Updated Prediction of Aggregators and Assay Interfering Substructures in Food Compounds

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

Positive outcomes in biochemical and biological assays of food compounds may appear due to the well-described capacity of some compounds to form colloidal aggregates that adsorb proteins, resulting in their denaturation and loss of function. This phenomenon can lead to wrongly ascribing mechanisms of biological action for these compounds (false positives), as the effect is non-specific and promiscuous. Similar false positives can show up due to chemical (photo)reactivity, redox cycling, metal chelation, interferences with the assay technology, membrane disruption, etc., which are more frequently observed when the tested molecule has some definite interfering substructures. Although discarding false positives can be achieved experimentally, it would be very useful to have in advance a prognostic value for possible aggregation and/or interference, based only in the chemical structure of the compound tested, in order to be aware of possible issues, help in prioritization of compounds to test, design of appropriate assays, etc. Previously, we applied cheminformatic tools derived from the drug discovery field to identify putative aggregators and interfering substructures in a database of food compounds, the FooDB, comprising 26457 molecules at that time. Here we provide an updated account of that analysis based on a current, much-expanded version of the FooDB, comprising a total of 70855 compounds. In addition, we also apply a novel machine learning model (the SCAM Detective) to predict aggregators with 46%-53% increased accuracies over previous models. In this way, we expect to provide the researchers in the mode of action of food compounds with a much improved, robust, and widened set of putative aggregators and interfering substructures of food compounds.

KEYWORDS: Food compounds, aggregators, interference filters, PAINS, assay interference, promiscuous compounds, cheminformatics, SCAM Detective

INTRODUCTION

There is currently a great research effort in the identification of the biological mechanisms of action of food compounds from a molecular point of view, in order to understand the beneficial or harmful effects of foods on human health, as well as finding novel nutraceuticals and scaffolds for drug design.[1–9] For that aim, biochemical and/or biological (cellular) assays directed towards different biological targets (typically proteins) are being conducted, so that specific macromolecule-food compound interactions can be identified. However, these assays are subject to compound-related artifacts. For example, compound aggregation is a well-described phenomenon that yields artifacts in biochemical and biological assays.[10–14] Some compounds, due to low water solubility, when tested at concentrations above a critical aggregation concentration (CAC),[15] form colloidal aggregates that adsorb biomacromolecules nonspecifically and alter their activities, in most cases inhibiting them through denaturation,[16] although in some cases activating them.[17] This effect can translate into misled interpretations of the biological mechanism of action of compounds, as it is wrongly ascribed to the target used in the assay while the aggregation is nonspecific. Aggregation is very dependent on the assay conditions (pH, buffer composition, testing concentration) and structure of the compound, rather than on the assay technology and target, and can be alleviated to some extent by the addition of nonionic detergents in the assay medium.

An alternative source of false positives in assays is the presence of substructures that make the molecule interact promiscuously with many targets, through mechanisms like (photo)chemical reactivity, large hydrophobicity, redox cycling, metal chelation, etc.; or

that provide it with some interfering properties with the assay technology (absorption/emission at reading wavelengths, membrane disruption, singlet-oxygen quenching or production, etc.).[18–26]

There is a soaring concern for the presence of increasing numbers of false positives in the scientific literature and the derived databases of bioactivities due to aggregation and/or interfering substructures, as it could severely hurt knowledge extraction, decision making, and lead to the waste of resources and time.[27,28] As a matter of fact, the American Chemical Society has provided specific guidelines[28] to follow when submitting manuscripts reporting biological or biochemical activities of compounds with putative aggregation and/or interfering substructures, such as performing additional orthogonal assays (with the same target but different technology) and/or counter-assays (with different target but the same technology), adding non-ionic detergents in the assay buffer, and reviewing reported activities for the same molecules in the literature.

In a previous work,[29] we performed an systematic analysis of a large database of food compounds, the FooDB,[30] aiming at identifying there both putative aggregators and interfering substructures. This effort would be useful for the scientific community aiming at deciphering the biological mechanisms of action of food compounds, as it would allow to point out possible issues with reported activities, guide in the prioritization of compounds to test, and help in the assay technology selection and design. We used for the analysis well-established and publicly available cheminformatic tools,[12,19,20,22] derived from the statistical and machine learning analysis of a large number of structure-activity datasets from both the high-throughput screening and

medicinal chemistry fields: to identify putative aggregators, we employed the Aggregator Advisor model,[12] while to find putative interfering substructures, we used three standard libraries of substructural filters derived through data mining efforts: the Pan-Assay INterference compoundS ("PAINS") set,[22] that of former-GlaxoWellcome (here called "Glaxo"),[19] and another from Pfizer ("LINT").[20]

Since 2020, the FooDB size has increased considerably, from about 26000 compounds at the time when the previous paper was published, to about 71000 to date. This suggested the need for an update of this analysis, as the number of compounds in the current version more than duplicates that of the previous one, leaving a lot of molecules without analysis for possible aggregation and/or interfering behavior. In addition, we took the opportunity to use more advanced approaches to predict aggregation. The Aggregator Advisor uses a simple and conservative approach for prediction, as it relies on a database of ~12000 known aggregators, and only marks a new compound as aggregator if its Tanimoto similarity to one or more know aggregators is > 0.85 and its logP > 3. Thus, this approach is restricted to a chemical space very close to the known aggregators, not being able to extrapolate to molecules outside this narrow space. It basically suggests putative aggregators with confidence, but potentially misses the vast majority of the chemical space, and contains no information about non-aggregators. On the contrary, a recent model based on machine learning, the so-called SCAM Detective,[31] is based on balanced datasets comprising tens of thousands of compounds for both the aggregator and non-aggregator classes, in two assays, each repeated with and without detergent, and at different concentrations of compound (here SCAM stands for "small, colloidally aggregating molecules"). The use of a much larger and balanced

dataset, together with a Random Forest as predictive model, as well as Extended-Connectivity Fingerprints of diameter 6 (ECFP6) to represent molecular structures, allows to increase the sensitivity and specificity of the predictions, resulting in improvements of total accuracy from 46% to 53% over previous models (e.g. the Aggregator Advisor), and to make it applicable to the whole chemical space. In addition, the SCAM Detective allows to generate probability maps for aggregation, that highlight regions in the molecule with high and low propensity for aggregation and help in the interpretation of the predictions in molecular terms.

Thus, in this work we attempt to provide an updated and more robust analysis of the putative aggregators and interference substructures in the current FooDB, comprising near 71000 molecules. We expect this work will be valuable for the experimentalist testing food compounds in different biological targets, in order to reduce the presence of false positive reports in the literature for these molecules. In turn, this will help globally to gain a better understanding of their true biological mechanisms of action and structure-activity relationships.

MATERIALS AND METHODS

All the analyses were performed in Python 3.9 and with the RDKit cheminformatic toolkit,[32] version 2021.03.3. The FooDB,[30] comprising a total of 70855 molecules (including all detected and quantified, detected but not quantified, expected but not quantified, and predicted) were kindly provided by Dr. Wishart group in SDF format. Structures were checked in a first step, and then standardized and solvent- and counterion-stripped (in the case complex molecules containing counterions and/or solvent molecules), to yield the corresponding parent compound, by using the ChEMBL Structure Pipeline.[33] A few compounds could not be processed by RDKit. Some compounds raising the Pipeline's exclusion flag (i.e. with transition metals, or > 7 of boron atoms, that cannot be properly standardized) or with penalty scores > 5 (due to severe structural inconsistencies) were discarded (molecules with penalty score = 5 were kept as these corresponded to stereo mismatches between InChi vs RDKit vs Mol representations of the molecules, with no impact in the analysis of aggregators or interference substructures). After that, duplicates were removed based on the InChiKey, resulting in a total of 69502 unique molecules. The application of this recently described structure normalization procedure changed slightly the statistics for the molecules analyzed in our previous work and reported therein,[29] that were now reanalyzed as they were a subset of the new FooDB, but we think this provides a more robust analysis of aggregators and interference substructures, without changing the former conclusions.

In some cases, for comparison purposes a list of drugs was used. This was obtained from the DrugBank,[34] using the small molecules in approved, not-withdrawn, and non-illicit

status as in our previous work, and following normalization procedure as with the FooDB. The resulting number of unique molecules was 2154.

For the SCAM Detective calculations to predict aggregators,[31] the code provided by the authors was locally installed and adapted for batch calculations. For comparison purposes with our previous work, the aggregators were also predicted through the Aggregator Advisor method,[12] implemented in a locally programmed script with RDKit as the Aggregator Advisor does not provide batch processing functionality. In this script, a molecule was assigned a "known" status if its Tanimoto similarity with a molecule in the list of 12642 Aggregator Advisor molecules was 1, a status of "probable" if its largest Tanimoto similarity with any molecule in the list of aggregators was ≥ 0.85 and its logP > 3, and an status of "none" for the rest of the molecules.

