson, B. P.; Khokhriakov, I.; Lawson, C. L.; Liang, Y.; Lowe, R.; Peisach, E.; Persikova, I.; Piehl, D. W.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Vallat, B.; Voigt, M.; Webb, B.; Westbrook, J. D.; Whetstone, S.; Young, J. Y.; Zalevsky, A.; Zardecki, C. RCSB Protein Data Bank (RCSB.Org): Delivery of Experimentally-Determined PDB Structures alongside One Million Computed Structure Models of Proteins from Artificial Intelligence/Machine Learning. *Nucleic Acids Research* **2023**, *51* (D1), D488–D508. https://doi.org/10.1093/nar/gkac1077.

(52) Openmm/Pdbfixer, 2024. https://github.com/openmm/pdbfixer (accessed 2024-09-11).

(53) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput Biol* **2017**, *13* (7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659.

(54) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12* (4), 1845–1852. https://doi.org/10.1021/acs.jctc.6b00049.

(55) Ravindranath, P. A.; Forli, S.; Goodsell, D. S.; Olson, A. J.; Sanner, M. F. AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLoS Comput Biol* **2015**, *11* (12), e1004586. https://doi.org/10.1371/journal.pcbi.1004586.

(56) 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.

(57) 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.

(58) 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.

(59) Mukherjee, G.; Braka, A.; Wu, S. Quantifying Functional-Group-like Structural Fragments in Molecules and Its Applications in Drug Design. *J. Chem. Inf. Model.* **2023**, *63* (7), 2073–2083. https://doi.org/10.1021/acs.jcim.3c00050.

(60) Ertl, P.; Schuhmann, T. A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products. *J. Nat. Prod.* **2019**, *82* (5), 1258–1263. https://doi.org/10.1021/acs.jnatprod.8b01022.

(61) Ertl, P.; Altmann, E.; McKenna, J. M. The Most Common Functional Groups in Bioactive Molecules and How Their Popularity Has Evolved over Time. *J. Med. Chem.* **2020**, *63* (15), 8408–8418. https://doi.org/10.1021/acs.jmedchem.0c00754.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

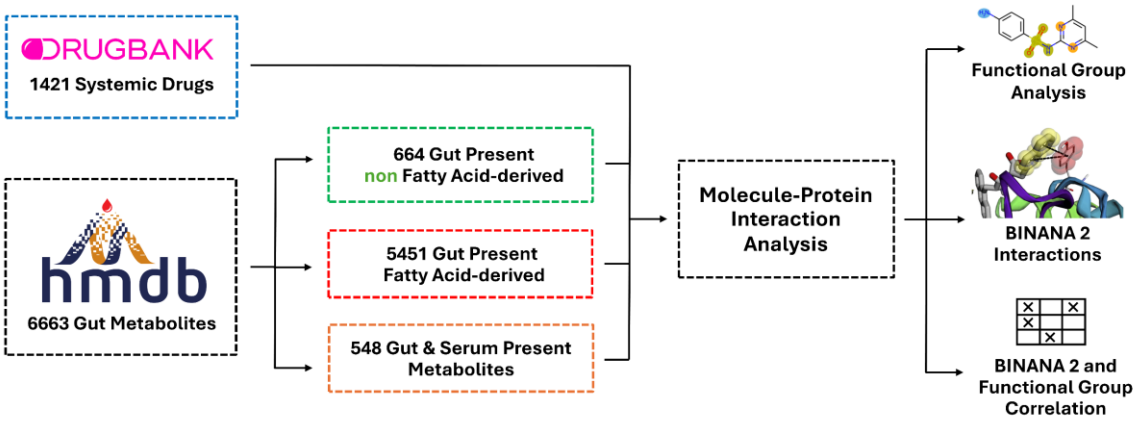
FUNDING SOURCES AKNOWLEDGEMENT

Grant PID2021-127318OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe". Both AS-R and MA have been hired under this project.

AS-R also 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, and a predoctoral grant (PIPF-2022/SAL-GL-26278). . A.D.A. was funded by Fundación Dadoris.

https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

TABLE OF CONTENTS GRAPHIC



https://doi.org/10.26434/chemrxiv-2024-q30x2 **ORCID:** https://orcid.org/0000-0002-8249-4547 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0