Consensus rank orderings of molecular fingerprints illustrate the 'most genuine' similarities between marketed drugs and small endogenous human metabolites, but highlight exogenous natural products as the most important 'natural' drug transporter substrates. bioRxiv version.

Steve O'Hagan[1,2] & Douglas B. Kell[1,2,3,*]

[1]*School of Chemistry,* [2]*Manchester Institute of Biotechnology,* [3]*Centre for the Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM), The University of Manchester, 131 Princess St, Manchester M1 7DN, UK*

*Corresponding Author: E-mail: dbk@manchester.ac.uk; Tel.: +44-161-306-4492 http://dbkgroup.org/ @dbkell

Abstract

We compare several molecular fingerprint encodings for marketed, small molecule drugs, and assess how their rank order varies with the fingerprint in terms of the Tanimoto similarity to the most similar endogenous human metabolite as taken from Recon2. For the great majority of drugs, the rank order varies very greatly depending on the encoding used, and also somewhat when the Tanimoto similarity (TS) is replaced by the Tversky similarity. However, for a subset of such drugs, amounting to some 10% of the set and a Tanimoto similarity of ~0.8 or greater, the similarity coefficient is relatively robust to the encoding used. This leads to a metric that, while arbitrary, suggests that a Tanimoto similarity of 0.75-0.8 or greater genuinely does imply a considerable structural similarity of two molecules in the drug-endogenite space. Although comparatively few (<10% of) marketed drugs are, in this sense, robustly similar to an endogenite, there is often at least one encoding with which they are genuinely similar (e.g. TS > 0.75). This is referred to as the Take Your Pick Improved Cheminformatic Analytical Likeness or TYPICAL encoding, and on this basis some 66% of drugs are within a TS of 0.75 to an endogenite.

We next explicitly recognise that natural evolution will have selected for the ability to transport dietary substances, including plant, animal and microbial 'secondary' metabolites, that are of benefit to the host. These should also be explored in terms of their closeness to marketed drugs. We thus compared the TS of marketed drugs with the contents of various databases of natural products. When this is done, we find that some 80% of marketed drugs are within a TS of 0.7 to a natural product, even using just the MACCS encoding. For patterned and TYPICAL encodings, 80% and 98% of drugs are within a TS of 0.8 to (an endogenite or) an exogenous natural product. This implies strongly that it is these exogeneous (dietary and medicinal) natural products that are more to be seen as the 'natural' substrates of drug transporters (as is recognised, for instance, for the solute carrier SLC22A4 and ergothioneine). This novel analysis casts an entirely different light on the kinds of natural molecules that are to be seen as most like marketed drugs, and hence potential transporter substrates, and further suggests that a renewed exploitation of natural products as drug scaffolds would be amply rewarded.

Keywords

drug transporters – cheminformatics – endogenites – metabolomics –encodings

Introduction

Given the overwhelming evidence [1-20] that pharmaceutical drugs must and do exploit endogenous transporters that normally transport biological metabolites, and that normally any diffusion of such drugs through the phospholipid bilayer portions of undamaged biological membranes is negligible [1; 3; 5-7; 10; 11; 13; 21], we [2; 22-24] and others (e.g. [16; 25-30]) have been assessing the extent to which marketed (hence successful) xenobiotic drugs are similar in structural terms to endogenous human metabolites (that we sometimes refer to as 'endogenites').

Chemical similarity is a slippery concept (see e.g. [31-35] and below) but, leaving aside descriptor-based vectors [36], it is most commonly assessed by encoding the molecules of interest into one or more fingerprints expressed as bitstrings, then comparing the bitstrings, again most commonly in terms of their Jaccard or Tanimoto similarity [37-39]. Our first detailed study [22] noted that the quantitative (Jaccard/Tanimoto) similarity varied markedly with the different (fingerprint-based) encodings used (and we reproduce the essential and Open Access findings in Fig 1A, below), just as does the appearance or otherwise of 'activity cliffs' [40-42]. To a certain degree, the shape of the profiles of rank-ordered drugs vs their Tanimoto similarity to the closest endogenous metabolite were smooth curves that differed somewhat. However, this of itself did not tell us – notwithstanding the numerical variation in Tanimoto similarity with each encoding – whether the rank order of individual drugs themselves was more or less well preserved for each encoding. In other words was the drug that was numerically most similar to an endogenite under the MACCS encoding also most similar under (say) the Atom Pair encoding?

The Tanimoto similarity is a true metric, and while it returns a numerical value between 0 and 1 the question also arises as to which values of the Tanimoto similarity genuinely count as 'significantly similar' [34] from a utilitarian point of view. Unlike QSAR and other 'supervised' methods where there is an objective function, for which the predictions of the model can be tested on unseen data (e.g. where a Tanimoto similarity of 0.85 to a 'hit' in a drug discovery assay increases the chance of another hit by 30-fold [43]), the pure notion of chemical similarity is really an 'unsupervised' method, and its numerical value is simply that.

Previously, apart from the addition of vitamins, we were rather restrictive about what might constitute an endogenous metabolite or 'endogenite', and we here recognise that this restriction was not only unnecessary but potentially very misleading, as any natural molecule with a high k_{cat} or k_{cat}/K_m [44-47] for a particular transporter might reasonably be regarded as a 'natural' substrate for it. In particular, we may suppose that there are or have been natural, bioactive/psychoactive dietary and medicinal components (and their and other microbiome-derived products) that are both beneficial and common enough that the host has essentially been exposed to them more or less regularly through evolutionary time, albeit they do not appear in the common models of human metabolism. Since useful bioactivity in tissues implies uptake, natural selection would then ensure that we had actually evolved transporters for them, and that these molecules, despite not being synthesised by the host, are properly to be seen as 'natural substrates' of such transporters. L-ergothioneine is a particularly clear example of this.

Some mammalian transporters with known selectivity for exogenous natural products

L-ergothioneine (2-mercaptohistidine trimethylbetaine; IUPAC name (2S)-3-(2-Thioxo-2,3-dihydro-1H-imidazol-4-yl)-2-(trimethylammonio)propanoate) is not synthesised by mammals, but exists in a wide range of foodstuffs (especially mushrooms) and may be highly concentrated in mammalian tissues [48; 49]. Several types of evidence imply that it has an important role *in vivo* as a natural antioxidant [50; 51]. First this activity may be measured directly [52-54]. Secondly, decreasing it leads to the accumulation of the products of the interaction of macromolecules with

hydroxyl radicals [55-57] and a decreased lifespan in model organisms [58]. Thirdly, it acts as a cytoprotectant [48; 59-63]. Our interest in it here comes from the fact that it has been found to be the natural (or at least most active) substrate of the concentrative, Na^+-dependent transporter SLC22A4 [64-66] (once referred to as OCTN1, now the ergothioneine transporter, which is also capable of transporting drugs such as the antidiabetic metformin [67]) (for SLC terminology see [68] and http://boparadigms.org/). It was known that OCTN1 transported organic cations, but not what the 'natural substrate' might be. Thus, Gründemann and colleagues [64] incubated cells with and without recombinant OCTN1 transporter expression in paired assays with diluted plasma (taken to contain all candidate substrate molecules) and compared differences in the uptake of the various compounds by mass spectrometry. The first substance identified was proline betaine (stachydrine). Subsequent tests on structurally related molecules showed that ergothioneine was much the best substrate, with an uptake activity almost 100-fold higher than those for tetraethyl ammonium and carnitine [64] (that were previously believed to be the 'main' substrates), and that cells lacking the transporter were virtually impermeable to ergothioneine. Since it does not seem to be essential for the growth of the host it has not attained the status of a vitamin, but it is clearly highly beneficial. (Its presence in almost all foodstuffs means that starvation for it specifically, the usual means of discovering or identifying a vitamin, has probably never occurred.) The same may generally be said to be true of other nutritionally beneficial molecules of plant origin, of which the flavonoids are among the best known.