To test differences in distributions of numeric variables in different groups, the non-parametric Kruskal-Wallis one-way analysis was used, followed by Conover's post-hoc test with p-values correction using Holm's approach. To test differences in proportions for pairs of samples, Z-test for proportions were used.

RESULTS

Physicochemical and structural analysis of the molecules in the new FooDB

Prior to conducting our analysis of aggregators and interfering substructures, it seemed necessary a characterization of the molecules incorporated in the new release of the FooDB, in order to see the extent of overlap with the chemical space of the previous FooDB molecules. A very large subset of the additional molecules in the new FooDB corresponds to acylglycerols (in what follows abbreviated AG), namely compounds with a glycerol backbone and with at least one fatty acid esterified to it, although other types of molecules (non-acylglycerols or NAG) are present. Figure 1 displays different in-silico calculated physicochemical and structural properties for four groups of molecules, specifically the molecules in the previous FooDB (corresponding to the "FD(OLD)" label in the abscissa), new non-acylglycerol molecules in FooDB (likewise, "FD(NEW/NAG)"), new acylglycerol molecules in FooDB ("FD(NEW/AG)"), and DrugBank ("DB") molecules. The descriptors calculated were: TPSA (topological polar surface area, TPSA), logP (LOGP), number of rotatable bonds (RB), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), molecular weight (MW), QED (Quantitative Estimation of Drug-likeness, QED[35]), number of rings (NRING), and fraction of sp3 carbons (FSP3).

We can see in Figure 1 that the dispersion of distributions is quite different for NAG and AG molecules in the new FooDB molecules. The main descriptive statistics and test for differences (omnibus and post-hoc) in the distributions of all the four groups of molecules can be obtained in Supporting Information (Tables S1 and S2). While the NAG tend to have a widening of their distributions as compared to the old FooDB, the AG

have usually very narrow distributions, with TPSA, HBD, HBA and NRING having a null inter-quartile range, which is expected given the high structural redundancy within this group as they are all acyl esters of glycerol. Both NAG and AG display a trend towards more and larger LOGP, RB, HBA, MW and FSP3. This is especially clear for the AG as regarding LOGP, RB and FSP3, and expected given their highly aliphatic structure. On the other hand, both NAG and AG display a reduction in QED and NRING, again not unexpected for the AG molecules.

As regarding TPSA, NAG show a marked increase compared to the old FooDB, while AG show an slight decrease of it; in the case of the HBD, the AG molecules have no one, while NAG show a clear trend towards higher values.

On the other hand, Figure 2 displays the distributions of the 10 most frequent Bemis-Murcko[36,37] scaffolds for FDB(OLD), FDB(NEW/NAG), and DB. AG molecules are not shown as any of them have scaffold (they are linear, branched molecules, and therefore have no Bemis-Murcko scaffolds, which are ring-based). Again, we observe quite different distributions for these sets, in spite the benzene ring being the most frequent scaffold in all of them. The percentage of molecules without scaffold is lowest in the drugs, but highest in the NAG (after AG, which are 100%). Following the benzene scaffold, the drugs show a variety of typical drug scaffolds; in decreasing order, two steroid scaffolds, pyridine, diphenylmethane, another steroid, etc. In turn, the old FooDB molecules have in common the presence of steroid scaffold, but the second most frequent one is cyclohexane, followed by tetrahydropyran, etc., while the NAG compounds display completely different scaffolds: e.g. dual aliphatic esters ending in furan rings, cyclopentamine, imidazole, etc.

194 In summary, we see that the new FooDB molecules, both AG and NAG, appear to occupy
195 a different region of the chemical space (including physicochemical properties and
196 structures), more separated to that of the drugs, as compared to the previous release
197 of FooDB. This observation stresses the need for an updated analysis of aggregators and
198 interference substructures in the FooDB.

199 Aggregators Analysis

200 In our previous work, the Aggregator Advisor method[12] was used to predict putative
201 aggregators and identify known aggregators. The Aggregator Advisor is based on a very
202 simple approach, comprising the calculation of the Tanimoto similarity of the tested
203 compound to a list of ~12600 know aggregators (using topological fingerprints), and if
204 the similarity is > 0.85 to at least one of these compounds, and its logP is > 3, it is assigned
205 a "possible aggregator" status; in the rest of the cases no prediction is made
206 ("unknown", or "non-aggregator" status). Thus, it is a very conservative approach that
207 is unable to give predictions of compounds lying outside the close structural space of
208 the aggregators list, although it is useful as a fast and easy-to-implement way to identify
209 close analogs to the known aggregators list with high risk of aggregation if lipophilic
210 enough.

211 On the contrary, the SCAM Detective[31] is a machine learning-based approach using
212 random forests[38] that is capable in principle to extrapolate to the whole structural
213 chemical space. The training sets for the SCAM Detective were obtained from pairs of
214 quantitative high-throughput screening (qHTS) campaigns in PubChem[39] run with the
215 same assay conditions but in the presence and absence of added detergent in the assay
216 buffer. By comparing the dose-response curves in each pair for each compound, in the

presence and in the absence of detergent, it was possible to mark it as putative aggregator or non-aggregator. Models were derived for both AmpC β-lactamase and the cysteine protease cruzain, which are frequently used counter-assays to discard false positives in assays.[40–42] The corresponding screens in PubChem were tested at different experimental conditions in terms both of assay buffer and dosing concentrations. This is an interesting feature as by comparing the predictions in both assays an approach to assess the robustness of the prediction is available.

In addition, in the development of the training set, a data-rebalancing approach was applied, so that in both training sets there were equal numbers of aggregators and non-aggregators. In this way, the SCAM Detective is able to *reliably predict both aggregators and non-aggregators*, and thus have a balanced sensitivity and specificity (0.72 and 0.73, respectively for the AmpC β-lactamase model, and 0.71 and 0.69 for the cruzain model).

The SCAM Detective also provides a measure of the reliability of its predictions, which is based on the so-called *applicability domain* (AD) of the model, defined as

$$D_{cutoff} = \langle D \rangle + 0.5S$$

where $\langle D \rangle$ and S are the average and SD of all the Euclidian distances in the descriptor space used between each compound and its nearest neighbors in the training set. New compounds with a minimum distance to the molecules in the training set $D > D_{cutoff}$ would be outside the AD of the model, meaning that the predictions for it would be in general less reliable.

On top of that, prediction of fragment contributions in the form of contour maps can be generated for the modeled molecules, aiding to provide an interpretation of the model

prediction through the groups in the compound most responsible for the aggregating or non-aggregating behavior.

By first applying the Aggregator Advisor method to the updated FooDB, we can identify a total of 92 known aggregators and 37 predicted aggregators. These concentrate mainly in the old FooDB database, since within the new compounds no one was a predicted compound, and only two were known aggregators. These were in the NAG group. By merging known and predicted aggregators, this leads to a 0.56% and a 0.004% of aggregator rate for the old and new FooDB molecules, respectively. These small numbers reflect the conservativeness of the Aggregator Advisor and that the novel FooDB molecules display even less overlap with the chemical space of the list of aggregators used by the method compared with the previous FooDB.

These numbers change dramatically upon application of the SCAM Detective. Table 1 displays the total number and percentages of predicted aggregators in both FooDB (and its different subsets) and DrugBank (for comparison purposes) for both the β-lactamase and cruzain models, as well as their intersection. The actual predictions for all the molecules in FooDB can be obtained in Supporting Information (Table S3). It can be seen that the aggregator rates for the SCAM Detective are much higher. For instance, the β-lactamase model predicts a 76.7% of aggregators in FooDB, that rises up to 95.39% for the AG subset. For the cruzain model, the aggregator rate is 40.82% for the whole FooDB, and 52.98% for the AG subset. In both models, the aggregator rates for the new FooDB molecules are significantly higher than for the old ones (43.17% and 19.95%, for β-lactamase and cruzain, respectively), both for NAG and AG, although especially for the later. This can be expected from the more lipophilic and flexible nature of these

molecules, with long aliphatic chains that makes them prone to aggregation in the AG. As a way of confirmation, Figure 3 displays the predicted fragment contribution maps for FDB00135, a predicted non-aggregator in both models from the old FooDB, and FDB080642, a doubly-predicted aggregator and an AG compound. We can see that while FDB000135 shows green contours indicative of non-aggregation contribution together with some weakly concentrated magenta contours, FDB080642 displays highly concentrated and dark-magenta contours, a signature of high-aggregating contribution, especially in its three polymethylene chains.

