A Systematic Analysis and Prediction of the Target Space of Bioactive Food Compounds: Filling the Chemobiological Gaps

Andrés Sánchez-Ruiz & Gonzalo Colmenarejo*

Biostatistics and Bioinformatics Unit

IMDEA Food

CEI UAM+CSIC

E28049 Madrid, Spain

*Corresponding Author

e-mail: gonzalo.colmenarejo@imdea.org

ABSTRACT

Food compounds and their molecular interactions are crucial for health and provide new chemotypes and targets for drug and nutraceutic design. Here we retrieve and analyze the complete set of published interactions of food compounds with human proteins, using the FooDB as compound set and ChEMBL as source of interactions. The data is analyzed in terms of 19 target classes and 19 compound classes, showing a small fraction of target assignment of the compounds (1.6%) and unraveling multiple gaps in the chemobiological space for these molecules. By using well established cheminformatic approaches (Similarity Ensemble Approach (SEA) combined with the maximum Tanimoto Coefficient to the nearest bioactive, "SEA + TC") we achieve a much enhanced target assignment (64.2%), filling many of the gaps with target hypothesis for fast focused testing. By publishing these datasets and analyses we expect to provide a set of resources to speed up the full clarification of the chemobiological space of food compounds, opening new opportunities for drug and nutraceutic design.

KEYWORDS: food compounds, chemical biology, cheminformatics, mode of action, drug design, nutraceutics

INTRODUCTION

There is currently a large research effort in the identification of the biological mechanisms of action of food compounds from a molecular point of view, in order to find novel nutraceuticals and scaffolds for drug design,[1–8] as well as understanding the beneficial or harmful effects of foods on human health.[9–18] A paradigm of this is caffeine, that has been applied as template to develop adenosine receptor antagonists and cyclic nucleotide phosphodiesterase inhibitors, among other biological targets.[18] For that aim, biochemical and/or biological (cellular) assays directed towards different biological targets (typically proteins) are being conducted, allowing specific protein-food compound interactions to be identified. However, these efforts are restricted to a reduced set of targets and compounds, and the assays performed (and therefore, targets tested) are frequently selected in a more or less *ad hoc* basis (many times based on target families and compounds previously studied), instead of being based on a systematic, evidence-based ranking of the whole set of targets in the human genome. In this respect, an obvious, albeit slow and costly approach would be the all vs all high-throughput testing of a comprehensive set of food compounds against a large representation of the human targets. Although in principle this "industrialized" pipeline is feasible, as has been demonstrated in the pharmaceutical and biotechnological R&D sectors,[19–21] it would require a huge investment and coordination of initiatives. An alternative approach would tap from cheminformatic models that can predict compound-target interactions based solely on the compound structure.[22–25,25–30] This approach, based on models trained in large chemobiological databases,[31,32] would provide a much cheaper and faster approach by prioritizing the interactions to test. By openly providing to the experimental research community these predictions, it will be possible for the multiple laboratories worldwide assaying food compounds to re-prioritize their assays in such a way that the identification of mechanism of action will be much accelerated. Similar approaches have been used recently to systematically find unexpected bioactivities of known drugs,[25] as well as

to provide target predictions to a large fraction of the ZINC database[33] of purchasable compounds.[34]

Our group is interested in the structure-activity analysis and modeling of food compounds as sources for the design of novel chemotypes for drugs and nutraceutics, by applying cheminformatic methods from the drug discovery field. In this regard, we have recently applied cheminformatic tools to identify putative interference substructures and aggregators in food compounds.[35,36] However, a crucial part of this effort is the reliable identification of known and predicted interactions with human targets and therefore, in the current work, we aim at conducting a systematic analysis of food compound vs human target interactions. In a first stage, we identify and analyze all the published experimental data for food compounds present in chemobiological databases (ChEMBL[31]). As source of food compounds we use the FooDB,[37] a comprehensive database of these molecules that comprises 70855 structures and is a subset of the Human Metabolome Database.[38] For comparison purposes, a similar data retrieval is performed also with the subset of small molecules in approved, not-withdrawn, and non-illicit status of the DrugBank[39] (2154 molecules). Then we follow this data collection with an analysis of the patterns in the data: chemobiological space covered, target classes and compound classes understudied vs well characterized, targets and compounds unique for food compounds vs shared with drugs. Afterwards, we use well established cheminformatic statistical models (the Similarity Ensemble Approach (SEA), combined with the maximum Tanimoto Coefficient (TC), labelled as "SEA+TC")[22,34] to predict interactions with human targets, and analyze the results as well in terms of both target classes and compound classes, favored vs disfavored combinations, enriched scaffolds, and in comparison with the published data. Finally, we provide examples of experimental validation for several of the predicted interactions, and use molecular docking for structural validation of them. We hope that this work, by openly sharing the data retrieved, will serve, on one hand, to provide a set of known and predicted interactions of food compounds with human targets, which will help in the easy prioritization of testing efforts of these

compounds to gain knowledge about their chemical biology. On the other hand, the analyses here performed will provide hints about the patterns found for the chemical biology of these compounds, as compared to those of drugs, which will help in the identification of new opportunities for drug and nutraceutic design.

RESULTS

Analysis of published bioactivities of food compounds

In a first stage, we performed a search of all the biological activities reported for food compounds in the FooDB. For that, we queried ChEMBL, a manually curated database of the bioactivities published in a large set of journals and maintained by the European Bioinformatics Institute. We used the 29th release, comprising a total of 2.1 million compounds, 14.5 K targets, and 18.6 million activities. From here, a total of 4472 unique interactions with food compounds were identified, arising from 1138 compounds and 759 target proteins. All of them are provided in Supporting Information Table S1. The distributions of both compounds per target, and targets per compound, display long right tails, with an average of 3.9 targets per compound (median of 2), and 5.9 compounds per target (median of 2). The compound with the largest number of targets was quercetin, with 84 targets, followed by ellagic acid, with 61. Both of them display interactions mainly with multiple kinases, targets in the "Others" group (see below), and lyases. These could correspond to promiscuous compounds. On the other hand, the target with the largest number of compounds was Prelamin-A/C, with 145 compounds, followed by the Thyroid stimulating hormone receptor, with 114 compounds. On average, we could find experimental target data (one or more targets) for only 1.6% of the FooDB molecules.

For comparison purposes, the same search was performed with compounds from the DrugBank. This gave a much larger total number of 10005 unique interactions, from 1230 unique compounds and 1296 human proteins, in spite of starting from a much more reduced compound set. In fact, this corresponds to target data (at least one assigned target) for about 57.8% of the DrugBank. Again, the distributions of targets per

compound and compounds per target are long right tailed, with a mean of 8.1 targets per compound (median of 3), and 7.7 compounds per target (median of 3). The compound with the largest number of targets is fedratinib (279 targets), followed by sunitinib (273). Both are competitive inhibitors of kinases, binding to its ATP binding pocket,[40,41] and the vast majority of their targets in ChEMBL are kinases; the large number of hits can be explained by the large conservation of the ATP binding site in this protein class, where is very frequent for this type of inhibitors to show pan-kinase activity. Regarding targets, the one with the largest number of compounds is again Prelamin-A/C, with 299 compounds, followed by Cytochrome P450 3A4, with 149. The latter is one of the cytochromes responsible for xenobiotics metabolism, and because of this has been studied and tested extensively, and therefore it is not surprising to find so many interactions with it. In addition, this cytochrome displays a very large and flexible active site, which accepts a wide variety of chemical diversity and physicochemical profiles (shows promiscuous ligand binding); as a matter of fact, it metabolizes ~75% of the drugs.[42] By considering the targets forming interactions with both compound sets (shared targets), a total of 571 targets were identified, corresponding to ~75% of the total targets interacting with food compounds, and ~50% of the total targets interacting with drug compounds. As regarding the shared compounds, a total of 203 compounds are shared between the two compound sets.

If we analyze the data by target classes, using the ChEMBL hierarchy of targets, we obtain the results displayed in Figure 1, where the number of unique targets with identified interaction is shown for the FooDB, the DrugBank, and the intersection of both groups. In all the target classes, the number of targets found is similar or slightly higher for drugs, with no target class showing more targets in FooDB than in DrugBank.

However, the class of kinases, being the most abundant for DrugBank compounds, shows an exceedingly higher number of targets in DrugBank, in comparison with FooDB. This could reflect the recent explosion in kinase drug discovery and development, especially in the area of cancer therapy,[43,44] that has focused a lot of effort in the last decade in this group of targets,[45] while research in the chemical biology of food compounds has remained focused in other target classes. In terms of the relative member sizes of target classes, the five most abundant ones in food compounds, in decreasing order, are: "Other" (139) > "7TM1" (105) > "Kinase" (86) > "LGIC" (56) > "Oxidoreductase" (51). In turn, the five most abundant target classes for drug compounds show this ordering: "Kinase" (417) > "Other" (217) > "7TM1" (142) > "LGIC" (59) > "Transferase" (57).

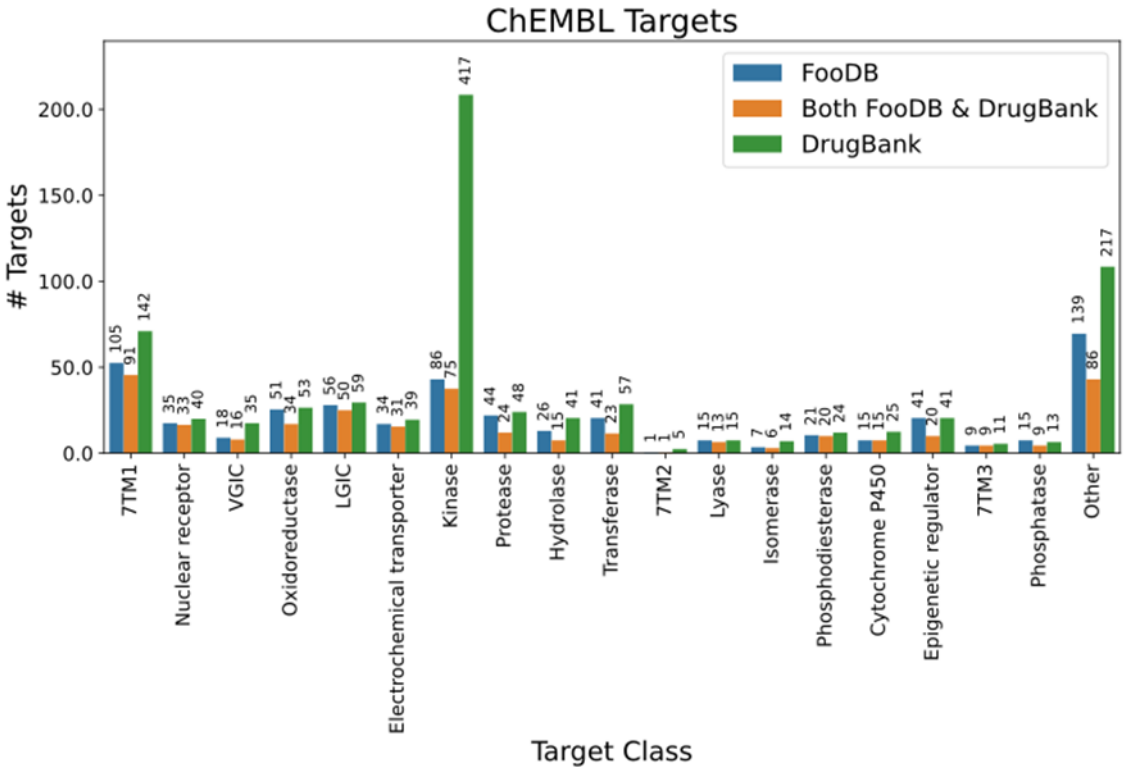


Figure 1. *Distribution of targets among target classes for interactions published for compounds in FooDB (blue bars), DrugBank (green bars), and their intersection*

(orange bars). Interactions retrieved from ChEMBL as described in Methods. The target class hierarchy of ChEBML has been used, arranged in the 19 classes in the plot (abscissa axis). 7TM, seven transmembrane class; VGIC, voltage-gated ion channel; LGIC, ligand-gated ion channel. Each target has been assigned a unique target class in the few cases where more than one was available in the hierarchy.

If, on the other hand, we focus on the distribution of *compounds* by the class of the interacted target, we get the barplot shown in Figure 2. Here, one compound can potentially interact with more than one target class; in addition, each compound is counted only once per target class, in spite being able to interact with several proteins in the same class. In this case, there are some target classes with more compounds in the FooDB than in DrugBank: namely “Oxidoreductase”, “Hydrolase”, “Lyase”, “Phosphodiesterase”, and “Phosphatase”. The five target classes with the largest number of food compounds are, in decreasing order: “Other” (522) > “7TM1” (295) > “Oxidoreductase” (253) > “Cytochrome P450” (213) > Epigenetic regulator (158); while in the case of drug compounds they are: “Other” (649) > “7TM1” (474) > “Cytochrome P450” (337) > “Oxidoreductase” (223) > Electrochemical transporter (216).