Indeed, in a similar way, it appears that specific transporters for flavonoid-type molecules also exist [69-71], albeit their molecular taxonomy remains unclear [72]. This said, a transporter in plants [73] shows significant homology to bilitranslocase, a liver uptake transporter for blood-derived bilirubin, and bilitranslocase has been shown to transport dietary flavonoids [74], in particular anthocyanins [75-78]. Thus we are led to the view that we should consider as substrates for mammalian transporters not only the known intermediary metabolites, but also a variety of (mainly plant and microbial) dietary molecules that are bioactive and beneficial, even if not essential. This is because organisms will have coevolved with them for millions of years since 'animals' began to consume plants [79-82] and to harbour microbes [83; 84]. Even stronger natural selection may be expected since the time that such plants actually began to be utilised in agriculture [85] or prescribed for medical benefit [86; 87], as in Ayurvedic [88-90] and Chinese Herbal Medicine [89; 91; 92] (ca 5-8000y BP). If this is the case, we would expect to find even more structural similarities between drugs and such natural products when these are compared to drug-endogenite similarities (and actually this proved to be the case in a pilot study; Fig 5C of [22]). One purpose of the present paper was to test this idea explicitly and in much more detail. Indeed, it transpires (see a detailed analysis in the body of this paper) that many of the least human-endogenite-like marketed drugs are considerably closer in structure to common plant and microbial secondary products than they are to endogenites. If we take the term 'natural substrate' to mean a substance to which an organism has been exposed and for which a transporter has a particularly high k_{cat} or k_{cat}/K_m, it is reasonable to refer to such a molecule as a 'natural substrate', as in the ergothioneine/SLC22A4 example above. Another exogenous molecule that seems similarly valuable to mammals [93-104] and other organisms [105], albeit its uptake transporter is

not yet known, is pyrroloquinoline quinone (PQQ) [106], also known as methoxatin, a redox cofactor normally associated with prokaryotes [107; 108].

How similar is similar?

Although the concept of what counts as 'significantly similar' must be recognised as highly important in cheminformatics, there has been surprisingly little work done on it; most of it has involved assessing the likelihood that a given similarity could be achieved from a (more or less random) distribution of chemicals [34; 109-113]. Given that our original, underlying interest is in understanding those features of drug and endogenous metabolite structures that tend to determine whether a drug is closer to or far from being most similar to a specific endogenous (or other) metabolite, the question is important (but the distribution of chemical structures is far from random, one having been selected by natural evolution, the other via the processes of drug discovery). Thus, the first part of the present paper analyses that question. The conclusion is that the rank order is reasonably preserved for only a small fraction – some 5-10% – of those drugs that are most similar to an endogenite, but that for the vast majority of drugs not only the numerical value of the (Tanimoto) similarity but also the rank order depends very strongly indeed on the encoding used. However, the fraction of drugs for which different fingerprinting methods of encoding do give consensual answers (Tanimoto similarity ≥ 0.8, for instance) provides a defensible cut-off for what really counts as 'significantly similar'. This leads to a second part, where we establish that plant- and microbially derived natural products have a much greater similarity to marketed drugs than do the endogenous metabolites of Recon2, and that they are in fact almost certainly the more common 'natural' substrates of the transporters on which pharmaceutical drugs hitchhike. This has profound implication for our understanding of the nature and evolution of human drug transporters.

Experimental

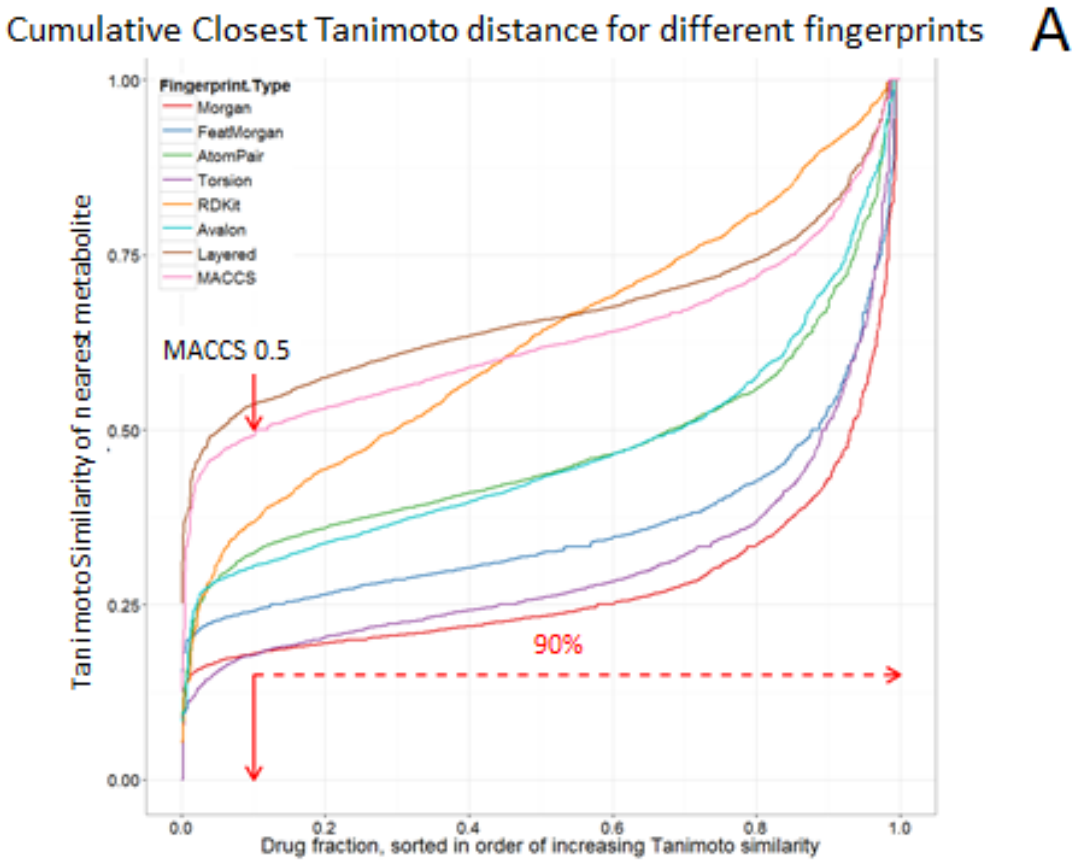
As previously [22-24; 114], we used the list of 1381 marketed drugs and 1113 Recon2-based endogenous metabolites as provided in the Supplementary information to [22]. A number of natural products and other databases exist [115-120]. We have here used the dataset for measured serum metabolites kindly provided by Prof David Wishart and colleagues [121], but removed all substances marked as drugs or that were in recon2. In addition, where noted, we also studied datasets such as UNPD http://pkuxxj.pku.edu.cn/UNPD/ [122] and ZINC [123; 124]. We also obtained a license for the (commercial) Dictionary of Natural Products [125] http://dnp.chemnetbase.com/intro/. All comparisons were done using KNIME-based workflows ([126-128] and www.knime.org/), and in particular we made use of the RDKit nodes [112; 129] (http://rdkit.org/).