On the other hand, the DrugBank display aggregator rates near the old FooDB molecules, slightly lower: 34.44% and 18.89% for β-lactamase and cruzain, respectively.

From Table 1 it can be observed that the β-lactamase model shows in all the compound sets an increased aggregator rate as compared to the cruzain model. This is in contrast to the original training datasets used, which showed comparable aggregator rates for β-lactamase and cruzain assays, although obviously the results obtained here depend on the chemical spaces of the molecules that were aggregators in one or the other training sets. In general, there is a highly significant association between the two models, with a 61% of molecules being aggregators or non-aggregators simultaneously in both models. This rises to 73% in the case of the DrugBank molecules. By considering the intersection between the models, which would correspond to a more robust (although more conservative) prediction of aggregation, the aggregation rate for the whole FooDB would be 39.25% (including 15.34%, 29.76% and 52.98% for the old FooDB, NAG, AG, respectively) vs a 13.32% for the DrugBank.

If we focus on the reliability of the predictions based on the AD of the different sets, we obtain the results collected in Table 2. Again, the number and percentage of molecules within the AD in one or the other model and the intersection of both, for the different compound sets, are shown. We can see very large percentages of molecules within the applicability domain in the updated FooDB (81.96% and 86.57% for β-lactamase and cruzain models, respectively), even higher for the new molecules, especially in the case of AG, which is very close to 100% (98.96% and 99.99%, respectively). For the old FooDB molecules the percentages are lower, 50.93% and 61.46%; these values are similar to those observed for the DrugBank (52.01% and 55.93%).

In this case, the percentage of molecules in the AD is slightly but significantly higher for the cruzain model compared to the β-lactamase model if we consider the whole set of molecules: 85.64% vs 81.05%, respectively. The agreement between both models as far as AD is concerned is very high, and ~94% of the molecules are within or without the AD of both models simultaneously. This is probably due to both models using similar collections of compounds in their training sets, as well as the very high percentages of molecules within the AD in both cases. By intersecting both models, there is just an slight decrease of percentages in all the sets, so that the whole FooDB has a 81.43% (including 49.51%, 78.07%, and 98.96% for old FooDB, NAG, and AG, respectively) vs a 47.58% for the DrugBank.

We could ask what are the SCAM Detective predictions for the compounds marked as known (92) or predicted (37) aggregators by the Aggregator Advisor. Of the 92 known aggregators, 23 are predicted aggregators by the β-lactamase model, and 15 by the cruzain model. Of the 37 predicted aggregators, 16 are predicted aggregators by the β-

lactamase model, while 16 are predicted aggregators by the cruzain model. Thus, there seems to be a modest agreement for the prediction of the aggregator class between both SCAM Detective and Aggregator Advisor models, although it must be taken into account that the datasets were derived in different conditions, with different definitions of aggregation (different predicted labels), and with different compound sets. In addition, the numbers used for the comparison are very small, given the tiny number of compounds predicted by the Aggregator Advisor, and we are only considering the aggregators class, not the non-aggregators.

In summary, we have obtained novel predictions for the updated FooDB through the machine learning approach used by the SCAM Detective, which was reported to provide balanced sensitivity/specificity predictions and an increase of accuracy from 46% to 53%[31] compared to other methods (e.g. Aggregator Advisor[12] and Hit Dexter[43], see SCAM Detective paper[31]). The FooDB show relatively large percentages of predicted aggregated molecules, 76.70% in β-lactamase and 40.82% in cruzain. The old fraction of FooDB displays clearly lower percentages (43.17% and 19.85%, respectively), while the new fraction of molecules shows increased values, especially in the case of AG, which are predicted aggregators in 95.39% of the cases by the β-lactamase model and 52.98% by the cruzain one. For the whole FooDB, a very large proportion of molecules appears to be within the AD of the models, 86.57% for the cruzain model and 81.96% for the β-lactamase. These predictions, provided as Supporting Information (Table S3), are expected to help the community of scientist aiming in understanding the biological mechanisms of action of food compounds to identify aggregators in their assays.

Analysis of interference substructures

In our previous work we also analyzed the presence of nuisance substructures in the FooDB. In this section we repeat that analysis for the updated FooDB. Three filter sets were used, namely PAINS, Glaxo, and LINT. The first one of these was derived by Baell and Holloway[22] after analysis of a series of high-throughput screens run with the AlphaScreen technology, and comprise a total of 481 filters, grouped in three families with decreasing statistical support: family A (16 filters), corresponding to the filters with the strongest support; family B (55 filters), of filters with median support; and family C, comprising 409 filters with the lowest statistical support. The PAINS filters were derived using a relatively "clean" screening collection developed after an in-silico effort to preclude the presence of inappropriately reactive functional groups, like epoxides, aziridines, alkyl halides, labile esters, etc.[22] Thus, in order to be able in our analysis for the detection of these substructures, we also included two additional more basic filter sets: Glaxo, corresponding to 55 filters derived in GlaxoWellcome,[19] and LINT, of 57 filters and generated in Pfizer.[20]

Figure 4 displays the 18 PAINS filters matched by at least one compound in FooDB, color coded by filter family (green for family A, blue for family B, and red for family C). We can see the same set of filters and almost the same distribution as the one observed previously,[29] dominated by "catechol_A(92)", followed by "quinone_A(370)", "imine_one_A(321)" and "azo_A(324)". No molecules in the updated FooDB match any more of the 18 matching filters in the old FooDB. The reason is that none of the AG molecules match any of these filters, while just a few set of 74 NAG match a reduced set of three filters observed before:[29] "catechol_A(92)", "imine_one_A(321)", and "quinone_A(370)". Because of this, the percentage of molecules filtered by PAINS filters

353 is reduced from 6.80% down to 2.23%. If we focus on the most reliable family A, the
354 percentage of filtered molecules is just a 0.37% (before it was 1.11%), corresponding to
355 ~17% of the total of matches by the PAINS set, while there are 6 filters in this family that
356 match at least one molecule.

357 Figure 5 shows the matches distribution for the Glaxo filter set. 35 filters match at least
358 one molecule, and now the distribution is overwhelmingly dominated by the first filter,
359 namely "I1 Aliphatic methylene chains 7 or more long", with a total of 51597 matches.
360 This is because of the new AG molecules, almost all of them matching this filter, as
361 expected due to the frequent presence in their structure of long polymethylenic chains.
362 Other than that, the AG do not match any other Glaxo filter. In the case of the NAG
363 molecules, again the most frequent filter is "I1 Aliphatic methylene chains 7 or more
364 long", but in this case the second one is no longer "N3 Saponin derivatives", which is
365 very unusual in these molecules, but instead "I15 Di and Triphosphates", followed by "I5
366 Thiols" and "N2 Polyenes". All in all, no additional Glaxo filter absent in the previous
367 study[29] matches any of the new FooDB molecules.

368 As regarding the LINT filters (Figure 6), a large increase of matched molecules is
369 observed for the first two filters, "long aliphatic chain, 6+" and "aliphatic ester, not
370 lactones", as compared to the matches in the previous FooDB,[29] due again to the large
371 number of new AG molecules that match these substructures. No additional filter is
372 matched by the AG compounds, while the NAG ones has as most frequently matched
373 filter "S/PO3 groups", followed by the previous "long aliphatic chain, 6+", "alkyl esters
374 of S or P" and "aliphatic ester, not lactones". As was observed with PAINS and Glaxo, no
375 new filter appears here that were not present in the previous study.

Table 3 collects the filter match statistics for all these filter sets. We can see that, as said before, the new compounds decrease the stringency of both the PAINS and PAINS-A sets, due to the little number of matching molecules in the new FooDB molecules for these sets (74 in total, all of them in the NAG group). The effect of this is to reduce the percentage of matched molecules from 6.80% to 2.23% (PAINS), and from 1.11% to 0.37% (PAINS-A). The contrary is observed for Glaxo and LINT, where the percentage of filtered molecules raises up to 78.45% and 85.11%, respectively, while in the previous work it was 36.18% and 55.43%. Similar dual effect is observed if we focus on the normalized percentage of matched molecules: now it is 0.0046%, 0.023%, 1.44%, and 1.49% for PAINS, PAINS-A, Glaxo, and LINT, while before it was 0.014%, 0.069%, 0.658%, and 0.973%. As in our previous work, the stringency order considering the percentage of filtered molecules is as follows: PAINS-A < PAINS < Glaxo < LINT. If instead we consider this percentage but normalized by the number of filters in the set, PAINS and PAINS-A switch order, but the rest remains the same: PAINS < PAINS-A < Glaxo < LINT. The same order is observed by the fraction of filters with at least one matching molecule in each set, that goes from 18 out of 481 in PAINS, to 49 out of 57 in the case of LINT, with 35 out of 55 in the case of Glaxo in between.