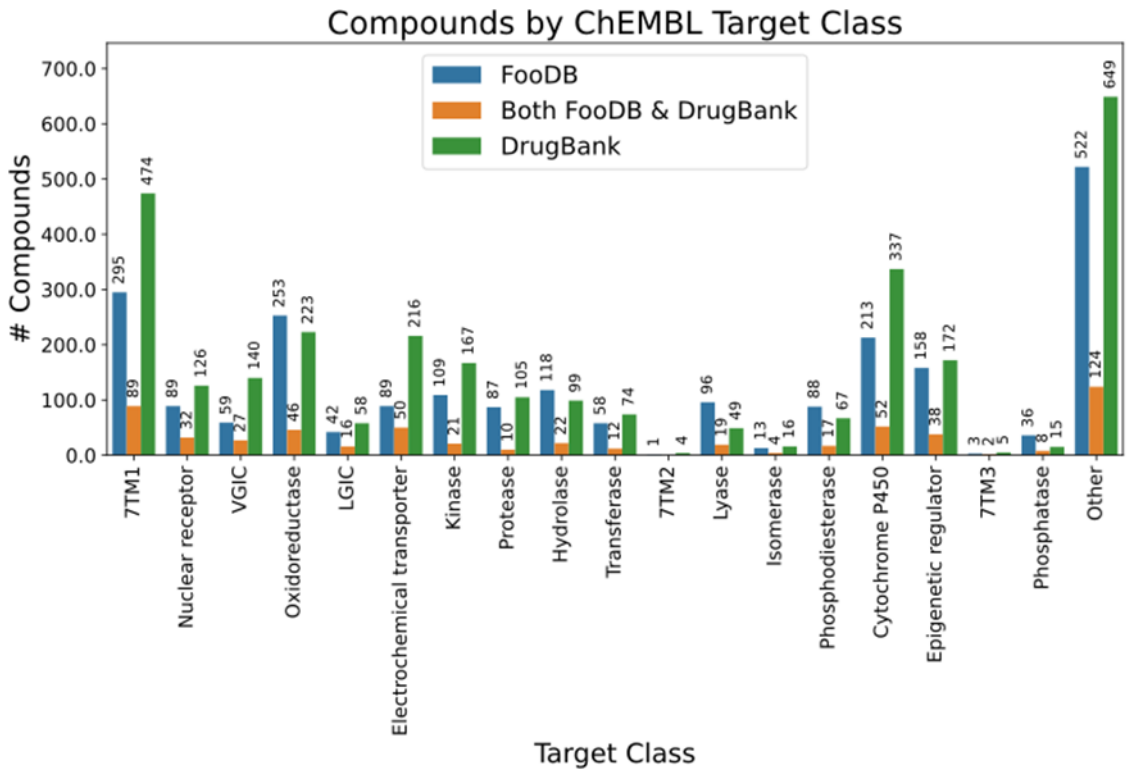


Figure 2. *Distribution of compounds among target classes for interactions published for compounds in FooDB (blue bars), DrugBank (green bars), and their intersection (orange bars). In this case, one compound can interact with multiple targets of the same or different target class, and only one instance of these per target class is counted. Interactions retrieved from ChEMBL as described in Methods.*

It is interesting to further investigate about the targets that show interactions *only* with food compounds. These are 188 human proteins that would correspond to a target space that is less interacted/tested with drug molecules. This could be because of varied reasons: they are less druggable, less relevant for diseases already treated with drugs, or simply less explored in the field of drug discovery. In turn, this can be due to e.g. resources limitation, behavioral ("rich-get-richer") type of phenomena, etc. In the latter case, the fact that they only interact with food molecules so far could be of interest for

using these compounds as source of chemotypes for drug design, or as tool compounds, if these targets become validated for diseases. The corresponding food compounds amount to a total 187 unique structures. Figure 3 displays the distribution of these targets and the corresponding FooDB compounds, vs the corresponding target classes (note that the sum of compound counts is larger than 187 as some of them interact with several target classes). The “Other” target class shows the largest share of targets and compounds. This is a mixed bag that comprises proteins of multiple different sub-families: e.g. transcription factors, enzymes, secreted proteins, membrane receptors, etc. Although this group has the second largest number of drugs from the point of view of established mechanism of actions (MoA, 12.8%), after 7TM1 (27.2%),[45] they correspond to a “non-privileged” target class as each of the subfamilies contain just a few examples of drugs. It is interesting to note that several target classes underrepresented for MoA of approved drugs (“Epigenetic regulator”, “Protease”, “Transferase”, “Hydrolase”, each of them corresponding to < 5% of the drug MoAs[45]) are enriched in the distribution of targets. They point towards a target space relevant for understanding the biological effect of food compounds at a molecular level, as well as for new opportunities for drug discovery as discussed above.

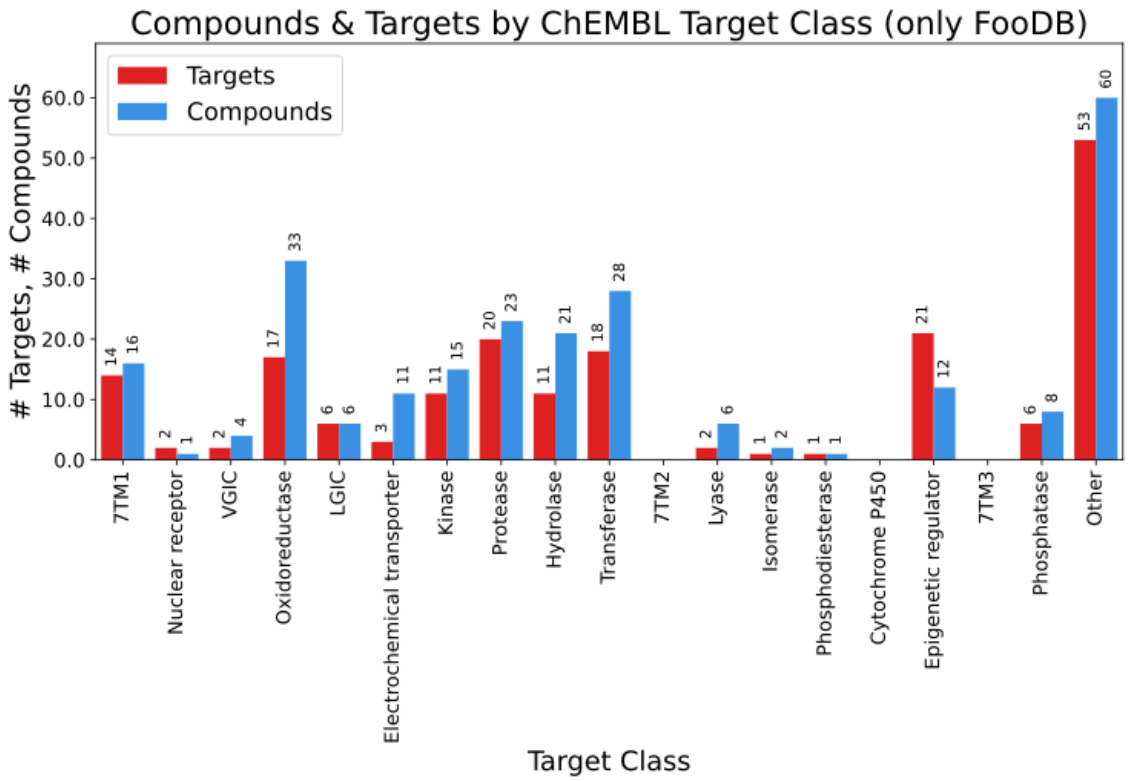


Figure 3. *Distribution of targets (red bars) and compounds (blue bars) among target classes for targets only interacted with FooDB compounds. Interactions retrieved from ChEMBL as described in Methods.*

From this analysis we can conclude that the target space of food compounds is mostly understudied. The vast majority of the food compounds (98.4%), as represented by the FooDB, lack target bioactivity information in ChEMBL, which collects this information from a large set of journals in the field of chemical biology and medicinal chemistry. Thus, in order to provide a better characterization of the target space for these molecules, we used target prediction methodologies to provide a set of putative interactions for a much larger fraction of the FooDB.

Analysis of predicted bioactivities of food compounds

As described in Methods, we used a combination of the SEA algorithm[22] with the maximum Tanimoto Coefficiengt similarity (TC) to provide target predictions for all these compounds. This approach has demonstrated recently improved sensitivity and specificity over the separate methods, as demonstrated through cross-validation analyses of compounds in ChEMBL.[34] Using this methodology, and adopting as thresholds pSEA ≥ 40, TC ≥ 0.4, and pCHEMLB ≥ 6 for food compounds, a total of 88550 interactions were predicted, that corresponded to 451 targets and 44825 compounds; all of them are provided in Supporting Information Table S2. We also provide the full set of predictions for all pSEA, TC and pCHEMBL values, in Supporting Information Table S3, in case the reader wanted to use the data with alternative thresholds. 239 of these interactions were already identified in the previous set of 4472 published interactions in ChEMBL, and thus the combination of them gives a total of 92783 interactions (4472 published + 88311 predicted). As regarding the targets, of the 451 predicted ones, 228 were already in the set of food targets described in the previous section. Thus, the total number of distinct targets of food compounds becomes 982 (759 published targets + 223 new predicted ones). On the other hand, of the 44825 compounds with predicted targets, 496 had at least one reported target in ChEMBL, so that in total, the number of compounds in FooDB with one or more target assignments becomes 45467 (1138 compounds with published targets + 44329 new compounds with predicted targets). This increases the percentage of food compounds with target assignment from the previous 1.6% to 64.2%. This represents a huge set of interaction hypotheses of high likelihood, which can orient the experimental work for a large set of previously unexplored food compounds and targets.

Considering only predicted interactions, the average number of compounds per target is 196.3 (median of 10); the reason for such high number is the presence in the FooDB of a large amount of glycerolipids (> 42K, see below), which are in most cases predicted to interact with LPAR3, FABP3, and, to a lesser extent, LPAR1, due to their structural redundancy. By removing these three targets the average of compounds per target becomes 7.9 (median of 1). Likewise, the average number of targets per compound is 1.97 (median of 2). The largest number is 43 targets per compound, obtained for three compounds (FDB112009, FDB006480, and FDB111642), all of them dipeptides, predicted to interact with different proteases, 7TM1 proteins, and proteins in the "Other" class, among others.

By considering the molecules in the DrugBank, for them a total of 2437 interactions were predicted, arising from 533 targets and 770 compounds. 677 of these interactions were already in the set of 10005 published interactions, and thus their combination makes a total of 11765 (10005 published + 1760 predicted). Focusing on the targets, of the 533 predicted ones, 407 were already seen in the set of drug experimental interactions, and therefore the total number of targets for drugs would rise to 1422 (1296 published + 126 new predicted targets). As regarding the compounds, of the 770 with predicted targets, 554 were within the list of compounds with published interactions; thus, the total number of drug compounds with one or more assigned target would rise to 1446 (1230 compounds with published targets + 216 new compounds with predicted targets), corresponding to a total of 67.9 % of drug compound with target assignment.

The distribution of predicted targets in different target classes can be analyzed as before. Figure 4 displays a barplot for the distribution of predicted targets in the 19

target classes above used, for both FooDB, DrugBank, and their intersection. For FooDB, "7TM1" is the class with the largest number of targets (99), followed by "Others" (76), and "Protease" (74). Then, a large decrease in the number of targets is observed, with 33 for "Oxidoreductase", and 23 for "Nuclear Receptor". The remaining target classes have 20 or less targets each. As regarding DrugBank, the most striking difference with food compounds is again the "Kinase" class, which is the most abundant here (116 targets), while in FooDB there are only 12, one of them shared with DrugBank. After this, the second most abundant class is "7TM1" (112), followed by "Other" (60), "Protease" (51), and "Oxidoreductase" (27), and the rest of classes have 20 or less targets.

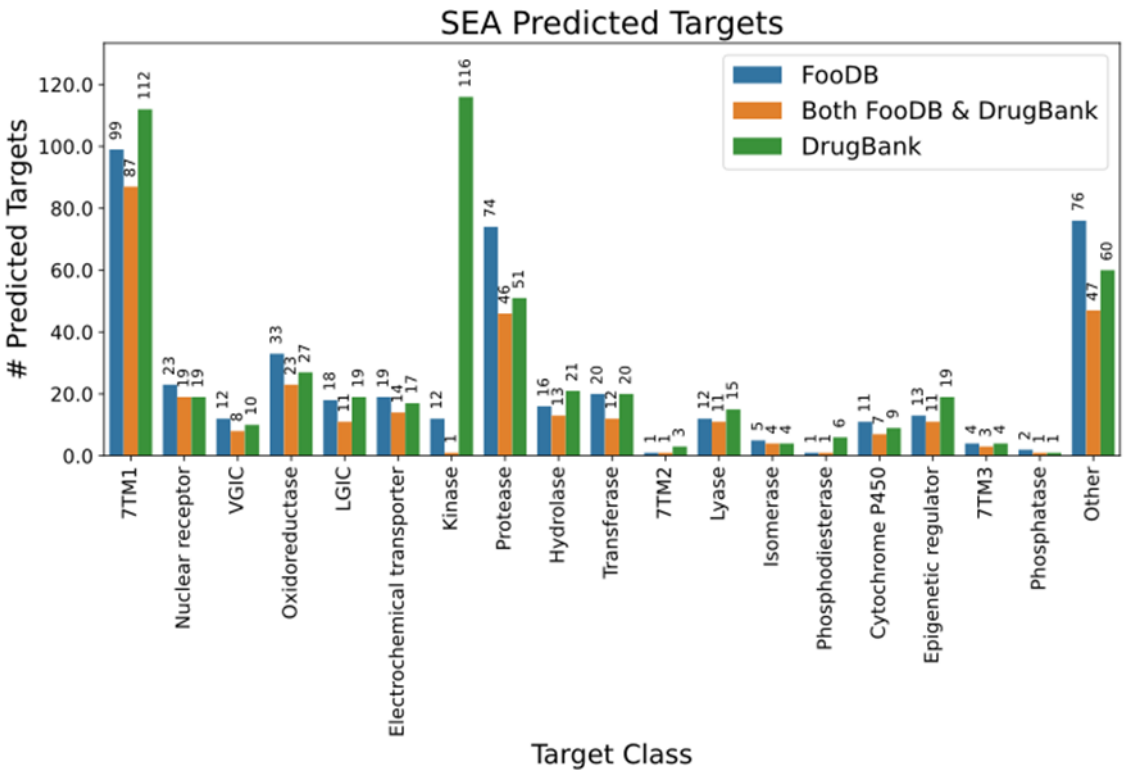


Figure 4. *Distribution of targets among target classes for interactions predicted for compounds in FooDB (blue bars), DrugBank (green bars), and their intersection (orange bars). Interactions predicted with the SEA + TC approach as described in Methods.*

In terms of the distribution of the corresponding compounds for these targets, Figure 5 displays the counts of compounds interacting with the different target classes, again for FooDB, DrugBank, and shared targets. There are two target classes, “7TM1” and “Other”, with huge numbers of FooDB compounds, 40710 the former, and 23407 the later. These correspond mostly to a large number of the glycerolipids above mentioned, that interact with the LPAR3 and LPAR1 targets in the “7TM1” class, and with FABP3 in the “Other” class. Between 600 and 900 compounds are observed for “Electrochemical transporter” (897), “LGIC” (819), “Oxidoreductase” (794), “Hydrolase” (692), “Protease” (644), and “Cytochrome P450” (576). The rest of the classes have less than 300 compounds, going from “Epigenetic regulator“, with 279, down to “7TM2” and “Phosphodiesterase”, both with only one compound.