Results and Discussion

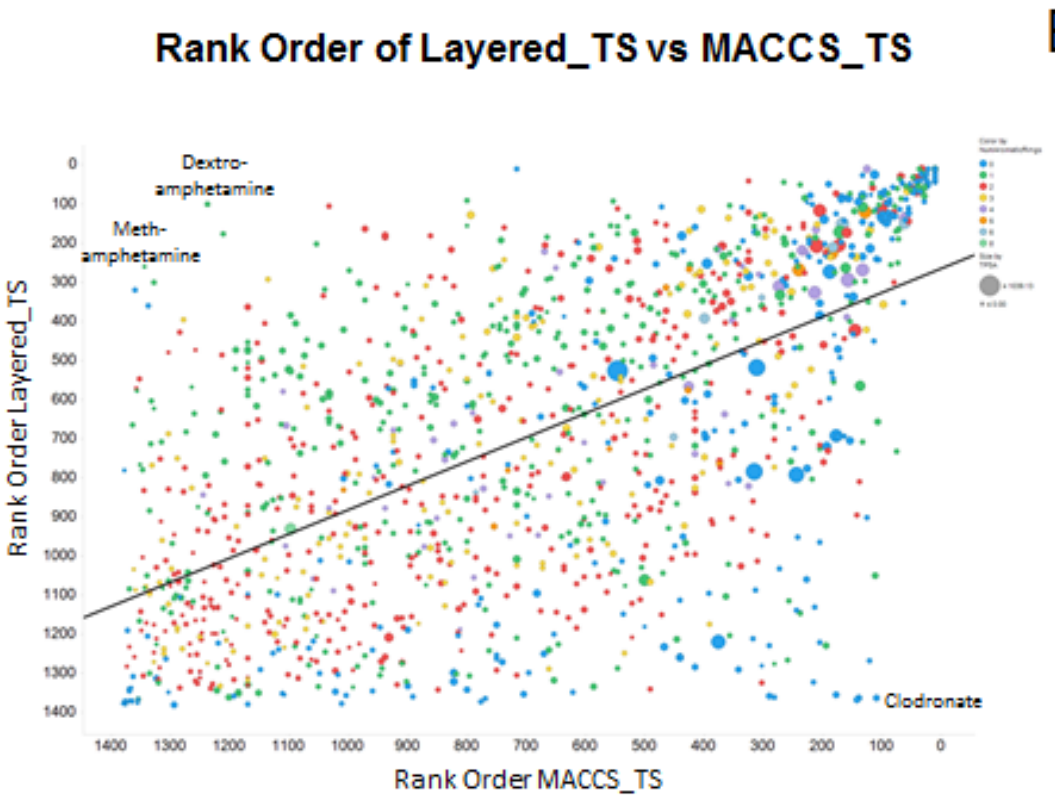
Variance in 'similarity' with different fingerprint encodings

Leaving aside molecules that are actually both drugs and metabolites, some drugs are clearly much more similar to one or more endogenous metabolites than are others, and this is true for a variety of fingerprint encodings [22-24] as provided via RDKit [112; 129]. The question thus arose as to whether these similarities extended to the actual rank orders of the drugs (with 1 always

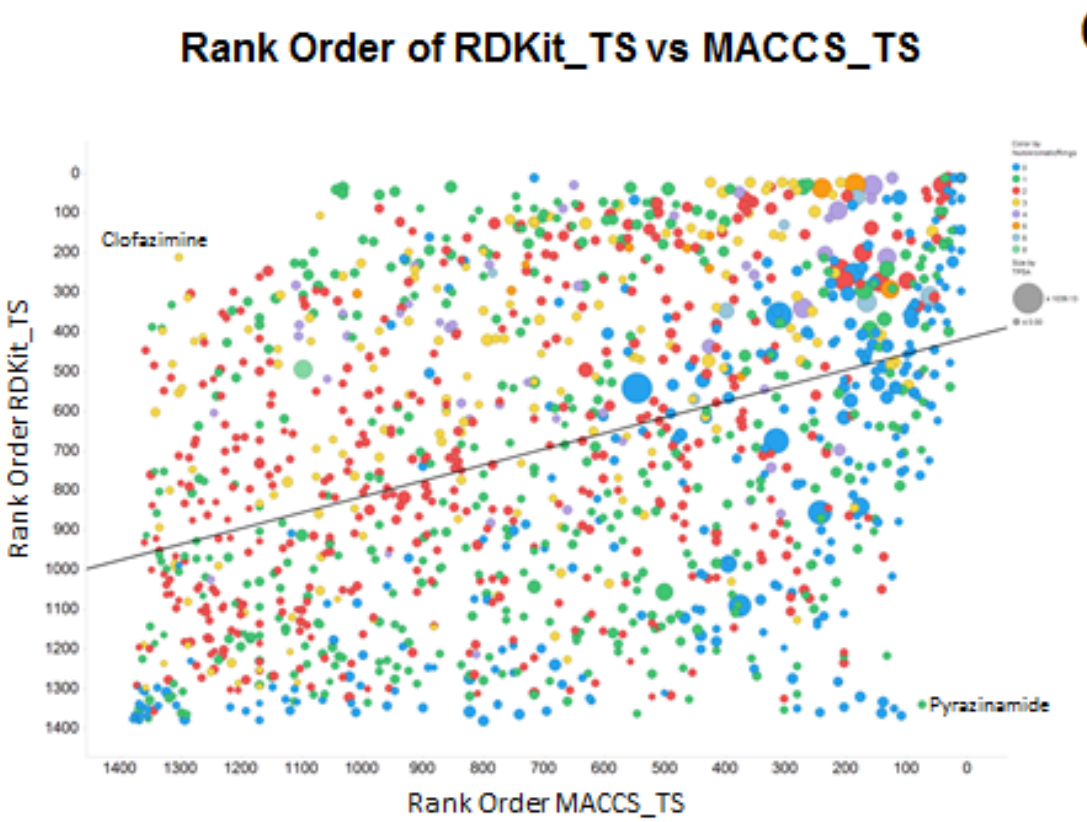
being the drug most similar to an endogenite). In other words, was the drug that was most similar to an endogenite when these were represented using the MACCS encoding also most similar with say the Atom Pair encoding? For ease of assessment, Fig 1A recapitulates the original analysis [22] (freely available under a CC-BY license). The three encodings that seemed to maximise the endogenite-likeness of marketed drugs in the earlier paper [22] were the MACCS, Layered and RDKit encodings in RDKit. Thus Fig 1B and 1C show, respectively, the relative rank orders of Layered and RDKit vs MACCS, all using Tanimoto as the metric of similarity. It is clear that while a small subset of the most endogenite-like drugs preserve their rank order between encodings, the rank order for the vast majority depends very strongly on the encoding used (cf. [112]). Also shown for Layered and RDKit (Figures 1B, 1C) are the names of a few drugs for which the differences in rank order are most extreme. The same kind of phenomena are true for Torsion vs MACCS (Fig 1D) and indeed for all the other comparisons tested (data not shown, but all of these data are provided as a spreadsheet via the Supplementary information (DvsMDrugRanks_Full_w_descriptors_hits_with_MACCS_TS.xls)). Overall, while the generation of fingerprints is entirely deterministic, we could discern no real molecular properties that would predict which TS values for a given drug would be 'high' or 'low' for the set of endogenites. This could be seen as giving weight to view that each is of value and might be used as required.