All these filter matches are collected in Supporting Information (Table S4). In the same way, it is expected that this will be useful to identify interferences in biochemical and biological assays of food compounds.

DISCUSSION

A large research effort is being devoted to the determination of the biological mechanism of action of food compounds, in order to understand the beneficial or harmful effect of foods in human health.[1–9,44] These studies are conducted by performing biochemical or biological (cellular) assays aiming to see if the food molecule interacts with some biological target, typically a protein. It is well known from the field of drug discovery that in some cases an assay can result in a false positive or misleading outcome due to some property of the tested molecule:[10,24,26–28,41] the molecule forms colloidal aggregates that denature the target protein or has some substructure that make it prone to membrane disruption, (photo)reactivity, redox cycling, etc., or rather to generate interferences with the assay signal.[18–23,26]

In a previous work,[29] we applied cheminformatic techniques from the drug discovery field to identify molecules prone to such false-positive behavior in a database of food compounds, the FooDB,[30] to find food compounds with these putative issues. The FooDB is (quoting from its web site) “the world’s largest and most comprehensive resource on food constituents, chemistry and biology. It provides information on both macronutrients and micronutrients, including many of the constituents that give foods their flavor, color, taste, texture and aroma”. Here we provide an update of that analysis after the FooDB more than duplicated its size (~26K compounds to ~71K compounds), that includes also the use of novel machine learning models to predict aggregation, as the method used before (Aggregator Advisor[12]) was not able to give predictions for the majority of food compounds. We opted to use the so-called SCAM Detective,[31] as it has been observed to yield accuracies ~50% above previous methods, including the

Aggregator Advisor, and as other methods like the Hit Dexter[43] show lower accuracy[31] and also predict a slightly different endpoint, namely hit promiscuity, which is a related but different label (aggregators are all promiscuous, but not all promiscuous compounds are such because of aggregating behavior).

From a practical point of view, the files provided here as Supplementary Information should be used by the experimenter to find if the tested compound appears there with one or more annotations for aggregation and/or interference. If the annotation is for aggregation, its presence can be experimentally checked by different approaches: decrease of activity after addition of small quantities of non-ionic detergents, counter-assay in aggregation-sensitive assays (e.g. β-lactamase), or detection of colloidal aggregates through dynamic light scattering. On the other hand, if the annotation is for a substructure that generates assay signal interferences (e.g. absorption, fluorescence, etc.), a possible solution is the test in an orthogonal assay using an alternative technology and signal. The third option are those interfering substructures that provide nonspecific activity (promiscuity) through variable mechanisms (e.g. membrane disruption); in that case it could be possible to run a counter-assay with a different and unrelated target and the same technology. In general, it is always advisable to check in public databases like ChEMBL[45] or PubChem[39] about previously reported activities of the compound, which would be informative about possible promiscuity issues if the molecule has shown activity against a wide set of unrelated targets. In addition, if chemical modifications are performed on the compound, finding a lack of a defined structure-activity relationship would be a signature of artifactual activity. More

thorough approaches to these issues have been described elsewhere;[46–48] and for more specifically referring to publication in ACS journals see reference (28).

The updated prediction performed in this work has increased considerably the number of putative aggregators in FooDB: 77% for the β-lactamase model and 41% for the cruzain model. One reason is the AG component of the new FooDB molecules, for which the β-lactamase model predicts a 95% of aggregators, and the cruzain a 53% (Table 2). Given the very large hydrophobicity of these molecules (median logP of 17.9) together with the extreme flexibility of their aliphatic structure, these predictions seem quite reasonable. The higher hydrophobicity, could also contribute to the increase in aggregator rates in NAG molecules over the old FooDB molecules, together with the significantly larger surface areas of the former group as compared to the latter and the higher number of rotatable bonds. In addition, another reason for getting many more aggregators is that the Aggregator Advisor had not been able to give predictions for the vast majority of the FooDB molecules as just a few of them were similar to one or more in the list of known aggregators: using a Tanimoto radius of 0.85, only 437 in the old FooDB, 20 for NAG, and none for the AG. This shows the advantages of the SCAM Detective, that is trained with a very large and diverse dataset of both aggregators and non-aggregators and gives predictions for the whole chemical space. Also, it is worth mentioning that a very large fraction of the FooDB molecules are within the AD of the SCAM Detective models, for which the predictions are expected to be more reliable, ranging from 51% to 100% depending on the subset and model. Thus, we expect this effort to give a much more reliable and comprehensive identification of putative aggregators in the food molecules of FooDB.

As regarding the interference filter analysis, in the previous work we applied as cheminformatic tool three well-known substructure filter sets derived from the high-throughput screening and medicinal chemistry fields: PAINS,[22] Glaxo[19] and LINT.[20] Here we have applied these filters to the updated FooDB too. As a result, for the PAINS we have observed a very small number of novel FooDB molecules matching them, so that the observed current distribution is very similar to the previous one; also the set of matched filters remains equal. The same set of matched filters are also observed again with the Glaxo and LINT filters, but in this case the distributions change significantly as the filters representing long aliphatic chains or aliphatic esters ("I1 Aliphatic methylene chains 7 or more long" in Glaxo and "long aliphatic chain, 6+" and "aliphatic ester, not lactones" in LINT) have an enormous increase of hits, corresponding mostly to the new AG compounds. As a result, the interference rates for these filter sets increase considerably from the previous analysis, resulting in a total of 78% and 85% for Glaxo and LINT, respectively (before they were 36% and 55%).

In general, we observe a decrease of drug-like properties in the new FooDB molecules, with a significant decrease of its drug-like "chemical beauty" (as measured by the Estimation of Drug-likeness (QED) descriptor[35], p-value < 0.001 in post-hoc test, see Supplementary Material, Table S2) and an increase of aggregator and interference rates (for the Glaxo and LINT filters). The new FooDB molecules (especially the AG component) tend to have more hydrophobicity, flexibility, and molecular weight; these factors make them more prone to aggregation on one hand (also the higher TPSA in the NAG molecules), and to display the long polymethylene-type of interfering substructure appearing in the Glaxo and LINT filter sets. The near complete absence of interferences

of the PAINS filters can on the other hand be explained by considering that the later were derived from a collection of relatively “clean” compounds previously filtered from more basic problematic substructures, to make them more amenable as starting points for drug development. The PAINS filters derived from that collection would be more specific to drug- or lead-like molecules, probably of a more synthetic origin, and corresponding to substructures scarcely present in the FooDB.

To conclude, we can say that the putative aggregators and interference matches for the new FooDB found in this work and available as Supporting Information would help to decrease the false positives in assays performed with food compounds, by applying, when present, the approaches discussed above, and therefore to gain a better and more robust understanding of their biological mechanisms of action, thus reducing the rates of false positive results in public databases like ChEMBL[45] or PubChem[39]. Other uses for these predictions would be the prioritization of compounds for testing, applications in large-scale data mining efforts for understanding structure-activity relationships, design of reliable nutraceutics, and selection of novel scaffolds for development of new drugs.

ABBREVIATIONS

American Chemical Society (ACS), Small, Colloidal, Interfering Molecules (SCAM) Detective, Pan-Assay INterference compoundS (PAINS), Invalid Metabolic PanaceaS (IMPs)

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-Vicente - 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 is thanked for providing the updated version of FooDB in SDF format. Dr Patrick Walters is acknowledged for providing the Glaxo, LINT and PAINS nuisance filters in SMARTS format adapted for RDKIT.

SUPPORTING INFORMATION

Statistical Analysis of PhysChem Distributions.xlsx. Both p.values for tests for comparison of physicochemical distributions between groups of compounds (omnibus and pairwise post-hoc), and descriptive statistics (median + interquartile range) for these distributions are collected here.

SCAM Detective Predictions for FooDB.xlsx Aggregator predictions for the FooDB using the SCAM Detective (both β-lactamase and cruzain models). The prediction result (0 non-aggregator, 1 aggregator) plus the AD result (Inside AD vs Outside AD) is shown for each FooDB compound.

Filter Matches for FooDB.xlsx. Filter matches for PAINS, Glaxo and LINT filter sets for the updated FooDB

REFERENCES

(1) 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.

(2) 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.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(7) 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.

(8) 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.

(9) Faudone, G.; Arifi, S.; Merk, D. The Medicinal Chemistry of Caffeine. *J. Med. Chem.* **2021**. https://doi.org/10.1021/acs.jmedchem.1c00261.

(10) 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.