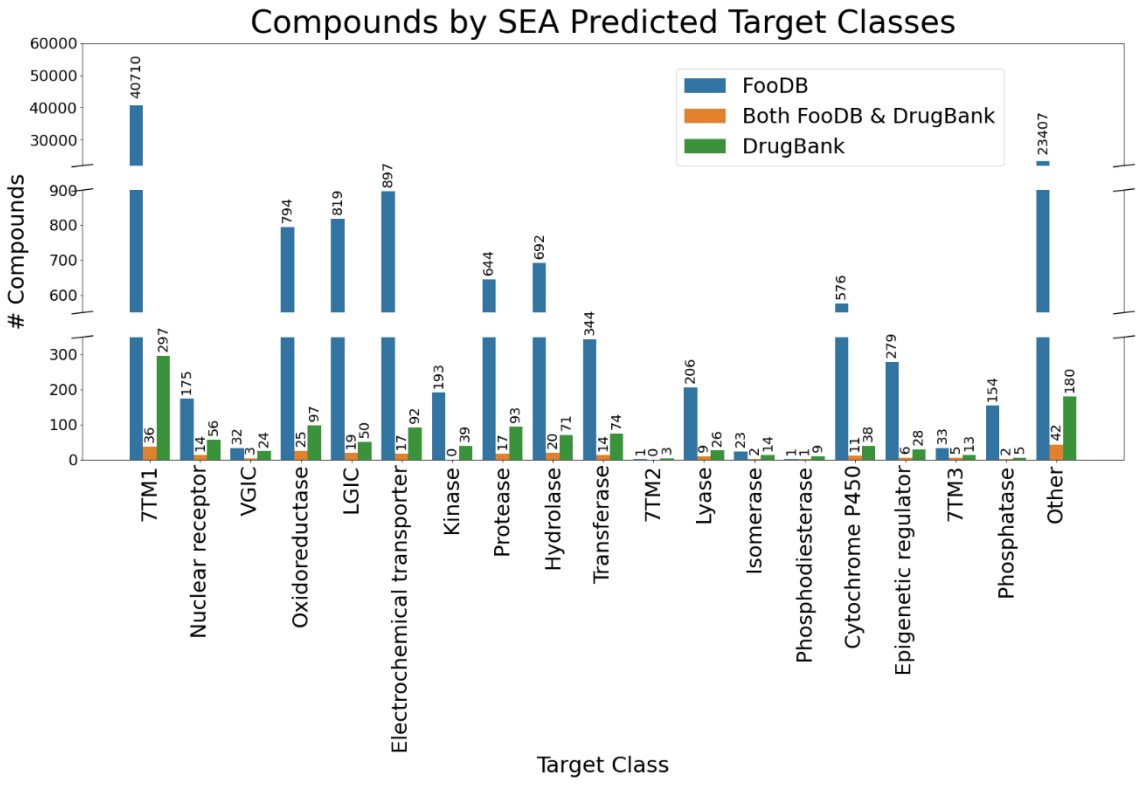


Figure 5. *Distribution of compounds among target classes for interactions predicted for compounds in FooDB (blue bars), DrugBank (green bars), and their intersection (orange*

bars). In this case, one compound can interact with multiple targets of the same or different target class, and only one instance of these per target class is counted. Interactions predicted with the SEA + TC approach as described in Methods.

In the case of DrugBank molecules, the number of compounds per predicted target class has "7TM1" as the most abundant class, with 297 unique compounds, followed in decreasing order by "Other" (180), "Oxidoreductase" (97), "Protease" (93), "Electrochemical transporter" (92), "Transferase" (74), "Hydrolase" (71), and "Nuclear receptor" (56). The rest of the classes display < 50 compounds each.

If we focus on the targets that are uniquely predicted for food compounds, and the corresponding compounds, we obtain the distributions shown in Figure 6. A total of 131 targets are uniquely predicted for food compounds; the latter amount to a total of 1857 structurally unique food compounds (again, the sum of compound counts in Figure 6, blue bars, is larger due to some compounds interacting with several target classes). The largest number of targets are in the "Other" and "Protease" classes, with 29 and 29 proteins, respectively. The following most abundant classes are "7TM1", "Kinase", and "Oxidoreductase", with 12, 11, and 10 targets. The rest of the target classes have less than 10 targets. Of these targets predicted uniquely for food compounds, 18 are within those observed experimentally in the same situation. Therefore, by adding the new ones we get a total of 301 "food-specific" targets (188 published + 113 new predicted). This expands the considerations above mentioned for new targets for drug discovery, as well as for gaining new knowledge on the chemical biology of food compounds, with a set of

novel target hypotheses that can guide the future experimental testing of food compounds and targets.

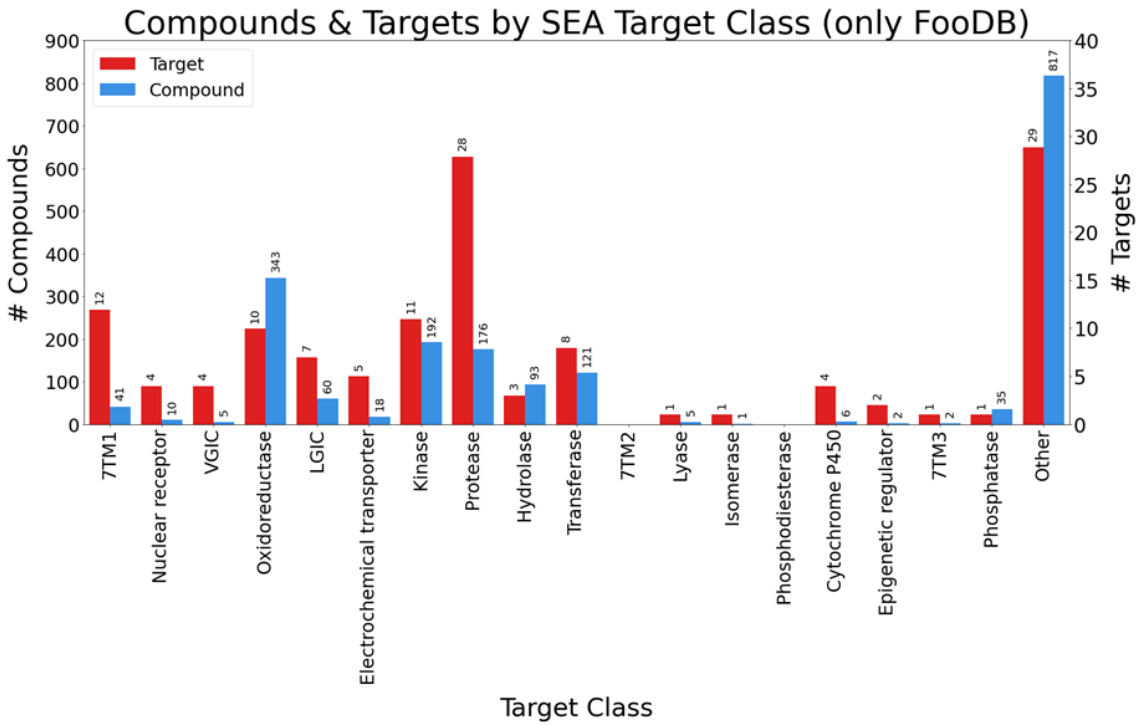


Figure.6. *Distribution of targets (red bars) and compounds (blue bars) among target classes for targets only interacted with FooDB compounds. Interactions predicted with the SEA + TC approach as described in Methods.*

With this augmented set of "food-specific" targets we explored the putative druggability of these proteins by comparing the distribution of their DrugEBIlity scores,[46] a structure-based index of druggability of proteins, vs those of "drug-specific" targets, only targeted by compounds in DrugBank. In this way, although the median score of "drug-specific" targets was larger, the application of a Mann-Whitney U test resulted in a not significant p.value (Figure S1 in Supplementary Information). This result suggests that these "food-specific" targets would be equally druggable, and would open the opportunity for their use in drug discovery should they become validated for particular diseases.

In addition, the corresponding *compounds* can be sources of new chemical diversity for drug discovery and/or as putative tool compounds. The most abundant share of these compounds is by large in the "Others" target class (817), followed by "Oxidoreductase" (343), "Kinase" (192), "Protease" (176) and "Transferase" (121). The rest of the classes have less than 100 compounds associated each. By comparing these structures with the 187 published ones found in ChEMBL (see previous section), we find an intersection of 43 compounds, making a total of 2001 food compounds predicted or known to interact with targets not interacted by drugs (187 published + 1814 new predicted ones).

Analysis of the target space of food compounds by chemotype

Given the large set of published and predicted interactions for food compounds above described, comprising a total of 92783 unique interactions, it seemed worth analyzing their distribution across *chemical* classes. From the Human Metabolome Database (HMDB),[38] of which the FooDB is a subset, we derived a classification of the food compounds in 19 classes that could be applied to the vast majority of FooDB compounds (only 5737 missed classification in the HMDB, being included here in a so-called "Unknown" class). Figure 7 displays the distribution of these classes ordered by decreasing member sizes. We can see that "Glycerolypids" (mainly acylglicerols) is by far the most abundant class of compounds (> 42K). These were identified in our previous work[36] and analyzed separately due to their abundance and structural redundancy. This class is followed by several others with > 1K compounds each, namely and in decreasing order, after the "Unknown" class: "Prenol lipids" (mainly terpene compounds, plus quinones and hydroquinones, 3190), "Phenylpropanoids and polyketides" (including

flavonoids, coumarins, macrolides and tannins, 2755), “Glycerophospholipids” (2749), “Fatty acyls” (2376), “Organoheterocyclic compounds” (2139), “Organic oxygen compounds” (1954), “Organic acids and derivatives” (1892), “Benzenoids” (1256), and “Steroids and steroid derivatives” (1164). The rest of the 19 classes (“Other Lipids”, “Organosulfur compounds”, “Nucleosides, nucleotides, and analogues”, “Lignans, neolignans and related compounds”, “Alkaloids and derivatives”, the mixed bag of “Others”, “Hydrocarbons”, and “Organic nitrogen compounds”) are less populated and have < 1K compounds each.

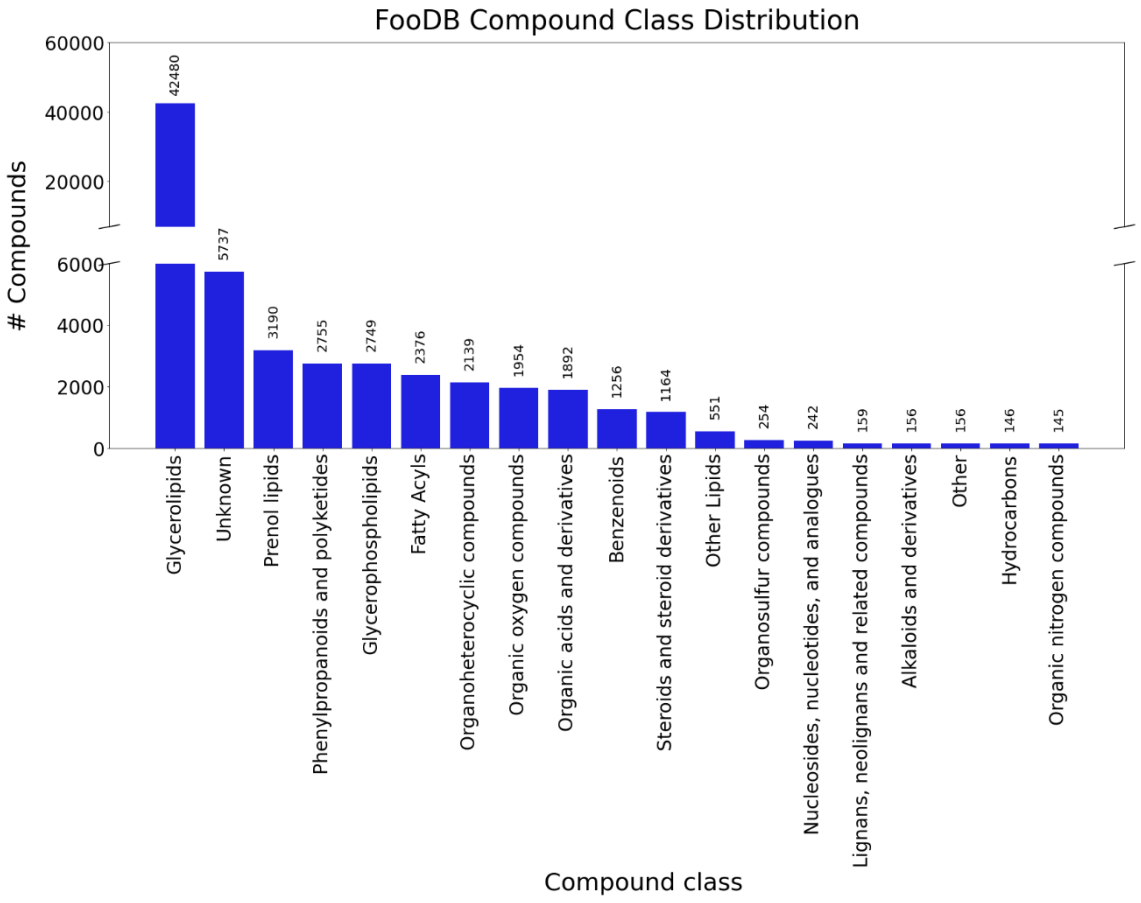


Figure 7. *Distribution of FooDB compounds across compound classes. The 19 compound classes derived from the HMDB as described in Methods.*

If from the set of 4472 published interactions we first plot the 2-way contingency table for compound class vs target class as a heatmap we obtain Figure 8. This corresponds to our current experimental knowledge of the target space of food compounds. In this plot we can see that some compound classes like “Glycerolipids”, “Glycerophospholipids”, “Other lipids”, “Organosulfur compounds”, “Lignans, neolignans and related compounds”, “Other”, “Hydrocarbons” display few (< 50) interactions. On the other extreme, three compound classes, “Phenylpropanoids and polyketides”, “Organoheterocyclic compounds”, and “Benzenoids”, comprise together 67% of the interactions, with counts of 1301, 832, and 848 interactions, respectively. In the case of “Phenylpropanoids and polyketides” the most populated (> 100 interactions) target classes correspond to “Oxidoreductase”, “Kinase”, “Lyase”, “Cytochome P450”, and “Other”. In the case of “Organoheterocyclic compounds” the target classes “7TM1” and “Other” are the ones having > 100 interactions, while in the case of “Benzenoids” they are “7TM1”, “Lyase”, and “Other”. In between these “extreme” compound classes we have the remaining ones with intermediate numbers of interactions. In decreasing number of counts they are (within parenthesis the most frequent target classes): “Steroids and steroid derivatives” (“Other” and “Nuclear receptor”); “Organic acids and derivatives” (“Other” and “7TM1”); “Unknown” (“Other”, “Oxidoreductase”, and “Cytochrome P450”); “Fatty acyls” (“Other”, “7TM1” and “Nuclear receptors”); “Organic oxygen compounds” (“7TM1” and “Other”); “Prenol lipids” (“Other” and “Nuclear receptors”); “Nucleosides, nucleotides and analogues” (“7TM1” and “Other”); “Organic nitrogen compounds” (“Lyase” and “7TM1”); “Alkaloids and derivatives” (“7TM1”, Cythochrome P450” and “Other”). As regards the average number of interactions per

compound, we see that the mean for all the compound classes is of 0.2, but this ranges from 0.000094 in the case of “Glycerolipids”, to 0.67 in the case of “Benzenoids”