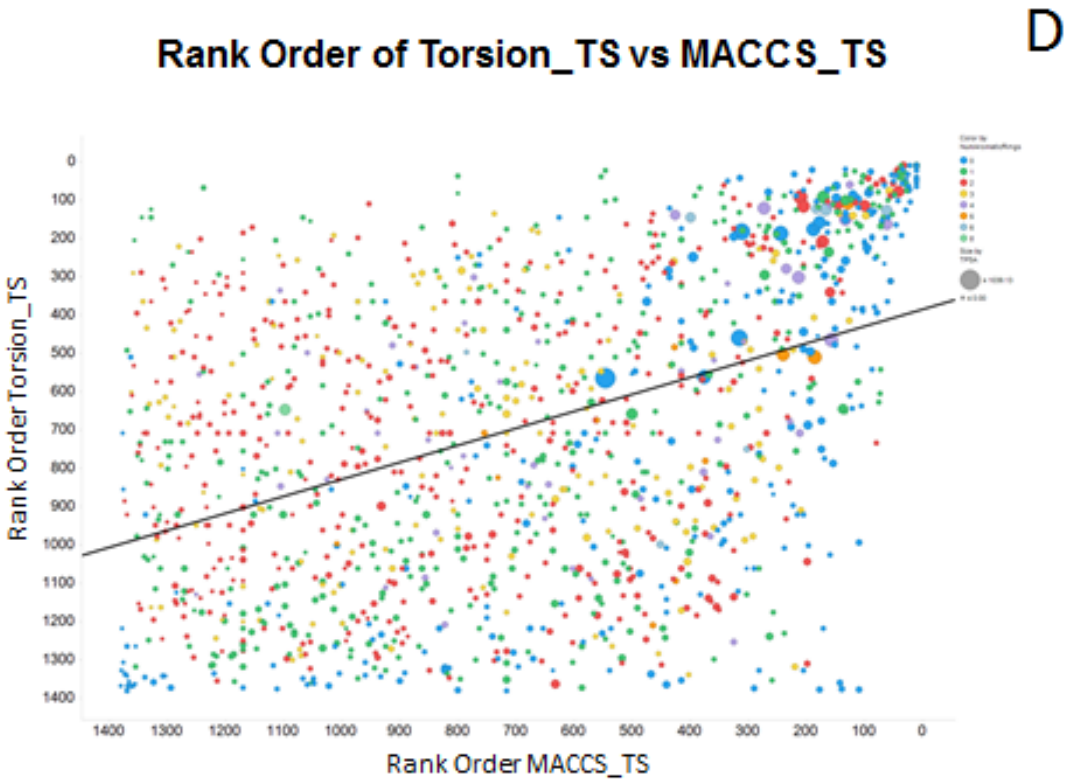
B



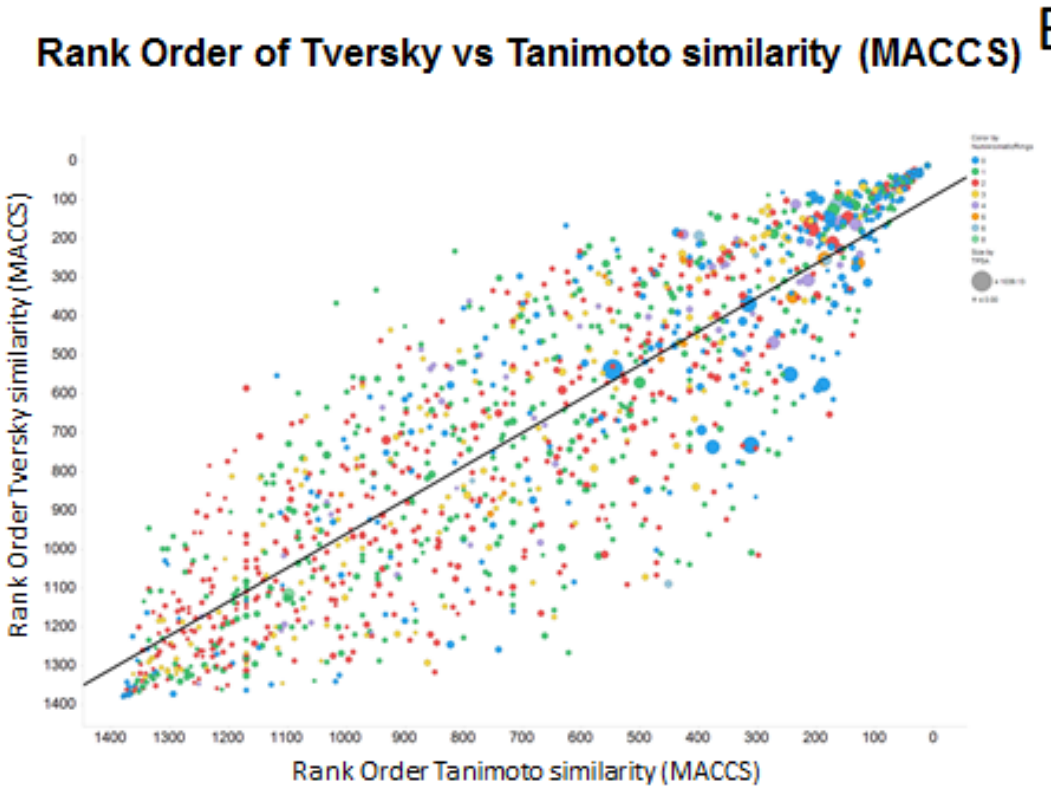
C