(11) Reker, D.; Bernardes, G. J. L.; Rodrigues, T. Computational Advances in Combating Colloidal Aggregation in Drug Discovery. *Nat. Chem.* **2019**, *11* (5), 402–418. https://doi.org/10.1038/s41557-019-0234-9.

(12) 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.
(13) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.* **2007**, *50* (10), 2385–2390. https://doi.org/10.1021/jm061317y.
(14) 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.
(15) 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.
(16) Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K. Promiscuous Aggregate-Based Inhibitors Promote Enzyme Unfolding. *J. Med. Chem.* **2009**, *52* (7), 2067–2075. https://doi.org/10.1021/jm801605r.
(17) Zorn, J. A.; Wille, H.; Wolan, D. W.; Wells, J. A. Self-Assembling Small Molecules Form Nanofibrils That Bind Procaspase-3 to Promote Activation. *J Am Chem Soc* **2011**, *133* (49), 19630–19633. https://doi.org/10.1021/ja208350u.
(18) Rishton, G. M. Reactive Compounds and in Vitro False Positives in HTS. *Drug Discovery Today* **1997**, *2* (9), 382–384. https://doi.org/10.1016/S1359-6446(97)01083-0.
(19) 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.
(20) 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.
(21) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, *3* (3), 435–444. https://doi.org/10.1002/cmdc.200700139.
(22) 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.
(23) Baell, J. B. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **2016**, *79* (3), 616–628. https://doi.org/10.1021/acs.jnatprod.5b00947.
(24) 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.
(25) Chakravorty, S. J.; Chan, J.; Greenwood, M. N.; Popa-Burke, I.; Remlinger, K. S.; Pickett, S. D.; Green, D. V. S.; Fillmore, M. C.; Dean, T. W.; Luengo, J. I.; Macarrón, R. Nuisance Compounds, PAINS Filters, and Dark Chemical Matter in

the GSK HTS Collection. *SLAS Discov* **2018**, *23* (6), 532–545. https://doi.org/10.1177/2472555218768497.
(26) Dahlin, J. L.; Auld, D. S.; Rothenaigner, I.; Haney, S.; Sexton, J. Z.; Nissink, J. W. M.; Walsh, J.; Lee, J. A.; Strelow, J. M.; Willard, F. S.; Ferrins, L.; Baell, J. B.; Walters, M. A.; Hua, B. K.; Hadian, K.; Wagner, B. K. Nuisance Compounds in Cellular Assays. *Cell Chemical Biology* **2021**, *28* (3), 356–370. https://doi.org/10.1016/j.chembiol.2021.01.021.
(27) Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature News* **2014**, *513* (7519), 481. https://doi.org/10.1038/513481a.
(28) 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.
(29) 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.
(30) FooDB https://foodb.ca/ (accessed 2021 -09 -13).
(31) 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.
(32) RDKit: Open-source cheminformatics https://www.rdkit.org/ (accessed 2021 -09 -03).
(33) 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.
(34) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Research* **2006**, *34* (suppl_1), D668–D672. https://doi.org/10.1093/nar/gkj067.
(35) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nature Chem* **2012**, *4* (2), 90–98. https://doi.org/10.1038/nchem.1243.
(36) 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.
(37) 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.
(38) Breiman, L. Random Forest. *Machine Learning* **2001**, *45*, 5–32.
(39) 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.

(40) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-Throughput Assays for Promiscuous Inhibitors. *Nat Chem Biol* **2005**, *1* (3), 146–148. https://doi.org/10.1038/nchembio718.
(41) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2010**, *53* (1), 37–51. https://doi.org/10.1021/jm901070c.
(42) Mott, B. T.; Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Ang, K. K.-H.; Leister, W.; Shen, M.; Silveira, J. T.; Doyle, P. S.; Arkin, M. R.; McKerrow, J. H.; Inglese, J.; Austin, C. P.; Thomas, C. J.; Shoichet, B. K.; Maloney, D. J. Identification and Optimization of Inhibitors of Trypanosomal Cysteine Proteases: Cruzain, Rhodesain, and TbCatB. *J. Med. Chem.* **2010**, *53* (1), 52–60. https://doi.org/10.1021/jm901069a.
(43) Stork, C.; Chen, Y.; Šícho, M.; Kirchmair, J. Hit Dexter 2.0: Machine-Learning Models for the Prediction of Frequent Hitters. *J. Chem. Inf. Model.* **2019**, *59* (3), 1030–1043. https://doi.org/10.1021/acs.jcim.8b00677.
(44) Chen, M.; Pan, D.; Zhou, T.; Gao, X.; Dang, Y. Novel Umami Peptide IPIPATKT with Dual Dipeptidyl Peptidase-IV and Angiotensin I-Converting Enzyme Inhibitory Activities. *J. Agric. Food Chem.* **2021**, *69* (19), 5463–5470. https://doi.org/10.1021/acs.jafc.0c07138.
(45) 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.
(46) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46* (21), 4477–4486. https://doi.org/10.1021/jm030191r.
(47) Auld, D. S.; Inglese, J.; Dahlin, J. L. Assay Interference by Aggregation. In *Assay Guidance Manual*; Markossian, S., Grossman, A., Brimacombe, K., Arkin, M., Auld, D., Austin, C. P., Baell, J., Chung, T. D. Y., Coussens, N. P., Dahlin, J. L., Devanarayan, V., Foley, T. L., Glicksman, M., Hall, M. D., Haas, J. V., Hoare, S. R. J., Inglese, J., Iversen, P. W., Kales, S. C., Lal-Nag, M., Li, Z., McGee, J., McManus, O., Riss, T., Saradjian, P., Sittampalam, G. S., Tarselli, M., Trask, O. J., Wang, Y., Weidner, J. R., Wildey, M. J., Wilson, K., Xia, M., Xu, X., Eds.; Eli Lilly & Company and the National Center for Advancing Translational Sciences: Bethesda (MD), 2004.
(48) Coussens, N. P.; Auld, D. S.; Thielman, J. R.; Wagner, B. K.; Dahlin, J. L. Addressing Compound Reactivity and Aggregation Assay Interferences: Case Studies of Biochemical High-Throughput Screening Campaigns Benefiting from the National Institutes of Health Assay Guidance Manual Guidelines. *SLAS Discov* **2021**, 24725552211026240. https://doi.org/10.1177/24725552211026239.

FUNDING SOURCES AKNOWLEDGEMENT

AS-R acknowledges the Consejería de Ciencia, Universidades e Innovación de la Comunidad de Madrid, Spain (Ref. PEJ-2020-AI/BMD-19384), for a research assistant contract.

FIGURE CAPTIONS

1. Figure 1. Boxplots for distributions of TPSA (topological polar surface area), LOGP (logarithm of the octanol/water partition coefficient), RB (number of rotatable bonds), HBD (number of hydrogen bond donors), HBA (number of hydrogen bond acceptors), MW (molecular weight), QED (quantitative estimation of drug-likeness), NRING (number of rings), and FSP3 (fraction of sp3-hybridized carbons), for the previous release of FooDB analyzed before[29] (FDB(OLD)), new non-acylglycerol FooDB molecules in the new release (FDB(NEW/NAG)), new FooDB acylglycerol molecules (FDB(NEW/AG)), and DrugBank molecules (DB). For clarity purposes, outliers have been removed from the plots.
2. Figure 2. Bemis-Murcko[36,37] scaffold distributions (10 top scaffolds only shown in decreasing frequency) for old FooDB, new FooDB (NAG), and DrugBank molecules.
3. Figure 3. Fragment contribution maps for SCAM Detective predictions for FDB00135 (from the former FooDB) and FDB080642 (a novel AG) in the β-lactamase and cruzain models. The former compound is a predicted non-aggregator in both models, while the later is a predicted aggregator in both models. The color and concentration of contours indicate the direction and strength of the contribution: magenta for aggregation, and green for non-aggregation. Strong contributions result in concentrated contours, weak in separated ones.
4. Figure 4. PAINS filter set distribution across FooDB matching molecules. Only the 18 filters with at least one match are displayed. Bars are color coded by filter family, where family A is green, family B is blue, and family C is red.

5. Figure 5. Glaxo filter set distribution across the FooDB matching molecules. Only the 35 filters with at least one match are displayed.
6. Figure 6. LINT filter set distribution across the FooDB matching molecules. Only the 49 filters with at least one match are displayed.