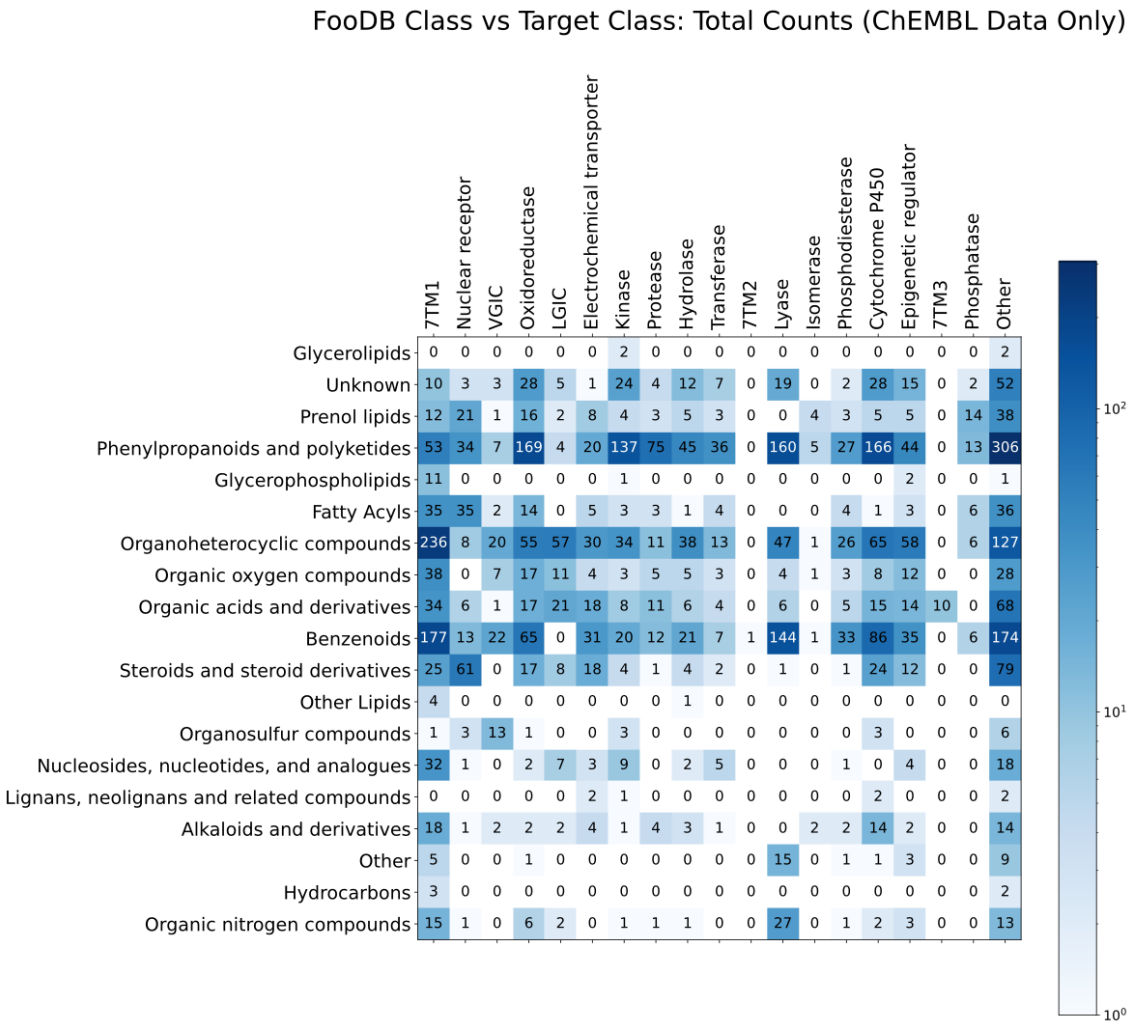


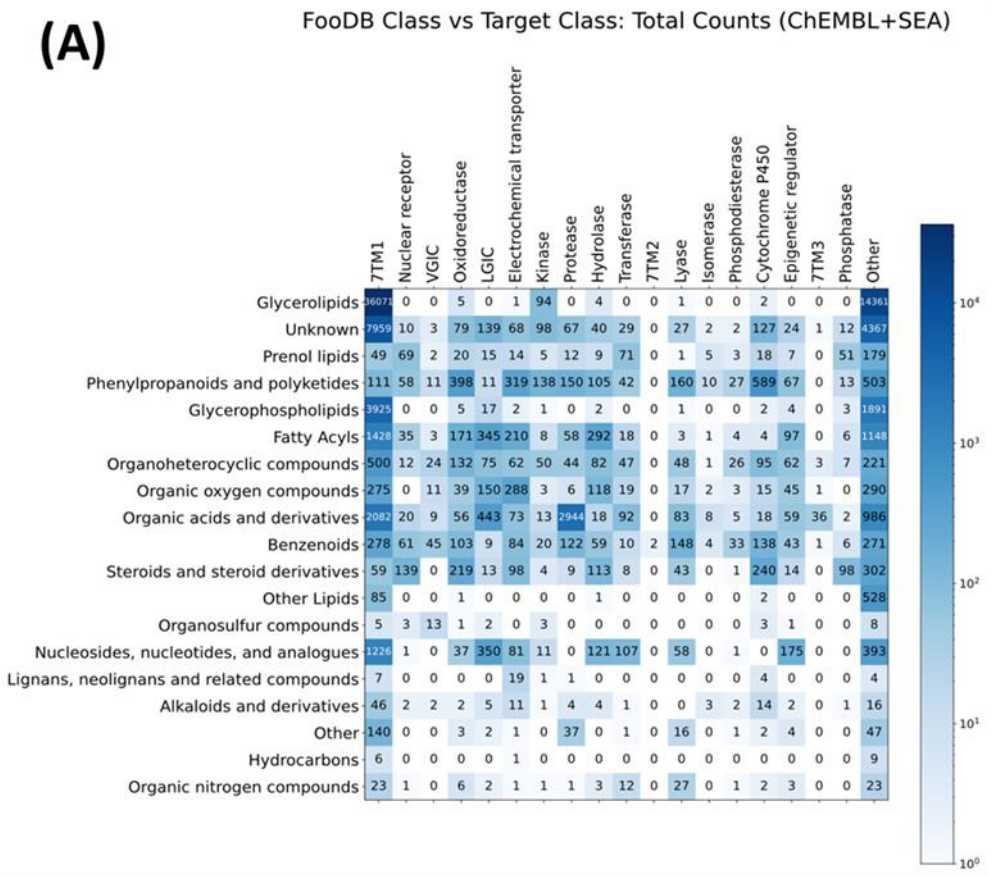
Figure 8. *Distribution of FooDB published interactions across compound classes and target classes. Interactions retrieved from ChEMBL as described in Methods.*

If we add to these published interactions those predicted by SEA + TC, we obtain the distribution displayed in Figure 9A as a heatmap, where multiple of the previously empty or scarcely populated cells appear now with many predicted interactions. We observe variable distributions for the different compound classes, both from the point of view of

the average number of interactions per compound, and from the interacted target classes. By considering the “Glycerolipids” group, the largest compound class, in spite of also having the largest share of interactions we also see a slightly smaller number of interactions per compound than the average of all compound classes (1.19 vs 1.63, respectively). Given that it is by far the most frequent compound class in FooDB, and that both “7TM1” and “Other” are the two most abundant target classes interacting with food compounds, it is not surprising to see that the two largest cell counts in the heatmap are obtained for the corresponding interactions, which represent basically complexes of acylglycerols with the LPAR3, FABP3, and LPAR1 targets described above. On the other hand, the interactions of this compound class with other target classes are comparatively few (< 100 each), resulting in a very unbalanced distribution. Similar unbalanced distributions, with interactions concentrated in “7TM1” and “Other” target classes, are observed for “Glycerophospholipids” and “Other lipids”, and to a lesser extent, for “Fatty Acyls”. For these compound classes, the average number of interactions per compound is 2.12, 1.12, and 1.61, respectively, not very different from “Glycerolipids”. In the case of another crowded group of lipids, “Prenol lipids”, the distribution is much more uniform, and with a much diminished average number of interactions per compound of 0.17, where only the “Other” target class has > 100 interactions. Similar balanced distribution is shown by the next less populated compound class, “Phenylpropanoids and polyketides”, where the “Cytochome P450”, “Other”, “Oxidoreductase”, and “Electrochemical transporter” target classes comprise > 300 interactions each, while the “Lyase”, “Protease”, “Kinase”, “7TM1” and “Hydrolase” have > 100 interactions each. In this case, however, the average number of interactions per compound is higher, of 0.98. In the case of “Organoheterocyclic

compounds", the most abundant interactions are with "7TM1", "Other", and "Oxidoreductase", all of them with > 100 interactions, while the "Organic oxygen compounds" class has five classes ("Other", "Electrochemical transporter", "7TM1","LGIC", and "Hydrolase") with > 100 interactions each, and therefore is more balanced. These two later compound classes have lower average number of interactions per compound, 0.70 and 0.66, respectively. It is worth mentioning the case of "Organic acids and derivatives", which has proportionately about twice as many interactions per compound than the average of all the classes (3.7 vs 1.63, see above): despite being the 9th most numerous compound class, it is the second in terms of number of interactions per compound. It shows > 2000 interactions with targets in the "Protease" and "7TM1" classes, and > 400 with the "Other" and "LGIC" classes; the rest of classes have < 100 interactions. The compound class displaying the largest number of interactions in relation to its number of compounds is "Nucleosides, nucleotides, and analogues", with an average of 10.6 interactions per compound. In this smaller group, with 242 compounds, the target classes "7TM1", "Other", "LGIC", "Epigenetic regulator", "Hydrolase", and "Transferase" all show > 100 interactions each. The "Benzenoids" and "Steroids and steroid derivatives" have similar average number of interactions per compound (1.14 and 1.17, respectively), but their most abundant target classes for interactions are different: in the case of "Benzenoids", "7TM1", "Other", "Lyase", "Cytochrome P450", "Protease", and "Oxidoreductase" have all > 100 interactions, while for "Steroids and steroid derivatives " these correspond to "Other", "Cytochrome P450", "Oxidoreductase", "Nuclear receptor", and "Hydrolase". For the rest of the compound classes, "Organosulfur compounds", "Lignans, neolignans and related compounds", "Alkaloids and derivatives", "Hydrocarbons", "Organic nitrogen

compounds" it is observed a clearly less than the average number of interactions per compound (0.15, 0.23, 0.74, 0.11, 0.73, respectively), in any case showing > 100 interactions for any target class, while the "Other" compound class have an average similar to the total one, 1.63, with "7TM1" displaying > 100 interactions.


(A)
FooDB Class vs Target Class: Total Counts (ChEMBL+SEA)
7TM1
Nuclear receptor
VGIC
Oxidoreductase
LGIC
Electrochemical transporter
Kinase
Protease
Hydrolase
Transferase
7TM2
Lyase
Isomerase
Phosphodiesterase
Cytochrome P450
Epigenetic regulator
7TM3
Phosphatase
Other
Glycerolipids 36071 0 0 5 0 1 94 0 4 0 0 1 0 0 2 0 0 14361
Unknown 7959 10 3 79 139 68 98 67 40 29 0 27 2 2 127 24 1 12 4367
Prenol lipids 49 69 2 20 15 14 5 12 9 71 0 1 5 3 18 7 0 51 179
Phenylpropanoids and polyketides 111 58 11 398 11 319 138 150 105 42 0 160 10 27 589 67 0 13 503
Glycerophospholipids 3925 0 0 5 17 2 1 0 2 0 0 1 0 0 2 4 0 3 1891
Fatty Acyls 1428 35 3 171 345 210 8 58 292 18 0 3 1 4 4 97 0 6 1148
Organoheterocyclic compounds 500 12 24 132 75 62 50 44 82 47 0 48 1 26 95 62 3 7 221
Organic oxygen compounds 275 0 11 39 150 288 3 6 118 19 0 17 2 3 15 45 1 0 290
Organic acids and derivatives 2082 20 9 56 443 73 13 2944 18 92 0 83 8 5 18 59 36 2 986
Benzenoids 278 61 45 103 9 84 20 122 59 10 2 148 4 33 138 43 1 6 271
Steroids and steroid derivatives 59 139 0 219 13 98 4 9 113 8 0 43 0 1 240 14 0 98 302
Other Lipids 85 0 0 1 0 0 0 0 1 0 0 0 0 0 2 0 0 0 528
Organosulfur compounds 5 3 13 1 2 0 3 0 0 0 0 0 0 0 3 1 0 0 8
Nucleosides, nucleotides, and analogues 1226 1 0 37 350 81 11 0 121 107 0 58 0 1 0 175 0 0 393
Lignans, neolignans and related compounds 7 0 0 0 0 19 1 1 0 0 0 0 0 0 4 0 0 0 4
Alkaloids and derivatives 46 2 2 2 5 11 1 4 4 1 0 0 3 2 14 2 0 1 16
Other 140 0 0 3 2 1 0 37 0 1 0 16 0 1 2 4 0 0 47
Hydrocarbons 6 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 9
Organic nitrogen compounds 23 1 0 6 2 1 1 1 3 12 0 27 0 1 2 3 0 0 23
10^4
10^3
10^2
10^1
10^0