D



E



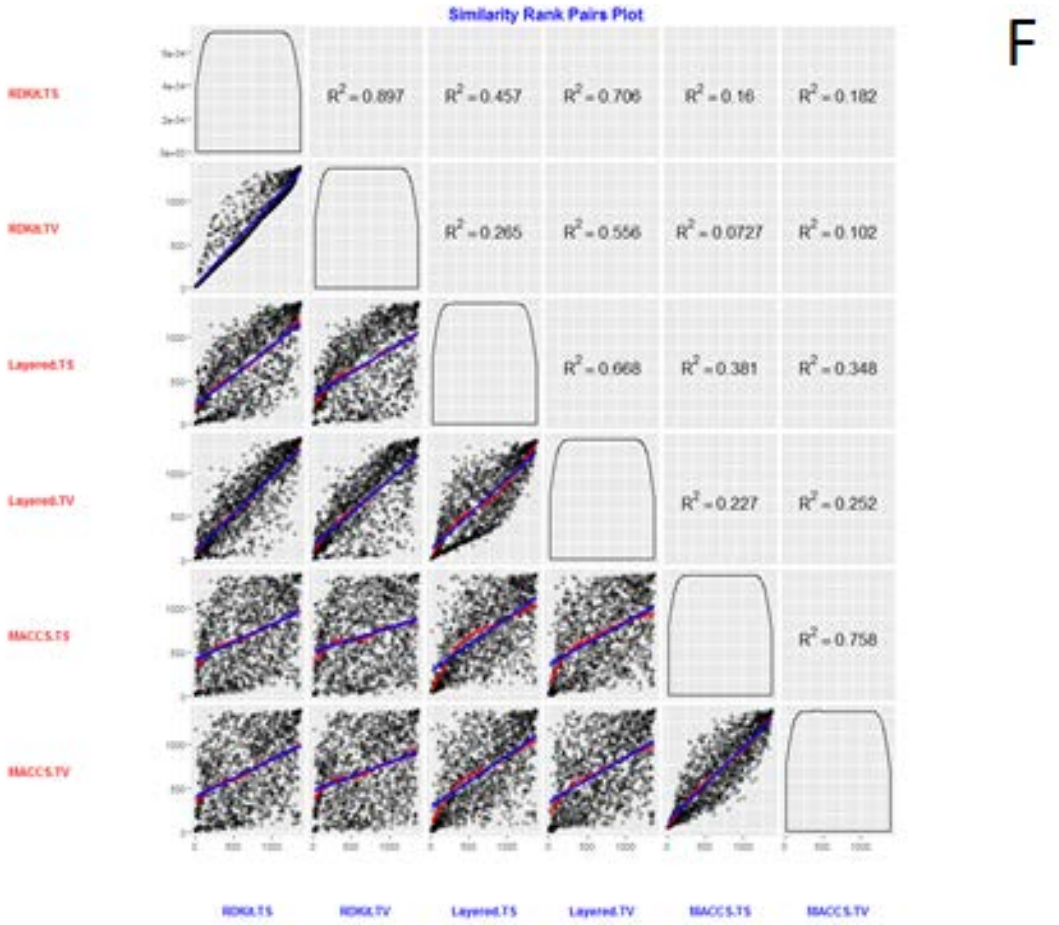
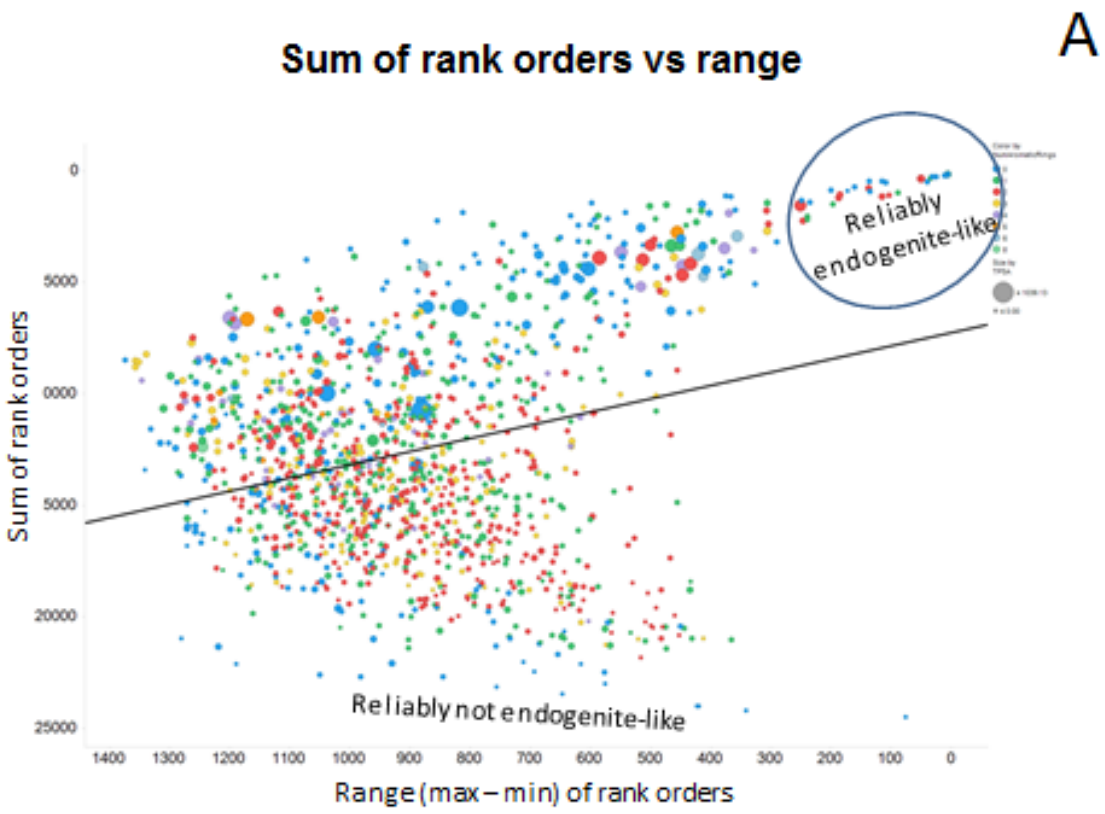