Table 1. Statistics of Prediction of Aggregators by the SCAM Detective[a]

Compound Set	β-lactamase	cruzain	Both
FooDB	54147 (76.70)	28815 (40.82)	27707 (39.25)
FooDB(OLD)	10124 (43.17)	4656 (19.85)	3598 (15.34)
FooDB(NEW/NAG)	2608 (69.94)	1160 (31.11)	1110 (29.76)
FooDB(NEW/AG)	41415 (95.39)	22999 (52.98)	22999 (52.98)
DrugBank	755(34.44)	414 (18.89)	292 (13.32)

[a] For the different compound sets, the number (percentage) of predicted aggregators in the β-lactamase and cruzain models, and the intersection of both models, are shown.

Table 2. Molecules Within Applicability Domain of SCAM Detective Models[a]

Compound Set	β-lactamase	cruzain	Both
FooDB	57859 (81.96)	61113 (86.57)	57487 (81.43)
FooDB(OLD)	11945 (50.93)	14415 (61.46)	11611 (49.51)
FooDB(NEW/NAG)	2950 (79.11)	3288 (88.17)	2912 (78.07)
FooDB(NEW/AG)	42964 (98.96)	43410 (99.99)	42964 (98.96)
DrugBank	1140 (52.01)	1226 (55.93)	1043 (47.58)

[a] For the different compound sets, the number (percentage) of compounds within the applicability domain in the β-lactamase and cruzain models, and the intersection of both models, are shown.

Table 3. Statistics of Matches of Filter Sets for FooDB[a]

Filter Set	# filters	# matching filters	# matching molecules	Filtered molecules (%)	Filtered molecules / filter (%)
PAINS	481	18	1554	2.23	0.0046
PAINS-A	16	6	260	0.37	0.023
Glaxo	55	35	54693	78.45	1.44
LINT	57	49	59337	85.11	1.49

[a]For the sets PAINS, PAINS-A, Glaxo, and LINT, the number of filters, number of matching filters, number of matching molecules, filtered molecules (%), and filtered molecules per filter (%) are displayed for the updated FooDB compound set.

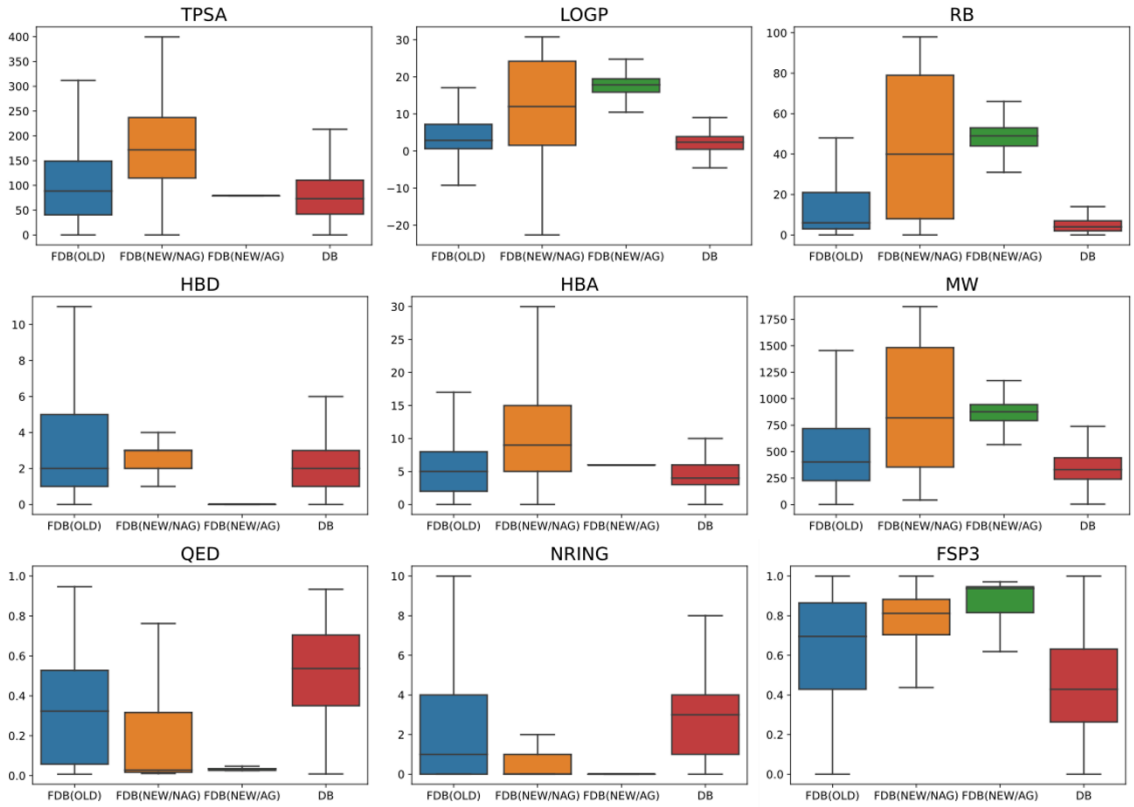

TPSA
LOGP
RB
HBD
HBA
MW
QED
NRING
FSP3
FDB(OLD)
FDB(NEW/NAG)
FDB(NEW/AG)
DB

Figure 1.

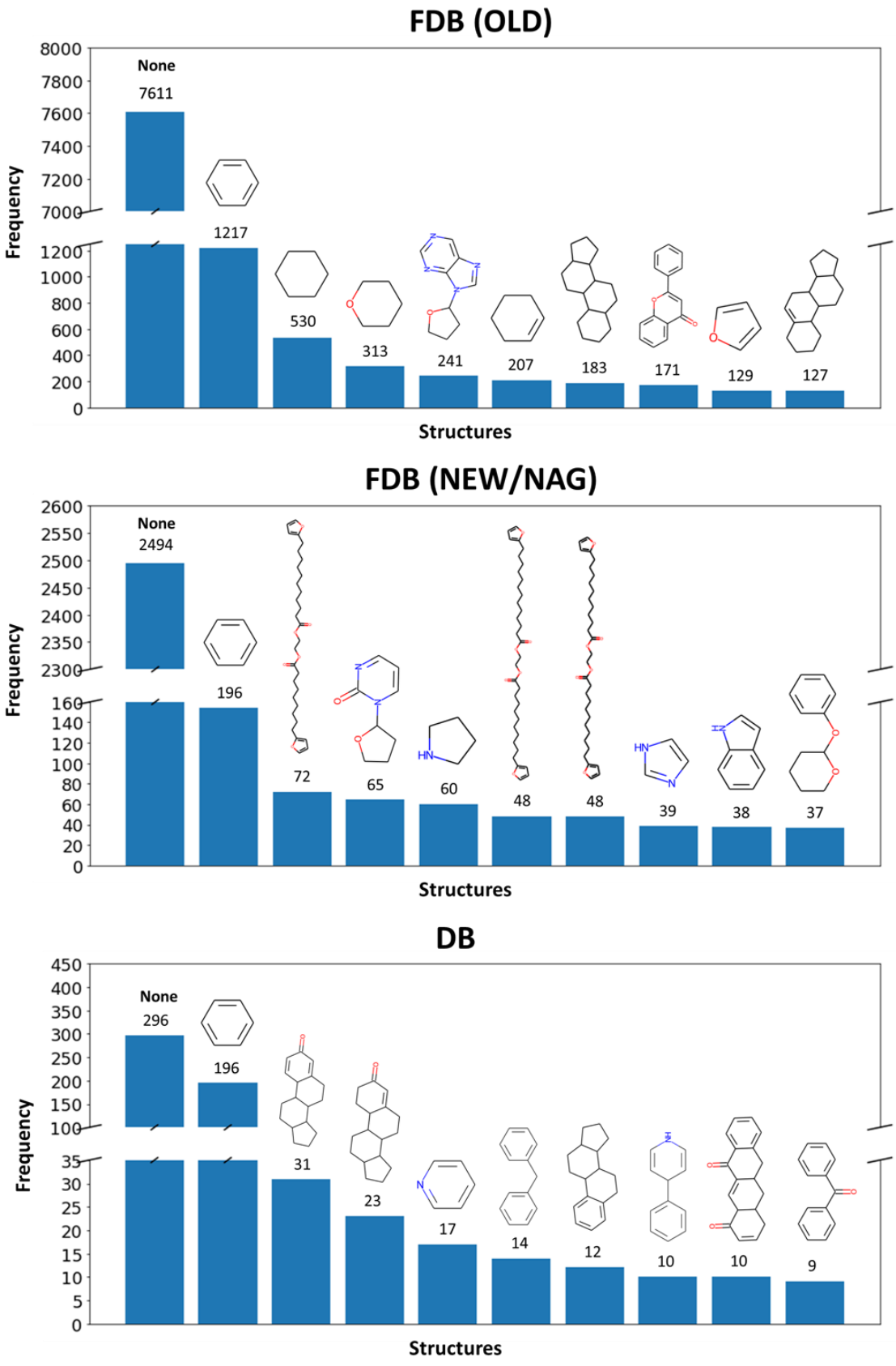

FDB (OLD)
Frequency
Structures
None
7611
1217
530
313
241
207
183
171
129
127
FDB (NEW/NAG)
Frequency
Structures
None
2494
196
72
65
60
48
48
39
38
37
DB
Frequency
Structures
None
296
196
31
23
17
14
12
10
10
9

Figure 2.