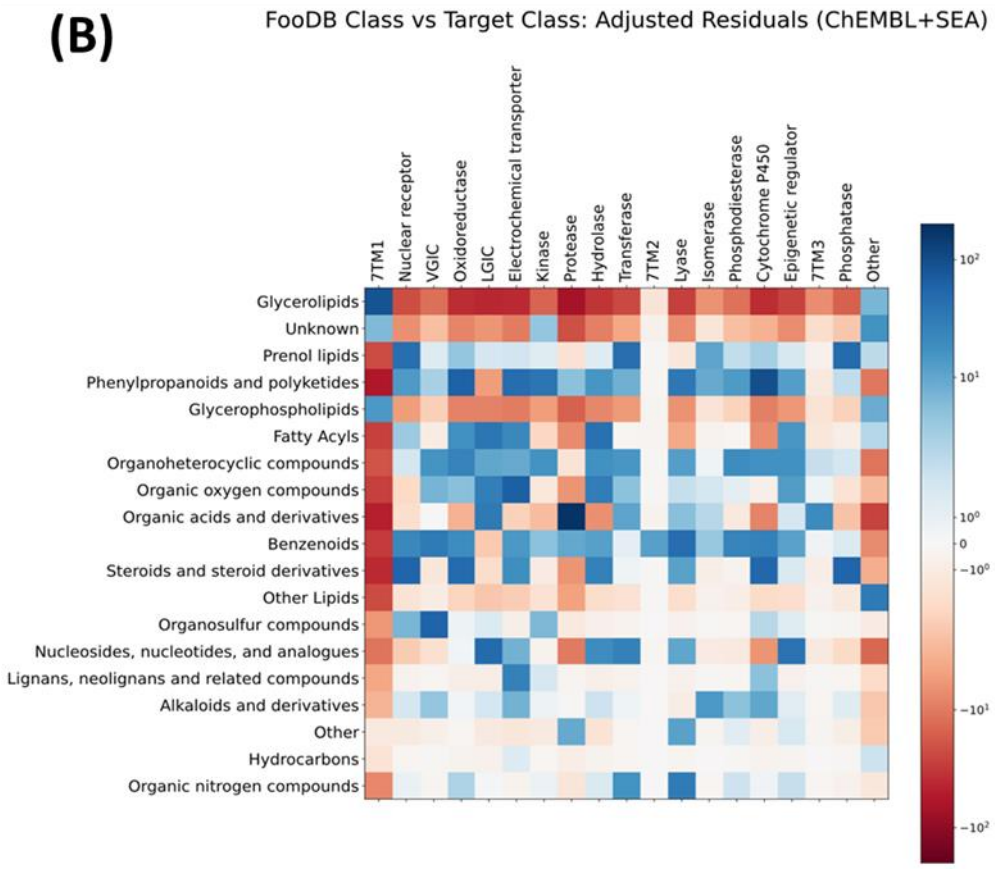

(B)
FooDB Class vs Target Class: Adjusted Residuals (ChEMBL+SEA)
7TM1
Nuclear receptor
VGIC
Oxidoreductase
LGIC
Electrochemical transporter
Kinase
Protease
Hydrolase
Transferase
7TM2
Lyase
Isomerase
Phosphodiesterase
Cytochrome P450
Epigenetic regulator
7TM3
Phosphatase
Other
Glycerolipids
Unknown
Prenol lipids
Phenylpropanoids and polyketides
Glycerophospholipids
Fatty Acyls
Organoheterocyclic compounds
Organic oxygen compounds
Organic acids and derivatives
Benzenoids
Steroids and steroid derivatives
Other Lipids
Organosulfur compounds
Nucleosides, nucleotides, and analogues
Lignans, neolignans and related compounds
Alkaloids and derivatives
Other
Hydrocarbons
Organic nitrogen compounds
10^2
10^1
10^0
0
-10^0
-10^1
-10^2

Figure 9. *Distribution of FooDB published plus predicted (total) interactions across compound classes and target classes. Published interactions retrieved from ChEMBL, and predicted interactions obtained with the SEA + TC approach as described in Methods. (A) Count distribution; (B) adjusted residuals (see Methods). A symmetric logarithmic scale has been used for the color ramp in the latter, with a central linear range from -5 to +5.*

The above analysis is based on the *counts* of unique interactions per compound class and target class. However, it is possible to alternatively base the analysis on the *adjusted residuals* for each cell, so that we can identify cells with counts above- or below-the-expected counts, the expectation being based on the number of compounds and targets interacting in each cell of the contingency table. The use of adjusted residuals, instead of raw or standardized residuals, normalizes also by the very different expected counts that the cells can have, that would result in a bias of cells of larger expected counts to have larger residuals. In this way, we can find combinations of compound classes with target classes highly "favored" or "disfavored", compared with the expected frequency.

Figure 9B shows a heatmap for these adjusted residuals. We can identify, on one hand, a series of compound class vs target class combinations that are much more (blue) or much less (red) frequent than expected. The top five favored interactions are, in decreasing order, the combinations: "Organic acids and derivatives" + "Protease" > "Phenylpropanoids and polyketides" + "Cytochrome P450" > "Glycerolipids" + "7TM1" > "Organic oxygen compounds" + "Electrochemical transporter" > "Phenylpropanoids and polyketides" + "Oxidoreductase". In turn, the top five disfavored interactions are, in

decreasing absolute value, these combinations: “Glycerolipids” + “Protease” > “Phenylpropanoids and polyketides” + “7TM1” > “Organic acids and derivatives” + “7TM1” > “Glycerolipids” + “LGIC” > “Steroids and steroid derivatives” + “7TM1”. Using a lower cutoff for adjusted residuals (> 40), we can find additional favored combinations. The top for the different target classes are: “Steroids and steroid derivatives” and “Prenol lipids” for “Nuclear receptor”; “Organosulfur compounds” for “VGIC”; “Steroids and steroid derivatives” for “Oxidoreductase”; “Nucleosides, nucleotides, and analogues” for “LGIC”; “Phenylpropanoids and polyketides” for “Electrochemical transporter”; “Fatty acyls” for “Hydrolase”; “Prenol lipids” for “Transferase”; “Benzenoids” for “Lyase”; “Steroids and steroid derivatives” for “Cytochrome P450”; “Steroids and steroid derivatives” for “Phosphatase”, and “Prenol lipids” for “Phosphatase”.

Looking with more detail into the structures of the different cells, we can gain some knowledge of the chemotypes enriched in the molecules of the corresponding target class vs chemical class combination as compared with the rest of the molecules. In this way, in Table S5 of the Supplementary Information we display, for each cell, the scaffold (as defined by Bemis and Murcko[47,48]) that shows the lowest p-value in a Fisher exact test for the presence/absence of the scaffold in the cell vs the rest of the molecules. In addition, in Figure 10, for each chemical class the most significant scaffold among all target classes is shown. From here, we can observe a large diversity of chemotypes, with variable amounts of heteroatoms, aromaticity, linker lengths, etc.

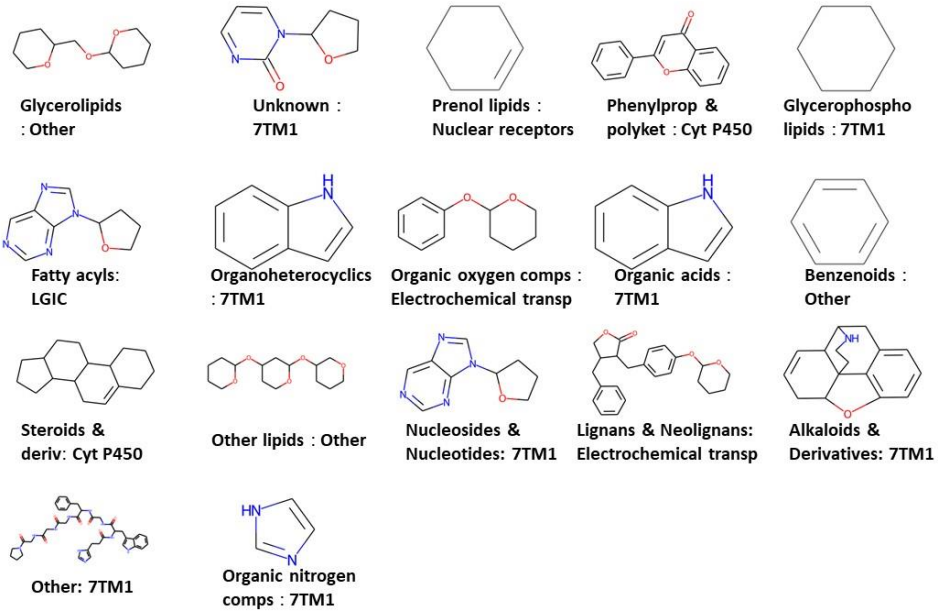


***Figure 10**. Set of most significant Bemis-Murcko scaffolds. For each chemical class, the most significantly enriched scaffold among all target classes is shown, and labeled with the corresponding chemical and target class. Scaffolds for neither hydrocarbons nor organosulfur compounds are shown as these target classes have only cells with < 10 molecules (see Methods).*

A final analysis based on compound classes can focus on the proportion of compounds in each of these classes having target assignment, whether published or predicted. This would give an idea of the relative extent of chemical biology knowledge for the different compound classes, both experimental and after the use of target predictive tools as we have done here. Figure 11 displays the percentages of target assigned compounds for the different compound classes. We can observe very different percentages of assignment. Among the classes with the largest fractions of target assigned compounds are “Glycerophospholips” (90.5%), “Nucleosides, nucleotides, and analogues” (89.7%),

the “Unknown” class (86.7%), and “Other lipids” (82.4%). On the other hand, the classes showing the lowest target assignment are “Organosulfur compounds” (7.5%), “Hydrocarbons” (8.2%), “Prenol lipids” (10.0%), and “Lignans, neolignans and related compounds” (14.5%). These percentages, together with the corresponding number of compounds in the different compound classes, can help to decide on the appropriate experimental approach for the different compound classes: on one hand, agnostic, high-throughput screening approaches (both experimental or virtual) on larger compound classes with little target assignment, given the scarcity of experimental data or even predictions for these compounds; versus focused, hypothesis based screen, for compound classes with large predicted target assignment but low experimental assignment.

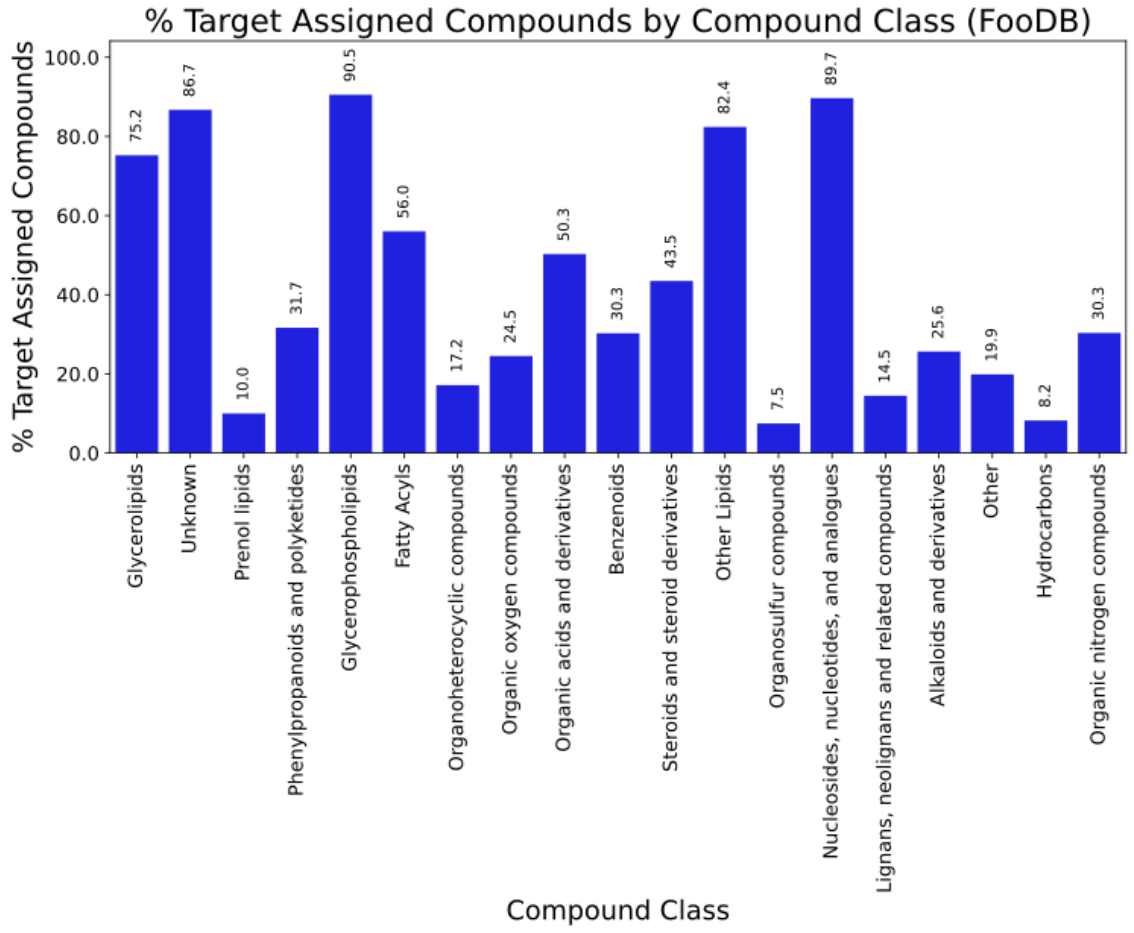


Figure 11. *Distribution of percentages of compounds with assigned targets (published and/or predicted) in FooDB across compound classes*

Experimental validation of the SEA + TC predictions

By taking advantage of the SEA software using data from a former release of ChEMBL (25) compared to the one we used to retrieve the published interactions (29), it was possible to obtain experimental confirmation of some of the predicted interactions for food compounds. In this way, by analyzing the set of 239 interactions both published and predicted, it was possible to observe that the TC value in the SEA + TC output (maximum Tanimoto similarity in the SEA set to the predicted compound) was 1 in the majority of the cases, meaning that the prediction was based (among others, as SEA uses similarities with sets of compounds) in the published interaction for the same compound and target. However, for a subset of 75 compounds the TC value was < 1, which corresponds to predictions based on non-identical compounds. Thus, the existence in the current ChEMBL release of the published interaction means that the prediction performed by SEA + TC based on non-identical compounds has been experimentally validated, giving confidence in the set of predictions for food compounds provided in this work. Figure 12 displays the histogram for the distribution of these 75 predictions, where it can be observed that the confirmed predictions were based on TC values as low as 0.4 (the lowest TC permitted by our settings of the SEA + TC method). The distribution of TC values has a mean and median of 0.70, and an standard deviation of 0.13, which is a rather low similarity

with the predicted compound. The set of 75 confirmed interactions is collected in Table S4 of the Supplementary Information.