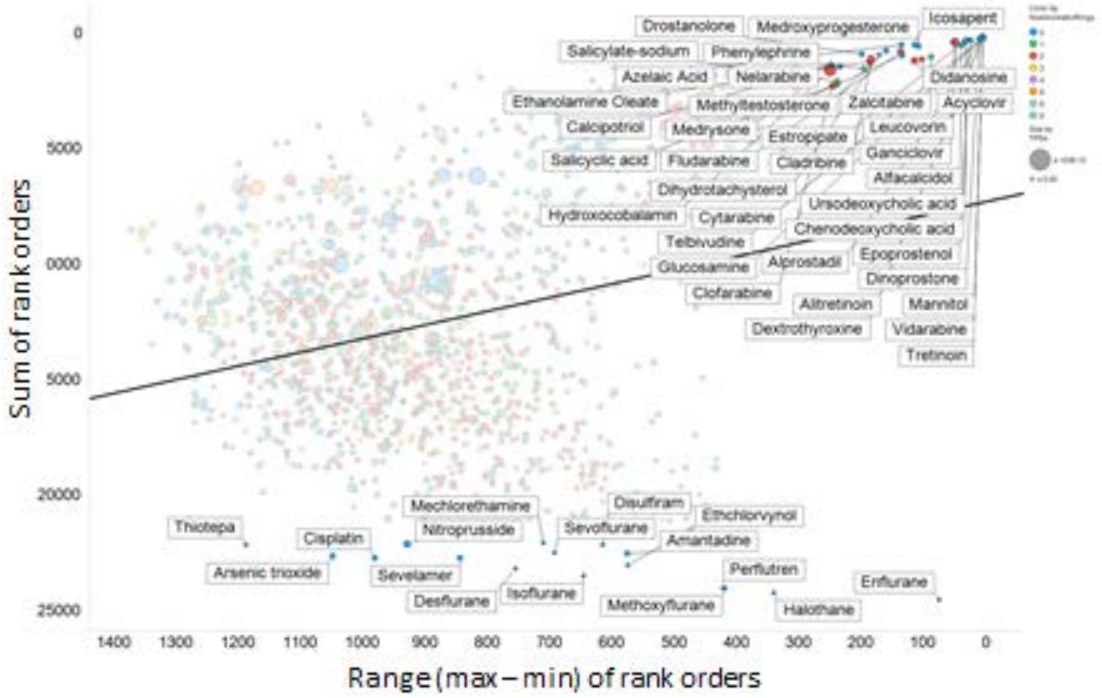
Figure 1. Cumulative similarity and rank order of various encodings. ***A****. Cumulative rank order (most similar on the right) of a drug to its closest endogenite for a series of encodings, Redrawn (under a CC-BY license) from [22]. For the other parts of this figure, each symbol represents the rank order of the encodings specified. In addition, although there were no observable trends, symbol size encodes total polar surface area, while colour encodes the number of aromatic rings in the drug (0 blue, 1 emerald, 2 red, 3 yellow, 4 lilac, 5 orange, 6 sapphire, 8 cyan), and these can help to identify individual molecules in different encodings.* ***B****. Layered vs MACCS encoding, Tanimoto similarities,* r^2 *= 0.38.* ***C****. RDKit vs MACCS encoding, Tanimoto similarities,* r^2 *= 0.16.* ***D****. Torsion vs MACCS encoding, Tanimoto similarities,* r^2 *= 0.20.* ***E****. Tversky (*α *=0.2,* β *=0.8) similarity vs Tanimoto similarity for MACCS encoding.* r^2 *= 0.76.* ***F****. Plot of multiple comparisons (blue best linear fit, red best LOESS fit).*

We also compared the Tversky similarities (α =0.2, β =0.8) (see [24; 114]) for the different encodings, with Fig 1E illustrating its comparison with the rank-ordered Tanimoto similarity for the MACCS encoding. It may again be concluded that while some drugs appear numerically similar to a given metabolite under the different metrics, many do not. However, in this case the correlation (r^2 = 0.76) is considerably better than that for comparisons of the different encodings. Finally, we illustrate several correlation plots together (Fig 1F).

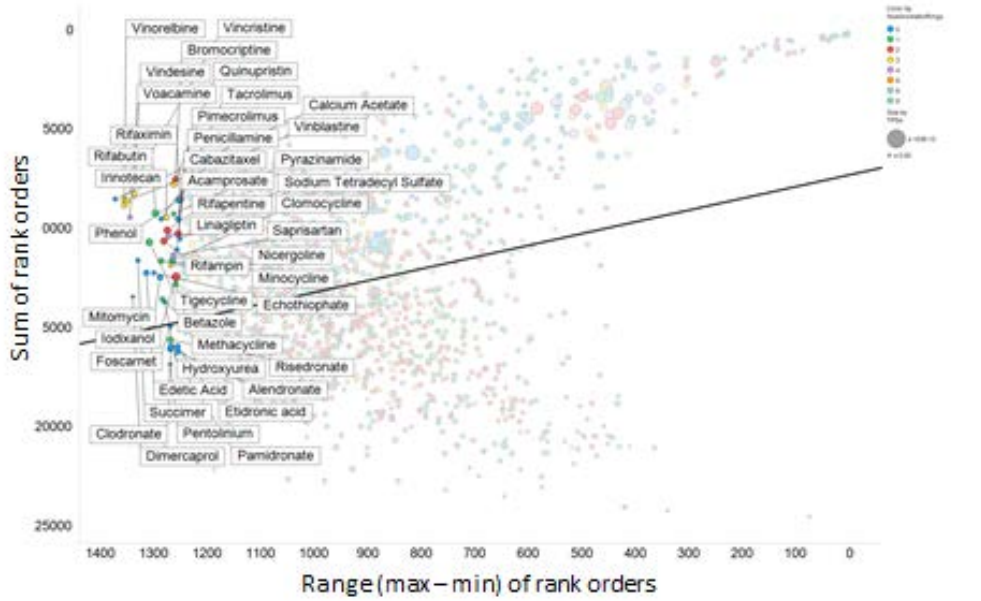
To encapsulate and to summarise all of the RDKit encodings used in one graph, we compared the sum of all the rank orders with their range (Fig 2A). Thus those at the top right of the plot (Fig 2B) are those drugs that are reliably of high rank order (most similar) whatever the encoding; there are only 44 where the cut-off was (somewhat arbitrarily) drawn. Similarly, a small subset are reliably of low rank order whatever the encoding (and include in particular 'drugs' such as fluorinated inhalational anaesthetics that are clearly very far from endogenites) (Fig 2B). Another subset (arbitrarily picked and illustrated in Fig 2C) contains drugs that are mainly not seen as very endogenite-like except in one or two encodings. However, it is obvious that for the vast majority of other drugs the rank order (and hence endogenite-likeness) depends very strongly upon the exact encoding used. For these, endogenite-likeness is not therefore a property of the drug *per se* but additionally (even particularly) of its encoding into whichever fingerprint is chosen. By contrast, the top 10% or so of drugs, that are within a MACCS Tanimoto similarity of ~0.8 to at least one endogenite, are relatively robust to the different encodings (Fig 2D), and one could argue that this relative independence from the nature of the encoding does seem to be a good metric of "similarity". Although this is something of a self-fulfilling prophecy, inspection of those drugs also clearly does show a metabolite-likeness, especially to endogenites such as nucleobases and sterols (Fig 2E). In a similar vein, although individual encodings can vary significantly, there is a good correlation between the average Tanimoto similarity (for the Morgan, FeatMorgan, AtomPair, Torsion, RDKit, Avalon, Layered, MACCS and Pattern encodings in RDKit) and the drugs that have a Tanimoto similarity ≥ 0.8 in the MACCS encoding (our standard benchmark) (Fig 2F). In Fig 2F, the overall correlation (r^2) = 0.77 (slope = 0.75), and the variance is much less than that of the rank order.