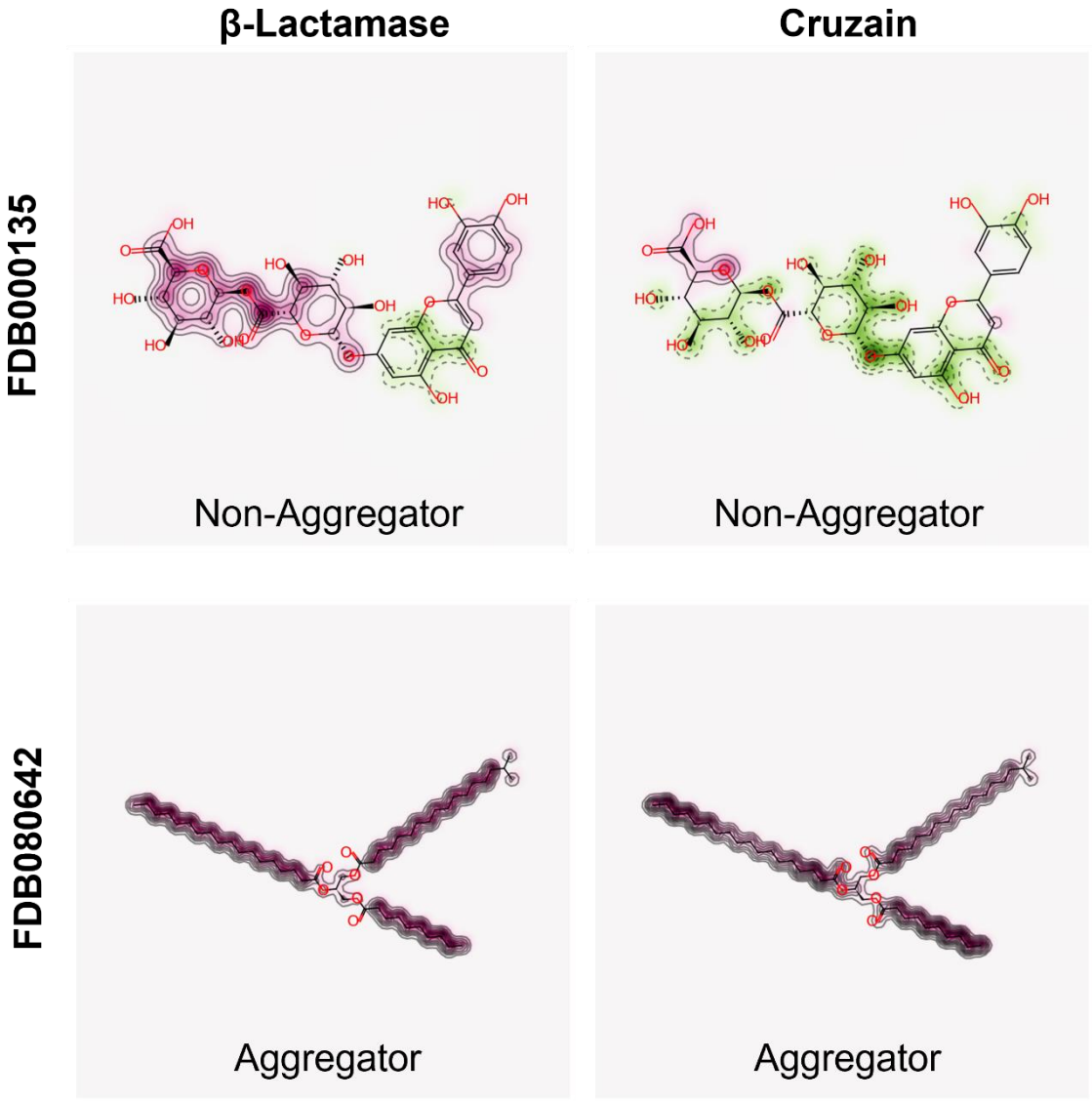

β-Lactamase
Cruzain
FDB000135
Non-Aggregator
Non-Aggregator
FDB080642
Aggregator
Aggregator

Figure 3.

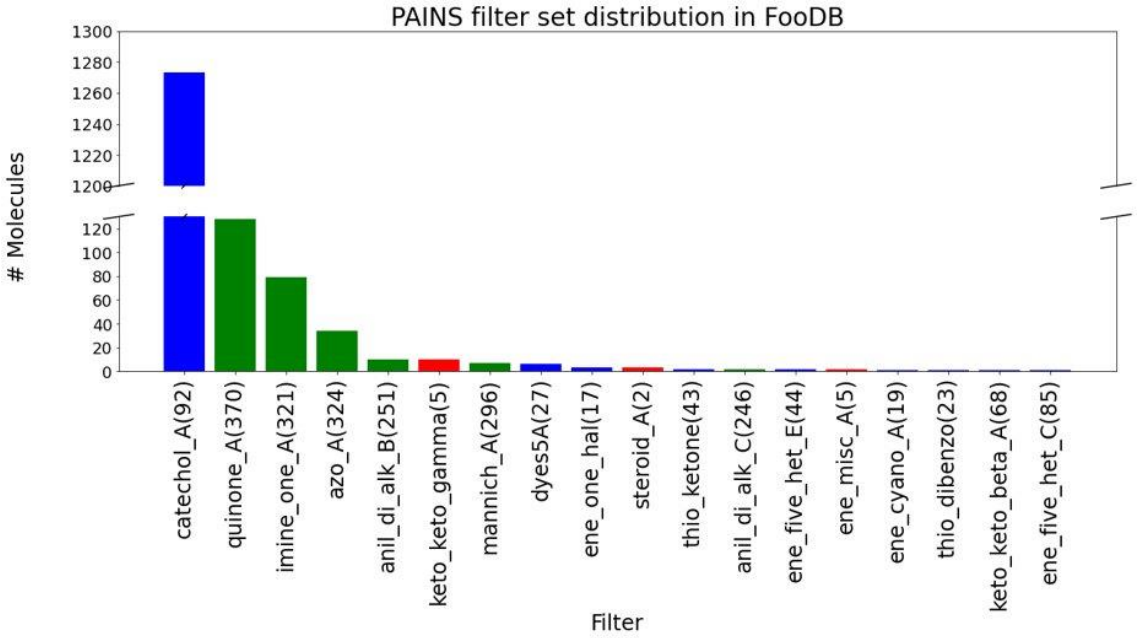

PAINS filter set distribution in FooDB
Molecules
Filter
1300
1280
1260
1240
1220
1200
120
100
80
60
40
20
0
catechol_A(92)
quinone_A(370)
imine_one_A(321)
azo_A(324)
anil_di_alk_B(251)
keto_keto_gamma(5)
mannich_A(296)
dyes5A(27)
ene_one_hal(17)
steroid_A(2)
thio_ketone(43)
anil_di_alk_C(246)
ene_five_het_E(44)
ene_misc_A(5)
ene_cyano_A(19)
thio_dibenzo(23)
keto_keto_beta_A(68)
ene_five_het_C(85)

Figure 4.