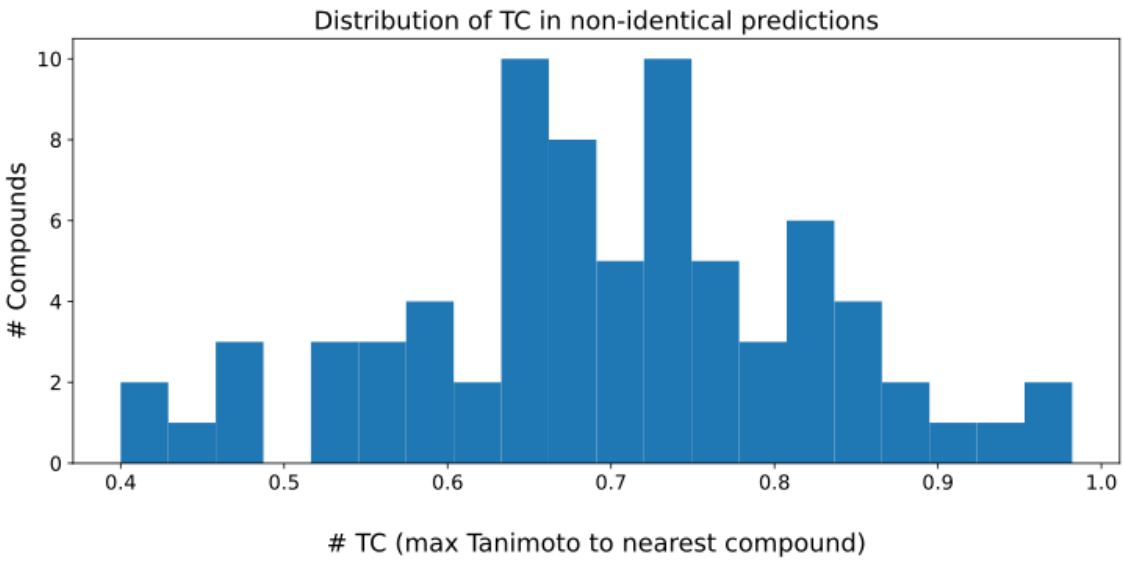


Figure 12. *Histogram for the distribution of TC values of SEA + TC predicted interactions based on non-identical compounds (TC < 1) that are confirmed in ChEMBL 29.*

Validation of the SEA + TC predictions through molecular docking

In addition to the experimental validation of predictions, we investigated the use of molecular docking to confirm and further characterize from a structural point of view some of these interactions. On one hand, we docked FDB002748 (pelargonidin 3-galactoside, a flavonoid glycoside present in fruits like gooseberry and vaccinium), into SLC5A2, the sodium/glucose cotransporter 2. The lowest-energy pose displayed an interaction energy of -9.6 kcal/mol, near that of the empagliflocin original ligand (-10.8 kcal/mol). In addition, it displayed a good fit to the binding pocket (Figure

13A), forming multiple hydrogen bonds with amino acids in the pocket, namely with asparagine 75, serine 287, lysine 321 and glutamine 457.

On the other hand, the docking of FDB022101 (adenylsuccinic acid) to DOT1L, a histone lysil methyltransferase, was also performed as validation test. Here, again, the lowest energy binding pose had good adaptation to the binding pocket and multiple hydrogen bonds were formed: with valine 135, glycine 137, threonine 139, glutamines 168 and 187, and phenylalanine 223 (Figure 12B). The energy value was of -9.7 kcal/mol, similar to that of S-adenosyl methionine, the original ligand (-9.2 kcal/mol).

Thus, we see that these SEA + TC predicted interactions can be explained by reasonable, low energy molecular interactions with the corresponding target, therefore providing additional confidence in the predictions.

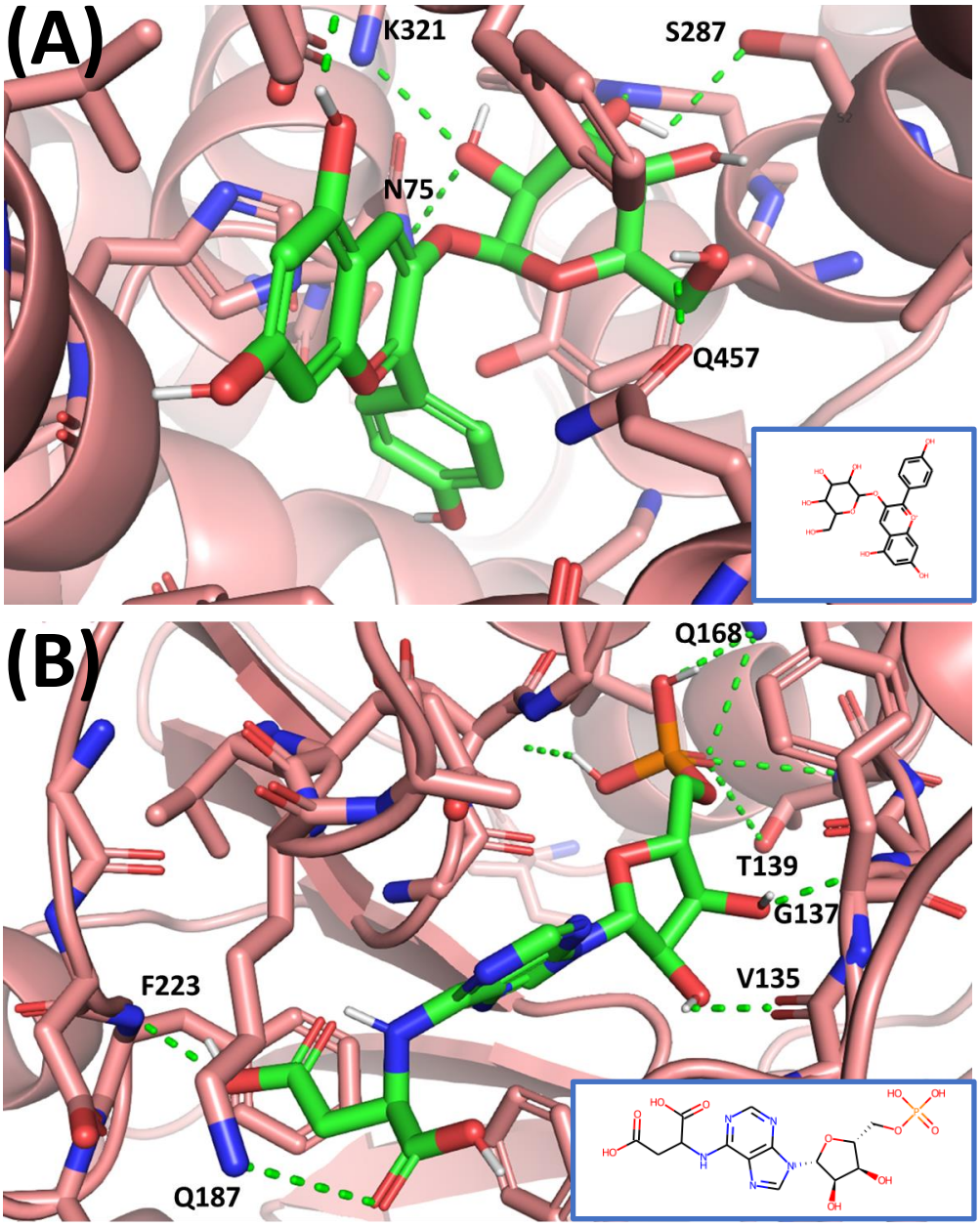

(A)
K321
S287
N75
Q457
(B)
Q168
T139
G137
V135
F223
Q187

Figure 13. *Molecular docking of food interaction predictions. (A) Docking of FDB002748 (pelargonidin 3-galactoside) into SLC5A2 (sodium/glucose cotransporter 2). Protein structure taken from entry 7VSI of PDB. Polar interactions represented as green dashed lines. Inset: chemical structure of FDB002748. (B) Docking of FDB022101 (adenylsuccinic acid) into DOT1L (DOT1 like histone lysine methyltransferase). Protein structure taken from entry 1NW3 of PDB. Polar interactions represented as green dashed lines. Inset: chemical structure of FDB022101.*

DISCUSSION

There is a large interest in understanding the biological effects of food compounds on human health in terms of molecular interactions with endogenous target proteins. Understanding the chemical biology of these important molecules will be highly beneficial in order to rationalize and personalize food consumption. Such knowledge can be exploited to improve human health through preventive approaches, by the design of better diets and the generation and use of appropriate food varieties. On the other hand, food compounds are sources of new chemotypes for drug and nutraceutic discovery and design, as exemplified by caffeine.[18] These molecules are typically safe and well tolerated, as their consumption through thousands of years has demonstrated. In addition, the knowledge about their interactions with human proteins could be useful to identify novel targets for use in the drug discovery field. Moreover, food molecules can provide tool compounds for doing basic or applied research on the targets they interact with.

In spite of this interest, the experimental testing in the area is mostly focused on a reduced number of chemotypes and targets, and the identification of new interactions is done at a slow pace. Similar behavior has been observed in the wider set of natural products,[49] and in kinase inhibitors.[50] Agnostic, high-throughput screening approaches, both in their in vitro and structure-based virtual screening modes, could be used in this area, where a compound collection of food compounds would be screened against a large representation of proteins in the human genome. However, in the current work, a faster and cheaper approach is used which represents an intermediate avenue between the two previous extremes. Here we openly provide the complete set of published interactions for food compounds, as present in the ChEMBL[31,51] chemobiological database and using the FooDB[37] as food compound collection (Supporting Information Table S1). In addition, this set is augmented with a set of predicted interactions through well-established cheminformatic approaches, namely (SEA + TC[22,34]), which is provided as well (Supporting Information Table S2). The combination of these two sets can

speed up the full clarification of the chemical biology of food compounds through diverse ways that we describe in what follows, together with putative applications in the drug discovery field. In the first place, the analysis of published interactions allows finding patterns like compound and target classes well characterized vs understudied ones, so that experimental effort aiming at finding novelties in both chemotypes and targets can focus on the unexplored regions of the chemobiological space. In this respect, we have observed that the target annotation for food compounds is scarce and in fact only a 1.6% of the FooDB (1138 compounds, out of 70855) have one or more significant interaction published with a human target. By analyzing this number by compound classes, we see that groups like “Glycerolipids”, “Glycerophospholipids”, “Other lipids” are extremely understudied, as for example only 2 out of 42480, and 6 out of 2749, and 2 out of 551 compounds, respectively, have at least one published target, while classes like “Organic nitrogen compounds”, “Benzenoids” and “Alkaloids and derivatives” are proportionally the most characterized food compounds, having 14-15% of their members with published targets. In terms of total counts of published interactions, the most populated compound classes are “Phenylpropanoids and polyketides”, “Organoheterocyclic compounds”, and “Benzenoids”, as they together concentrate 67% of all the these interactions.

As regarding the targets for the published interactions, these comprise a set of 759 different proteins, which were grouped in 19 classes. The classes having the largest number of different targets are “Other” (139), “7TM1” (105), and “Kinase” (86), while in terms of unique interactions the most characterized ones are “Other” (975), “7TM1” (709), “Lyase” (423), “Cytochrome P450” (420), and “Oxidoreductase” (410). On the opposite extreme, “7TM3”, “Isomerase”, “7TM2” are the classes with the lowest number of both members and interactions. Looking at the target vs compound class combinations, the most characterized ones are the “Phenylpropanoids and polyketides” with target classes “Other”, “Oxidoreductase”, “Cytochrome P450”, and “Lyase”; “Organoheterocyclic compounds” with “7TM1”; and “Benzenoids” with “7TM1” and “Other”.

In addition, we have also identified a set of 131 targets of food compounds that are not interacted with drugs. The corresponding 187 food molecules interacting with them could be a source of new chemotypes for drug discovery efforts, should the corresponding targets become validated for the treatment of a particular disease. Alternatively, the compounds themselves could be used as tool compounds in assays designed against these targets.

On the other hand, the large set of 88550 predicted interactions provide a much enhanced collection of hypothesis to test experimentally, that together with the published interactions span ~62.4% of the FooDB and for a total of 982 targets. The SEA method used for target prediction is especially powerful for its improved capability of extrapolation in the chemical space,[22,25,34,52,53] by using set-wise similarities together with statistical tests instead of simple compound-wise similarities. It has been used to predict unexpected polypharmacologies of drugs,[22,25] or biological activities in drug inactive ingredients.[53] In addition, its combination with the TC criterion and stringent criteria (pSEA ≥ 40, TC > 0.4, pK_i ≥ 6) as used here has recently shown to improve the method in both sensitivity and specificity,[34] thus making these a set of robust and novel interactions for focused screens. All in all, we provide the full set of predictions for all pSEA, TC and pCHEMBL values, in Supporting Information Table S3, so that the experimentalist could try different thresholds in order to better adapt to different experimental settings: i.e. one that emphasizes having higher precisions, at the cost of lower recalls, or the other way around.

It is clear that the use of target prediction methodologies, always based in modeling previous structure-activity patterns, has the drawback of reducing the probability of finding completely novel or unexpected interactions, which would only be identifiable with fully agnostic high throughput screening approaches. However, as said before, the SEA + TC method used stands out for its previous success in identifying new interactions for compounds relatively dissimilar to those in the training set,[25,52,53] and in addition the low TC used here (0.4) will only discard compounds plainly dissimilar to the training set. As a matter of fact, some of the experimentally

validated interactions corresponded to TC values that low, providing further evidence of the extrapolative capability of the SEA + TC predictive method for this particular dataset.

The resulting set of new predictions fill many holes in the chemobiological space of published interactions, making them high priorities for low-cost, focused testing. In particular, predictions for which there were no published interactions at all are provided for a total of new 42 target class vs compound class combinations (compare Fig 8 vs Fig 9A). Some of the compound classes most benefited are within the most understudied ones: for “Glycerophospholipids” and “Glycerolipids”, interactions in 7 and 6 previously empty compound class vs target class combinations are provided, respectively. In terms of target classes, the most benefited are “Lyase” and “LGIC”, both with predicted interactions in 5 previously empty compound class vs target class combinations. In addition to providing putative interactions for these previously empty combinations, the predictions also vastly enlarge the interaction counts in combinations with a few previous published examples. Some of the most striking cases are the combinations “Glycerolipids” + “Other”, “Unknown” + “Other, and “Organic acids and derivatives“ + “Protease”.

Additionally, the predictions increase the number of targets unique for food compounds to 301, and the corresponding food compounds that interact with them to 1857, thus largely expanding the possibilities above described. The druggability predictions performed over these proteins suggest that these proteins would be no less druggable than the “drug-specific” targets, which is good news for drug discovery.

Besides providing target hypothesis , the predicted interactions were further analyzed to identify putative “favored” vs “disfavored” target vs compound class combinations, as displayed in Fig 9B. The presence of large adjusted residuals in some combinations could be related to the presence in the compound classes of some selective “privileged scaffolds”, as those described from the analysis of ChEMBL data,[54] and for natural products from proprietary screening data.[55] The compound classes used here, however, are not restricted to single scaffolds or

substructures, and therefore other factors could be of importance in this regard. As a matter of fact, the scaffolds identified in the current work (Figure 10 and Table S5) are the most significant ones for each target class vs chemical class combination, and thus provide an initial glimpse on the chemical diversity of the involved chemotypes and their relation to the activity. A thorough analysis, however, is currently being conducted in our group to further analyze the multiple significant scaffolds available in each cell combination.