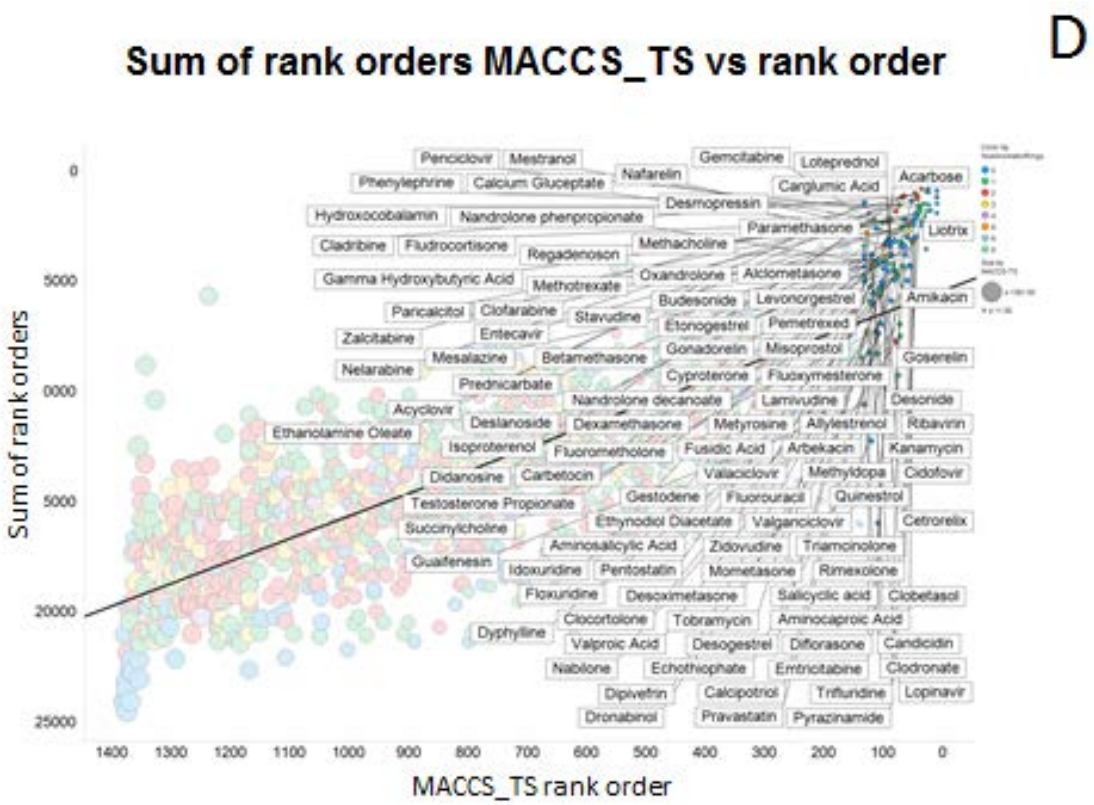
B



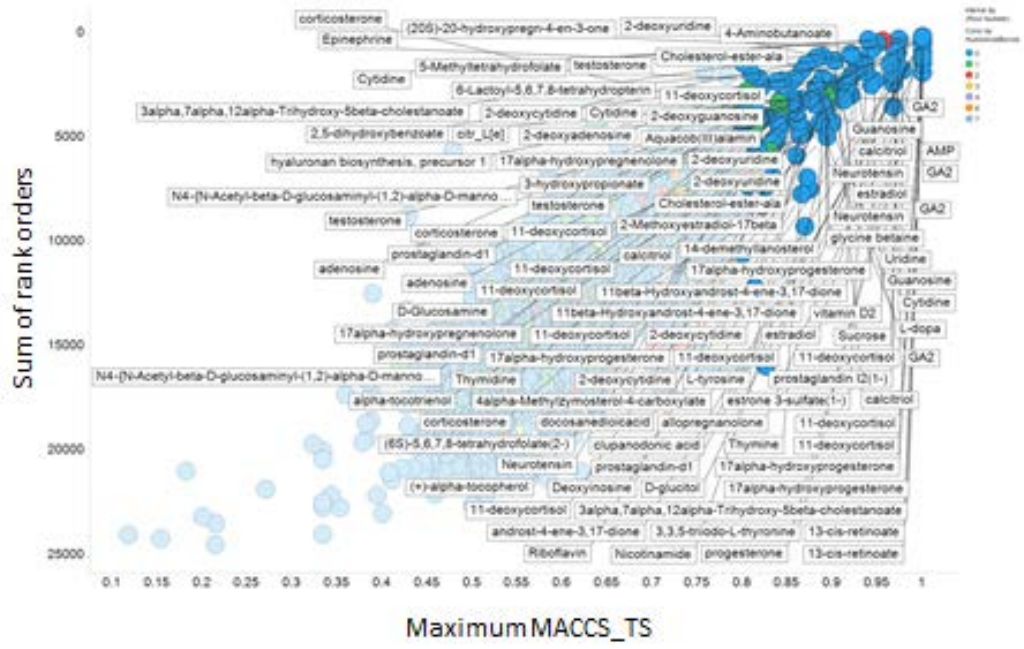
C



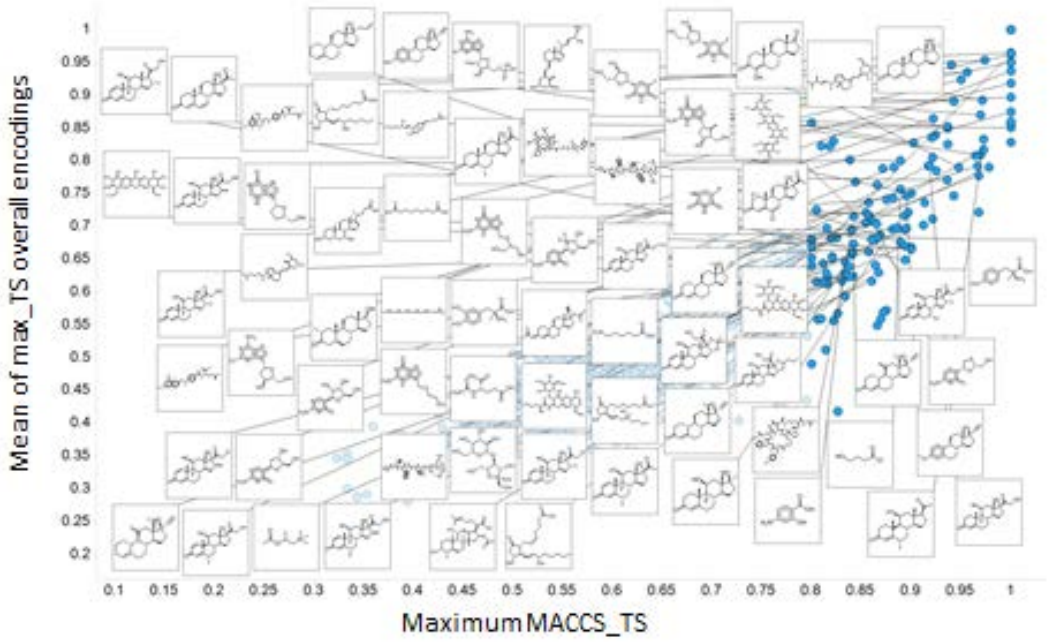