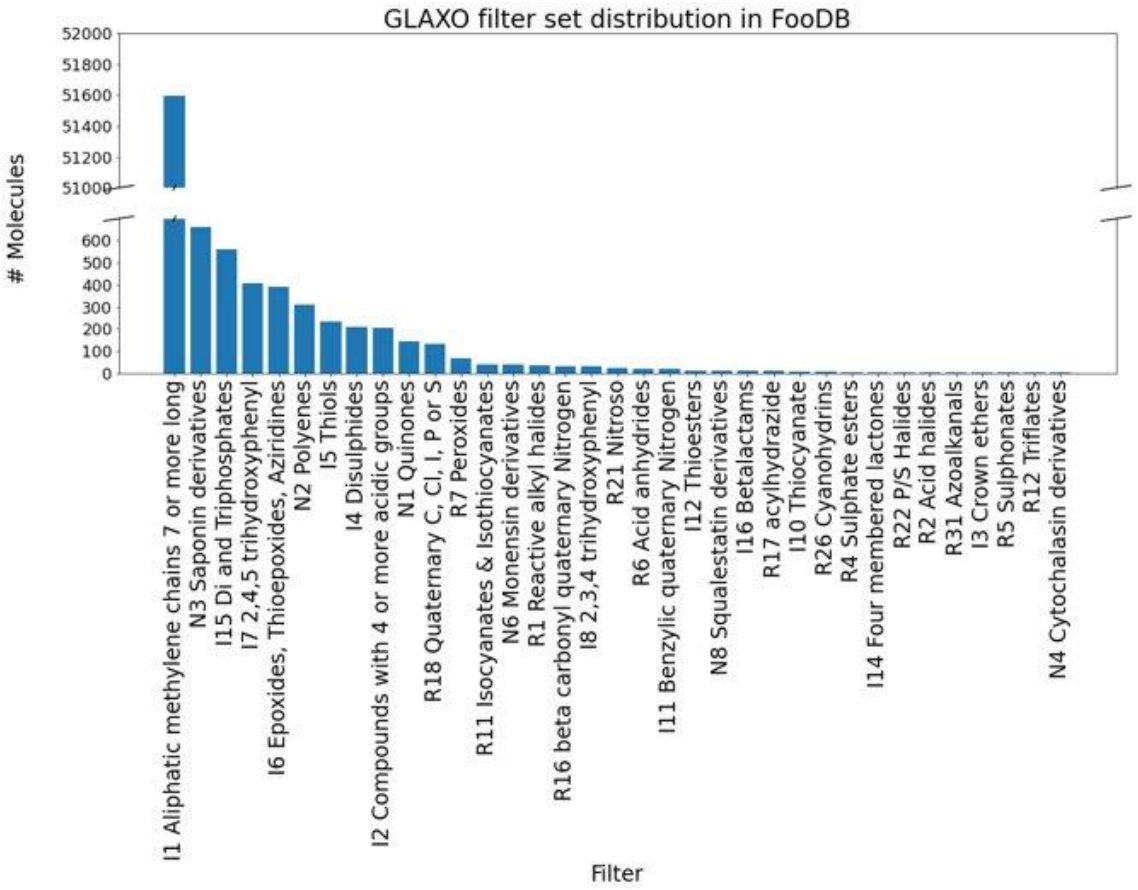
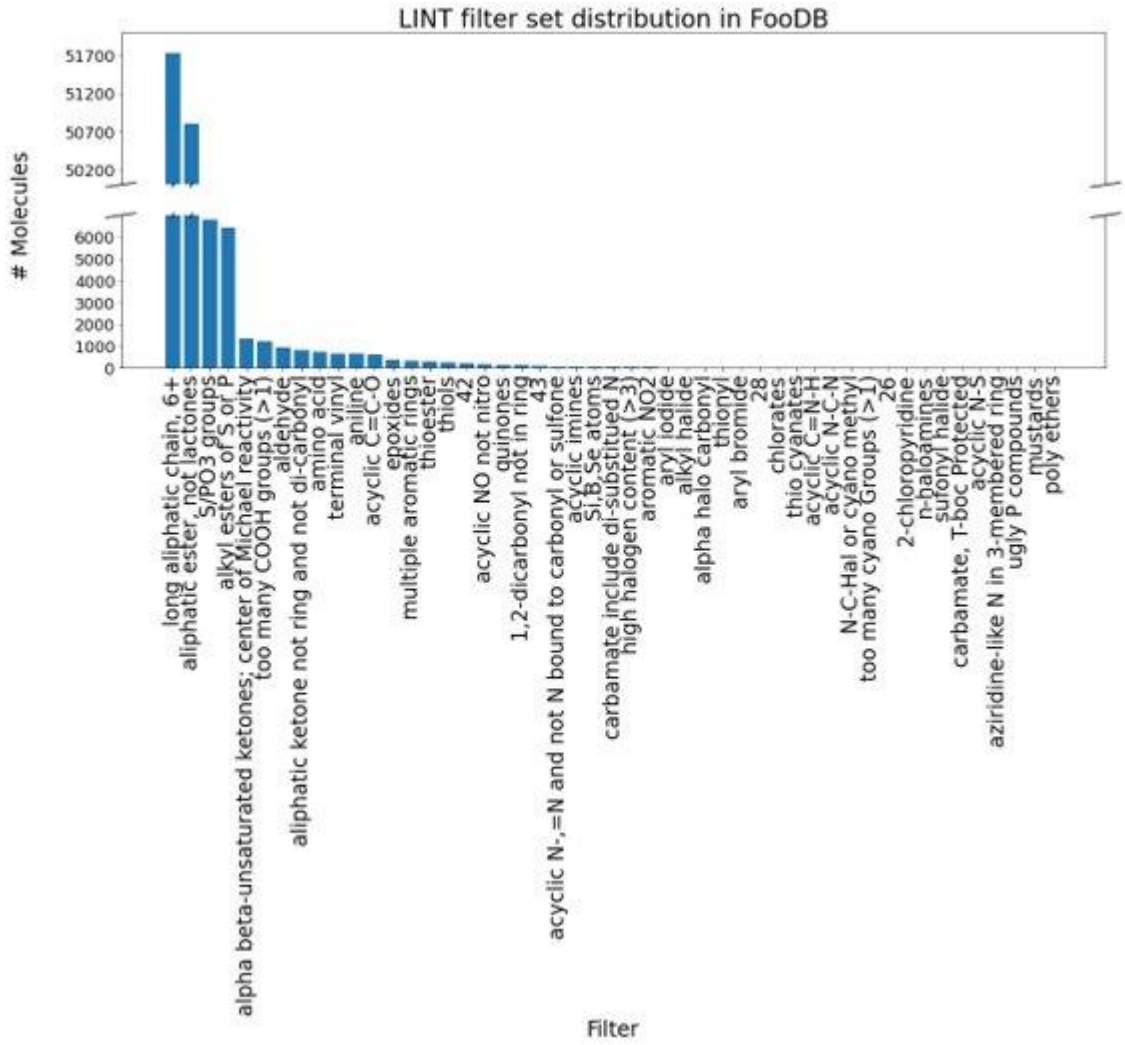

GLAXO filter set distribution in FooDB
Molecules
52000
51800
51600
51400
51200
51000
600
500
400
300
200
100
0
I1 Aliphatic methylene chains 7 or more long
N3 Saponin derivatives
I15 Di and Triphosphates
I7 2,4,5 trihydroxyphenyl
I6 Epoxides, Thioepoxides, Aziridines
N2 Polyenes
I5 Thiols
I4 Disulphides
I2 Compounds with 4 or more acidic groups
N1 Quinones
R18 Quaternary C, Cl, I, P or S
R7 Peroxides
R11 Isocyanates & Isothiocyanates
N6 Monensin derivatives
R1 Reactive alkyl halides
R16 beta carbonyl quaternary Nitrogen
I8 2,3,4 trihydroxyphenyl
R21 Nitroso
R6 Acid anhydrides
I11 Benzylic quaternary Nitrogen
I12 Thioesters
N8 Squalestatin derivatives
I16 Betalactams
R17 acylhydrazide
I10 Thiocyanate
R26 Cyanohydrins
R4 Sulphate esters
I14 Four membered lactones
R22 P/S Halides
R2 Acid halides
R31 Azoalkanals
I3 Crown ethers
R5 Sulphonates
R12 Triflates
N4 Cytochalasin derivatives
Filter

Figure 5.


LINT filter set distribution in FooDB
Molecules
51700
51200
50700
50200
6000
5000
4000
3000
2000
1000
0
long aliphatic chain, 6+
aliphatic ester, not lactones
S/PO3 groups
alkyl esters of S or P
alpha beta-unsaturated ketones; center of Michael reactivity
too many COOH groups (>1)
aldehyde
aliphatic ketone not ring and not di-carbonyl
amino acid
terminal vinyl
aniline
acyclic C=C-O
epoxides
multiple aromatic rings
thioester
thiols
42
acyclic NO not nitro
quinones
1,2-dicarbonyl not in ring
43
acyclic N-,=N and not N bound to carbonyl or sulfone
acyclic imines
Si,B,Se atoms
carbamate include di-substitued N
high halogen content (>3)
aromatic NO2
aryl iodide
alkyl halide
alpha halo carbonyl
thionyl
aryl bromide
28
chlorates
thio cyanates
acyclic C=N-H
acyclic N-C-N
N-C-Hal or cyano methyl
too many cyano Groups (>1)
26
2-chloropyridine
n-haloamines
sufonyl halide
carbamate, T-boc Protected
acyclic N-S
aziridine-like N in 3-membered ring
ugly P compounds
mustards
poly ethers
Filter

755

Figure 6.TABLE OF CONTENTS GRAPHIC