The adjusted residuals used in Figure 9B come from non-independent instances, since the interactions among different target classes can show some association, thus precluding the calculation of p-values through typical post-hoc analysis of contingency tables.[56,57] They should be considered as some crude approximation to a rigorous estimation of enrichments over the product of marginal probabilities.

The predictions here provided should be taken into account, when tested, in combination with predictions for interference and aggregation in assays which we provided in previous publications.[35,36] It is well known that the presence of some substructures are associated to a tendency in the compounds containing them to give false positive results in assays for different reasons: redox cycling, interference with assay signal, chemical reactivity, etc.[58–63] Similarly, some compounds have shown propensity to form colloidal aggregates that adsorb the tested protein and produce its denaturation, again resulting in a false positive assay reading.[64–66] Food compounds are not without these issues, and therefore these predictions using well known cheminformatic approaches[59,60,67–69] can help to identify these possible issues. By applying appropriate counter assays and orthogonal assays, as described elsewhere,[63] it will be possible to confirm the compound as a true hit or discard it as a false positive.

To conclude, the current work is expected to speed up the complete characterization of the chemobiological space of food compounds, by providing as resources these analyses and datasets, which identify many opportunities for fast focused screens based on robust target hypothesis. As a result, the design of new drugs and nutraceutics based on these compounds

and targets will be accelerated, in addition to multiple applications in the optimization of personalized diets and food varieties to improve health.

METHODS

All the data analyses were conducted with Python 3.9. The RDKit cheminformatic toolkit,[70] version 2021.09.4, was used throughout. As source of food compounds, the latest FooDB was used (70855 molecules), which was kindly provided by Dr Wishart group in SDF format. Structures were processed with the ChEMBL Structure Pipeline, the open source curation pipeline used by ChEMBL,[31,51] as described before.[36,71] As source of drugs, the subset of small molecules in approved, not-withdrawn, and non-illicit status of the DrugBank[39] was used (2154 molecules). Bioactivities were retrieved from ChEMBL 29 by using its Python webresource client,[72] through querying the InChiKey of the compounds. Multiple pCHEMBL values for each InChiKey/target pair were averaged. Afterwards, only pChEBML ≥ 5 values and for human protein targets were kept. For mixtures, both the mixture and the parent compound (obtained through the ChEMBL Structure Pipeline[71]) were queried, and at the end all the bioactivities were aggregated by parent compound.

To obtain target predictions for the different compounds, the Similarity Ensemble Approach (SEA) method, in combination with the maximum Tanimoto Coefficient (TC) to the nearest bioactive, was applied,[22,34] by using the tldr software provided by Irwin and Shoichet labs (tldr.docking.org).[73] This software uses ECFP4 fingerprints[74] to compute similarities. Previously we evaluated other options for target prediction, namely SwissTargetPrediction,[75] PPB2,[76] and PASS,[77] but these software's do not allow the batch processing of molecules required for this dataset (in the case of PASS there is a batch processing version but requires a paid license). An alternative option that allowed batch processing was OCEAN,[78] but it uses a very similar approach to SEA, and in addition the open access version, available through github for installation

(https://github.com/rdkit/OCEAN), is based on a very outdated ChEMBL snapshot (17). On the contrary, the tldr software, a public access service, is based on ChEMBL 25 and is fully functional and supported. In addition, the SEA method has demonstrated capabilities for predicting "unexpected" interactions based on compounds with rather low similarities to the evaluated compound.[22,52,53]

The SEA approach is based on the set-wise chemical similarity between the query compound and the ligands of a target, its distribution being compared through a statistical test to that with a random set of compounds. TC assigns to a query compound the target of the nearest bioactive above a given threshold, and if the largest Tanimoto similarity is below the threshold, no prediction is made. The combined SEA+TC approach used here was recently shown to outperform both SEA itself and a naïve-Bayes classifier in a 5-fold cross-validation of ChEMBL bioactivities.[34] In order to obtain reliable predictions, with a reasonable balance of sensitivity and precision, only predictions with negative log SEA p.value (pSEA) ≥ 40 and TC ≥ 0.4, with a $pK_i \geq 6$, were marked as "hits", as used in ([34]) These thresholds correspond to a precision of the method of 0.19 and a recall of 0.96 from cross-validation analyses (see Figure 1B in 34) which seems appropriate for focused screening efforts. However, to allow the reader to use alternative thresholds depending on different experimental situations, we provide the full SEA output of the tldr software as Table S3 in Supplementary Information.

To calculate druggability scores for "food-specific" targets we used the DrugEBIlity software as implemented in the ChEMBL tractability pipeline,[46,79] which is based on the atomic-resolution structure of the corresponding protein. For comparison purposes, a

similar calculation was performed on the set of "drug-specific" target, and a Mann-Whitney test was performed to test for possible differences in the distributions.

The target class assignment of the targets was based on the PROTEIN_FAMILY_CLASSIFICATION table in ChEMBL. In turn, the assignment of compound class for food compounds was based on the classification provided by the Human Metabolome Database,[38] of which FooDB is a subset.

The *post-hoc* analysis of contingency tables was based on the calculation of cell-specific adjusted residuals. These are defined as:[56]

$$r_{ij} = \frac{n_{ij} - \widehat{\mu_{ij}}}{\sqrt{\widehat{\mu_{ij}}(1 - \frac{n_{i+}}{N})(1 - \frac{n_{+j}}{N})}}$$

Where n_{ij} is the count of the ij-th cell, $\widehat{\mu_{ij}}$ is its expected count, n_{i+} and n_{+j} are the i-th row marginal and j-th column marginals, respectively, and N is the total sample size. The denominator in the previous expression is the standard error of the numerator, so that in this way the raw residual in numerator is normalized to account for very different expected cell values, since otherwise raw residuals for large expected counts would tend to be larger as well.

The scaffold analysis was based on the use of Bemis-Murcko scaffolds[47,48] as implemented in RDKit.[70] For each chemical class vs target class combination, we selected the scaffold yielding the largest odds ratio (lowest right-tailed Fisher test p-value) for the counts of the scaffold in that combination over the rest of the combinations (basically we retrieved the most enriched scaffold in the chemical class vs target class combination as compared to the rest of the molecules). Only combinations with > 10

molecules were considered for the analysis, and for scaffolds displayed by more than two molecules.

Molecular docking analyses were performed with smina.[80] Protein structures were obtained from the Protein Data Bank[81] (PDB) and prepared through AutoDock4[82] receptor preparation scripts. Docking was performed with the default settings for smina and using a binding pocket defined by a box 8 Å around the original ligand in the crystal structure. Vinardo potential[83] was used as scoring function. The convenience of these settings was confirmed by redocking the original ligand and ensuring that the lowest-energy binding pose was reasonably similar to the one in the crystal structure. In the case of SLC5A2 (sodium/glucose cotransporter 2) the PDB entry used was 7VSI, which has as ligand empagliflocin, while the docked ligand was. For DOT1L (DOT1 like histone lysine methyltransferase) the PDB entry was 1NW3 with has S-adenosyl methionine as ligand), and the docked ligand was FDB022101 (adenylsuccinic acid). Molecular visualizations were conducted with Pymol 2.3.4.

DATA AND SOFTWARE AVAILABILITY

RDKit 2021.09.4: https://www.rdkit.org (accessed 2022-02-08; conda install -c conda-forge rdkit).

ChEMBL webresource client: https://github.com/chembl/chembl_webresource_client (accesed 2022-02-08; pip install chembl_webresource_client).

ChEMBL Structure Pipeline: https://github.com/chembl/ChEMBL_Structure_Pipeline (accesed 2022—02-08; conda install -c conda-forge chembl_structure_pipeline).

SEA+TC: https://tldr.docking.org/ (accesed 2022-02-08)

ChEMBL: https://www.ebi.ac.uk/chembl/ (accesed 2022-02-08)

FooDB: https://www.foodb.ca/ (accesed 2022-02-08)

Human Metabolome Database: https://hmdb.ca/ (accesed 2022-02-08)

All the results of the ChEMBL data retrieval and the SEA+TC target prediction are provided as Supplementary Information.

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:

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

The Dr Wishart Research Group (University of Alberta, Edmonton, Canada) is thanked for providing the updated version of FooDB in SDF format. Dr John Irwin and collaborators (University of California, San Francisco, USA) are thanked for providing help with the SEA+TC calculations.

REFERENCES

(1) Herrera-Arozamena, C.; Estrada-Valencia, M.; López-Caballero, P.; Pérez, C.; Morales-García, J. A.; Pérez-Castillo, A.; Sastre, E. del; Fernández-Mendívil, C.; Duarte, P.; Michalska, P.; Lombardía, J.; Senar, S.; León, R.; López, M. G.; Rodríguez-Franco, M. I. Resveratrol-Based MTDLs to Stimulate Defensive and Regenerative Pathways and Block Early Events in Neurodegenerative Cascades. *J. Med. Chem.* **2022**, *65* (6), 4727–4751. https://doi.org/10.1021/acs.jmedchem.1c01883.

(2) Zhu, Y.; Fu, J.; Shurlknight, K. L.; Soroka, D. N.; Hu, Y.; Chen, X.; Sang, S. Novel Resveratrol-Based Aspirin Prodrugs: Synthesis, Metabolism, and Anticancer Activity. *J. Med. Chem.* **2015**, *58* (16), 6494–6506. https://doi.org/10.1021/acs.jmedchem.5b00536.

(3) Gu, C.; Stashko, M. A.; Puhl-Rubio, A. C.; Chakraborty, M.; Chakraborty, A.; Frye, S. V.; Pearce, K. H.; Wang, X.; Shears, S. B.; Wang, H. Inhibition of Inositol Polyphosphate Kinases by Quercetin and Related Flavonoids: A Structure–Activity Analysis. *J. Med. Chem.* **2019**, *62* (3), 1443–1454. https://doi.org/10.1021/acs.jmedchem.8b01593.

(4) Selvaraj, S.; Krishnan, U. M. Vanadium–Flavonoid Complexes: A Promising Class of Molecules for Therapeutic Applications. *J. Med. Chem.* **2021**, *64* (17), 12435–12452. https://doi.org/10.1021/acs.jmedchem.1c00405.

(5) Long, H.; Hu, X.; Wang, B.; Wang, Q.; Wang, R.; Liu, S.; Xiong, F.; Jiang, Z.; Zhang, X.-Q.; Ye, W.-C.; Wang, H. Discovery of Novel Apigenin–Piperazine Hybrids as Potent and Selective Poly (ADP-Ribose) Polymerase-1 (PARP-1) Inhibitors for the Treatment of Cancer. *J. Med. Chem.* **2021**, *64* (16), 12089–12108. https://doi.org/10.1021/acs.jmedchem.1c00735.

(6) Jafari, Z.; Bigham, A.; Sadeghi, S.; Dehdashti, S. M.; Rabiee, N.; Abedivash, A.; Bagherzadeh, M.; Nasseri, B.; Karimi-Maleh, H.; Sharifi, E.; Varma, R. S.; Makvandi, P. Nanotechnology-Abetted Astaxanthin Formulations in Multimodel Therapeutic and Biomedical Applications. *J. Med. Chem.* **2022**, *65* (1), 2–36. https://doi.org/10.1021/acs.jmedchem.1c01144.

(7) Estrela, J. M.; Mena, S.; Obrador, E.; Benlloch, M.; Castellano, G.; Salvador, R.; Dellinger, R. W. Polyphenolic Phytochemicals in Cancer Prevention and Therapy: Bioavailability versus Bioefficacy. *J. Med. Chem.* **2017**, *60* (23), 9413–9436. https://doi.org/10.1021/acs.jmedchem.6b01026.

(8) Sassetti, E.; Clausen, M. H.; Laraia, L. Small-Molecule Inhibitors of Reactive Oxygen Species Production. *J. Med. Chem.* **2021**, *64* (9), 5252–5275. https://doi.org/10.1021/acs.jmedchem.0c01914.

(9) Polya, G. M. *Biochemical Targets of Plant Bioactive Compounds: A Pharmacological Reference Guide to Sites of Action and Biological Effects*; Taylor & Francis: London ; New York, 2003.

(10) Hu, J.; Wang, J.; Gan, Q.; Ran, Q.; Lou, G.; Xiong, H.; Peng, C.; Sun, J.; Yao, R.; Huang, Q. Impact of Red Yeast Rice on Metabolic Diseases: A Review of Possible Mechanisms of Action. *J. Agric. Food Chem.* **2020**, *68* (39), 10441–10455. https://doi.org/10.1021/acs.jafc.0c01893.

(11) Teodoro, A. J. Bioactive Compounds of Food: Their Role in the Prevention and Treatment of Diseases. *Oxidative Medicine and Cellular Longevity* **2019**, *2019*, 1–4. https://doi.org/10.1155/2019/3765986.

(12) Mbachu, O. C.; Howell, C.; Simmler, C.; Malca Garcia, G. R.; Skowron, K. J.; Dong, H.; Ellis, S. G.; Hitzman, R. T.; Hajirahimkhan, A.; Chen, S.-N.; Nikolic, D.; Moore, T. W.; Vollmer, G.; Pauli, G. F.; Bolton, J. L.; Dietz, B. M. SAR Study on Estrogen Receptor α/β Activity of (Iso)Flavonoids: Importance of Prenylation, C-Ring (Un)Saturation, and Hydroxyl Substituents. *J. Agric. Food Chem.* **2020**, *68* (39), 10651–10663. https://doi.org/10.1021/acs.jafc.0c03526.

(13) Perez-Gregorio, R.; Simal-Gandara, J. A Critical Review of Bioactive Food Components, and of Their Functional Mechanisms, Biological Effects and Health Outcomes. *Current Pharmaceutical Design* **2017**, *23* (19), 2731–2741.
(14) Sharifi-Rad, J.; Quispe, C.; Zam, W.; Kumar, M.; Cardoso, S. M.; Pereira, O. R.; Ademiluyi, A. O.; Adeleke, O.; Moreira, A. C.; Živković, J.; Noriega, F.; Ayatollahi, S. A.; Kobarfard, F.; Faizi, M.; Martorell, M.; Cruz-Martins, N.; Butnariu, M.; Bagiu, I. C.; Bagiu, R. V.; Alshehri, M. M.; Cho, W. C. Phenolic Bioactives as Antiplatelet Aggregation Factors: The Pivotal Ingredients in Maintaining Cardiovascular Health. *Oxidative Medicine and Cellular Longevity* **2021**, *2021*, e2195902. https://doi.org/10.1155/2021/2195902.
(15) Sharifi-Rad, J.; Cruz-Martins, N.; López-Jornet, P.; Lopez, E. P.-F.; Harun, N.; Yeskaliyeva, B.; Beyatli, A.; Sytar, O.; Shaheen, S.; Sharopov, F.; Taheri, Y.; Docea, A. O.; Calina, D.; Cho, W. C. Natural Coumarins: Exploring the Pharmacological Complexity and Underlying Molecular Mechanisms. *Oxidative Medicine and Cellular Longevity* **2021**, *2021*, e6492346. https://doi.org/10.1155/2021/6492346.
(16) Gry, J.; Black, L.; Eriksen, F. D.; Pilegaard, K.; Plumb, J.; Rhodes, M.; Sheehan, D.; Kiely, M.; Kroon, P. A. EuroFIR-BASIS – a Combined Composition and Biological Activity Database for Bioactive Compounds in Plant-Based Foods. *Trends in Food Science & Technology* **2007**, *18* (8), 434–444. https://doi.org/10.1016/j.tifs.2007.05.008.
(17) Faudone, G.; Arifi, S.; Merk, D. The Medicinal Chemistry of Caffeine. *J. Med. Chem.* **2021**. https://doi.org/10.1021/acs.jmedchem.1c00261.
(18) Faudone, G.; Arifi, S.; Merk, D. The Medicinal Chemistry of Caffeine. *J. Med. Chem.* **2021**. https://doi.org/10.1021/acs.jmedchem.1c00261.
(19) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nature Reviews Drug Discovery* **2011**, *10* (3), 188–195. https://doi.org/10.1038/nrd3368.
(20) Zeng, W.; Guo, L.; Xu, S.; Chen, J.; Zhou, J. High-Throughput Screening Technology in Industrial Biotechnology. *Trends in Biotechnology* **2020**, *38* (8), 888–906. https://doi.org/10.1016/j.tibtech.2020.01.001.
(21) 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.
(22) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat Biotechnol* **2007**, *25* (2), 197–206. https://doi.org/10.1038/nbt1284.
(23) Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A Network Integration Approach for Drug-Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information. *Nature Communications* **2017**, *8* (1), 573. https://doi.org/10.1038/s41467-017-00680-8.
(24) Yamanishi, Y.; Kotera, M.; Moriya, Y.; Sawada, R.; Kanehisa, M.; Goto, S. DINIES: Drug–Target Interaction Network Inference Engine Based on Supervised Analysis. *Nucleic Acids Research* **2014**, *42* (W1), W39–W45. https://doi.org/10.1093/nar/gku337.
(25) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature* **2012**, *486* (7403), 361–367. https://doi.org/10.1038/nature11159.
(26) Sureyya Rifaioglu, A.; Nalbat, E.; Atalay, V.; Jesus Martin, M.; Cetin-Atalay, R.; Doğan, T. DEEPScreen: High Performance Drug–Target Interaction Prediction with Convolutional Neural Networks Using 2-D Structural Compound Representations. *Chemical Science* **2020**, *11* (9), 2531–2557. https://doi.org/10.1039/C9SC03414E.

(27) Wang, X.; Shen, Y.; Wang, S.; Li, S.; Zhang, W.; Liu, X.; Lai, L.; Pei, J.; Li, H. PharmMapper 2017 Update: A Web Server for Potential Drug Target Identification with a Comprehensive Target Pharmacophore Database. *Nucleic Acids Research* **2017**, *45* (W1), W356–W360. https://doi.org/10.1093/nar/gkx374.

(28) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. *J Cheminform* **2015**, *7* (1), 51. https://doi.org/10.1186/s13321-015-0098-y.

(29) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/C8SC00148K.

(30) Sydow, D.; Burggraaff, L.; Szengel, A.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P.; Volkamer, A. Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.* **2019**, *59* (5), 1728–1742. https://doi.org/10.1021/acs.jcim.8b00832.

(31) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res* **2017**, *45* (Database issue), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(32) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Research* **2021**, *49* (D1), D1388–D1395. https://doi.org/10.1093/nar/gkaa971.

(33) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675.

(34) Irwin, J. J.; Gaskins, G.; Sterling, T.; Mysinger, M. M.; Keiser, M. J. Predicted Biological Activity of Purchasable Chemical Space. *J. Chem. Inf. Model.* **2018**, *58* (1), 148–164. https://doi.org/10.1021/acs.jcim.7b00316.

(35) 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.

(36) 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.

(37) *FooDB*. https://foodb.ca/ (accessed 2021-09-13).

(38) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res* **2018**, *46* (D1), D608–D617. https://doi.org/10.1093/nar/gkx1089.

(39) 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.; Iynkkaran, 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.

(40) Davis, R. R.; Li, B.; Yun, S. Y.; Chan, A.; Nareddy, P.; Gunawan, S.; Ayaz, M.; Lawrence, H. R.; Reuther, G. W.; Lawrence, N. J.; Schönbrunn, E. Structural Insights into JAK2 Inhibition by Ruxolitinib, Fedratinib, and Derivatives Thereof. *J. Med. Chem.* **2021**, *64* (4), 2228–2241. https://doi.org/10.1021/acs.jmedchem.0c01952.

(41) McTigue, M.; Murray, B. W.; Chen, J. H.; Deng, Y.-L.; Solowiej, J.; Kania, R. S. Molecular Conformations, Interactions, and Properties Associated with Drug Efficiency and Clinical Performance among VEGFR TK Inhibitors. *PNAS* **2012**, *109* (45), 18281–18289. https://doi.org/10.1073/pnas.1207759109.

(42) Di, L.; Kerns, E. H. *Drug-like Properties: Concepts, Structure Design and Methods: From ADME to Toxicity Optimization*, Second edition.; Elsevier/AP: Amsterdam ; Boston, 2016.

(43) Cohen, P.; Cross, D.; Jänne, P. A. Kinase Drug Discovery 20 Years after Imatinib: Progress and Future Directions. *Nat Rev Drug Discov* **2021**, *20* (7), 551–569. https://doi.org/10.1038/s41573-021-00195-4.

(44) Bhullar, K. S.; Lagarón, N. O.; McGowan, E. M.; Parmar, I.; Jha, A.; Hubbard, B. P.; Rupasinghe, H. P. V. Kinase-Targeted Cancer Therapies: Progress, Challenges and Future Directions. *Molecular Cancer* **2018**, *17* (1), 48. https://doi.org/10.1186/s12943-018-0804-2.

(45) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A Comprehensive Map of Molecular Drug Targets. *Nat Rev Drug Discov* **2017**, *16* (1), 19–34. https://doi.org/10.1038/nrd.2016.230.

(46) *tractability_pipeline_v2/ot_tractability_pipeline_v2 at master · chembl/tractability_pipeline_v2*. GitHub. https://github.com/chembl/tractability_pipeline_v2 (accessed 2022-07-08).

(47) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. https://doi.org/10.1021/jm9602928.

(48) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42* (25), 5095–5099. https://doi.org/10.1021/jm9903996.

(49) Bisson, J.; McAlpine, J. B.; Friesen, J. B.; Chen, S.-N.; Graham, J.; Pauli, G. F. Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery? *J. Med. Chem.* **2016**, *59* (5), 1671–1690. https://doi.org/10.1021/acs.jmedchem.5b01009.

(50) Fedorov, O.; Müller, S.; Knapp, S. The (Un)Targeted Cancer Kinome. *Nat Chem Biol* **2010**, *6* (3), 166–169. https://doi.org/10.1038/nchembio.297.

(51) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Research* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(52) Gregori-Puigjané, E.; Setola, V.; Hert, J.; Crews, B. A.; Irwin, J. J.; Lounkine, E.; Marnett, L.; Roth, B. L.; Shoichet, B. K. Identifying Mechanism-of-Action Targets for Drugs and Probes. *Proceedings of the National Academy of Sciences* **2012**, *109* (28), 11178–11183. https://doi.org/10.1073/pnas.1204524109.

(53) Pottel, J.; Armstrong, D.; Zou, L.; Fekete, A.; Huang, X.-P.; Torosyan, H.; Bednarczyk, D.; Whitebread, S.; Bhhatarai, B.; Liang, G.; Jin, H.; Ghaemi, S. N.; Slocum, S.; Lukacs, K. V.; Irwin, J. J.; Berg, E. L.; Giacomini, K. M.; Roth, B. L.; Shoichet, B. K.; Urban, L. The Activities of Drug Inactive Ingredients on Biological Targets. *Science* **2020**, *369* (6502), 403–413. https://doi.org/10.1126/science.aaz9906.

(54) Schneider, P.; Schneider, G. Privileged Structures Revisited. *Angewandte Chemie International Edition* **2017**, *56* (27), 7971–7974. https://doi.org/10.1002/anie.201702816.

(55) Coma, I.; Bandyopadhyay, D.; Diez, E.; Ruiz, E. A.; de los Frailes, M. T.; Colmenarejo, G. Mining Natural-Products Screening Data for Target-Class Chemical Motifs. *J Biomol Screen* **2014**, *19* (5), 749–757. https://doi.org/10.1177/1087057114521463.

(56) 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.
(57) Sharpe, D. Chi-Square Test Is Statistically Significant: Now What? *Practical Assessment, Research, and Evaluation* **2019**, *20* (1). https://doi.org/10.7275/tbfa-x148.
(58) Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature News* **2014**, *513* (7519), 481. https://doi.org/10.1038/513481a.
(59) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 897–902. https://doi.org/10.1021/ci990423o.
(60) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740. https://doi.org/10.1021/jm901137j.
(61) Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chem. Biol.* **2018**, *13* (1), 36–44. https://doi.org/10.1021/acschembio.7b00903.
(62) J., F. B. Identification and Evaluation of Molecular Properties Related to Preclinical Optimization and Clinical Fate. *Medicinal Chemistry* **2005**, *1* (6), 649–655.
(63) Aldrich, C.; Bertozzi, C.; Georg, G. I.; Kiessling, L.; Lindsley, C.; Liotta, D.; Merz, K. M.; Schepartz, A.; Wang, S. The Ecstasy and Agony of Assay Interference Compounds. *ACS Cent. Sci.* **2017**, *3* (3), 143–147. https://doi.org/10.1021/acscentsci.7b00069.
(64) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45* (8), 1712–1722. https://doi.org/10.1021/jm010533y.
(65) Owen, S. C.; Doak, A. K.; Ganesh, A. N.; Nedyalkova, L.; McLaughlin, C. K.; Shoichet, B. K.; Shoichet, M. S. Colloidal Drug Formulations Can Explain "Bell-Shaped" Concentration–Response Curves. *ACS Chem. Biol.* **2014**, *9* (3), 777–784. https://doi.org/10.1021/cb4007584.
(66) Coan, K. E. D.; Shoichet, B. K. Stoichiometry and Physical Chemistry of Promiscuous Aggregate-Based Inhibitors. *J. Am. Chem. Soc.* **2008**, *130* (29), 9606–9612. https://doi.org/10.1021/ja802977h.
(67) Blake, J. F. Identification and Evaluation of Molecular Properties Related to Preclinical Optimization and Clinical Fate. *Med Chem* **2005**, *1* (6), 649–655. https://doi.org/10.2174/157340605774598081.
(68) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58* (17), 7076–7087. https://doi.org/10.1021/acs.jmedchem.5b01105.
(69) Alves, V. M.; Capuzzi, S. J.; Braga, R. C.; Korn, D.; Hochuli, J. E.; Bowler, K. H.; Yasgar, A.; Rai, G.; Simeonov, A.; Muratov, E. N.; Zakharov, A. V.; Tropsha, A. SCAM Detective: Accurate Predictor of Small, Colloidally Aggregating Molecules. *J. Chem. Inf. Model.* **2020**, *60* (8), 4056–4063. https://doi.org/10.1021/acs.jcim.0c00415.
(70) *RDKit: Open-source cheminformatics*. https://www.rdkit.org/ (accessed 2021-09-03).
(71) 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.
(72) *ChEMBL Webresource Client*; The ChEMBL Group, 2022.
(73) *tldr*. https://tldr.docking.org/ (accessed 2022-07-08).
(74) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(75) Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: Updated Data and New Features for Efficient Prediction of Protein Targets of Small Molecules. *Nucleic Acids Research* **2019**, *47* (W1), W357–W364. https://doi.org/10.1093/nar/gkz382.
(76) Awale, M.; Reymond, J.-L. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.* **2019**, *59* (1), 10–17. https://doi.org/10.1021/acs.jcim.8b00524.
(77) Lagunin, A.; Filimonov, D.; Poroikov, V. Multi-Targeted Natural Products Evaluation Based on Biological Activity Prediction with PASS. *CPD* **2010**, *16* (15), 1703–1717. https://doi.org/10.2174/138161210791164063.
(78) Czodrowski, P.; Bolick, W.-G. OCEAN: Optimized Cross REActivity EstimatioN. *J. Chem. Inf. Model.* **2016**, *56* (10), 2013–2023. https://doi.org/10.1021/acs.jcim.6b00067.
(79) Brown, K. K.; Hann, M. M.; Lakdawala, A. S.; Santos, R.; Thomas, P. J.; Todd, K. Approaches to Target Tractability Assessment – a Practical Perspective. *Med. Chem. Commun.* **2018**, *9* (4), 606–613. https://doi.org/10.1039/C7MD00633K.
(80) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53* (8), 1893–1904. https://doi.org/10.1021/ci300604z.
(81) Bank, R. P. D. *RCSB PDB: Homepage*. https://www.rcsb.org/ (accessed 2021-04-30).
(82) mgl-admin. Download AutoDock4. *AutoDock*.
(83) Quiroga, R.; Villarreal, M. A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLOS ONE* **2016**, *11* (5), e0155183. https://doi.org/10.1371/journal.pone.0155183.

FUNDING SOURCES AKNOWLEDGEMENT

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

SupportingInformation.xlsx:

Table S1: FooDB compounds with published interactions, aggregated by inchikey. For each inchikey, the foodb_id, compound class, target, and target class are provided.

Table S2: FooDB compounds with predicted interactions, aggregated by inchikey. For each inchikey, the foodb_id, compound class, target, and target class are provided.

Table S4. Set of predicted interactions in SEA + TC (that uses ChEMBL25) experimentally validated in ChEMBL29.

SupportingInformation Table S3.csv

Table S3: Full tldr predictions for all FooDB compounds. For each interaction, the compound identified, target name, affinity threshold, SEA p.value, TC, target description, SMILES, inchikey are provided.

Table S5.pptx

Table S5. Set of most significant scaffolds for food compounds-target interactions. For each compound class, the most significant scaffold is shown for each target class, for compound class vs target class combinations with > 10 molecules and for scaffolds in > 2 molecules (see Methods).

Figure S1.pdf

Figure S1. Boxplot with distributions of drugEBIlity scores for both “food-specific” targets and “drug-specific” targets.

TABLE OF CONTENTS GRAPHIC