D



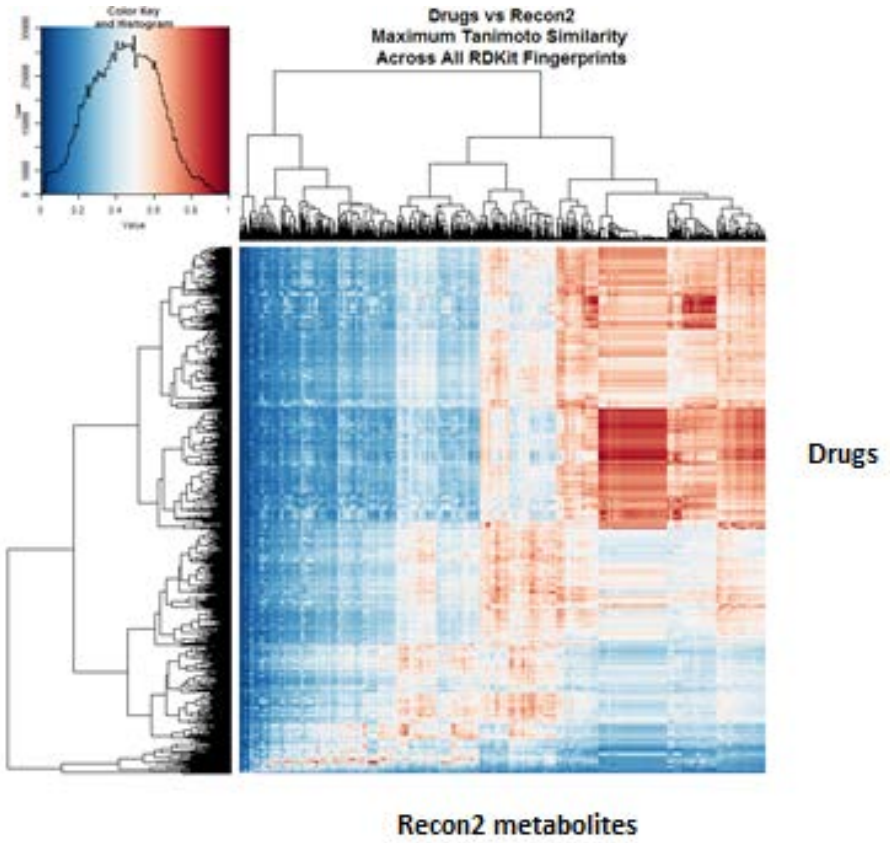
E



F



G



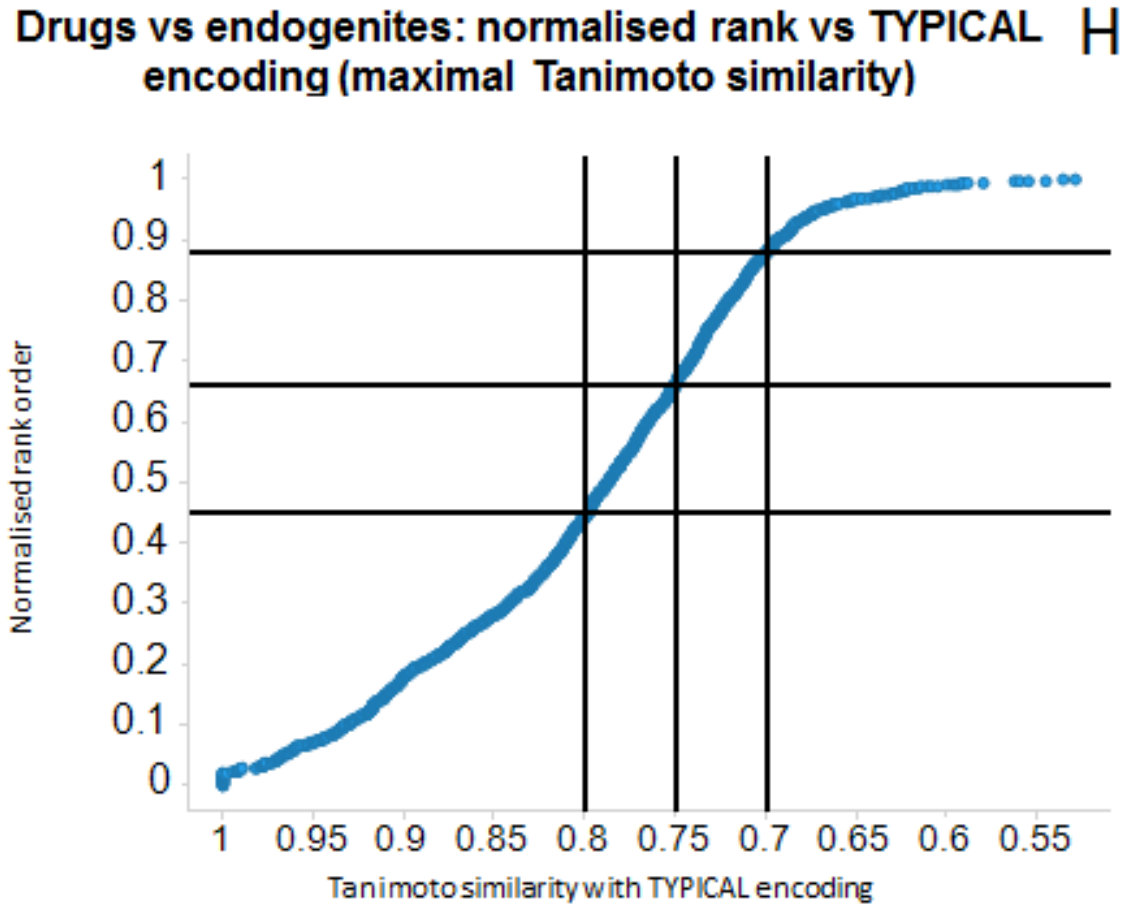


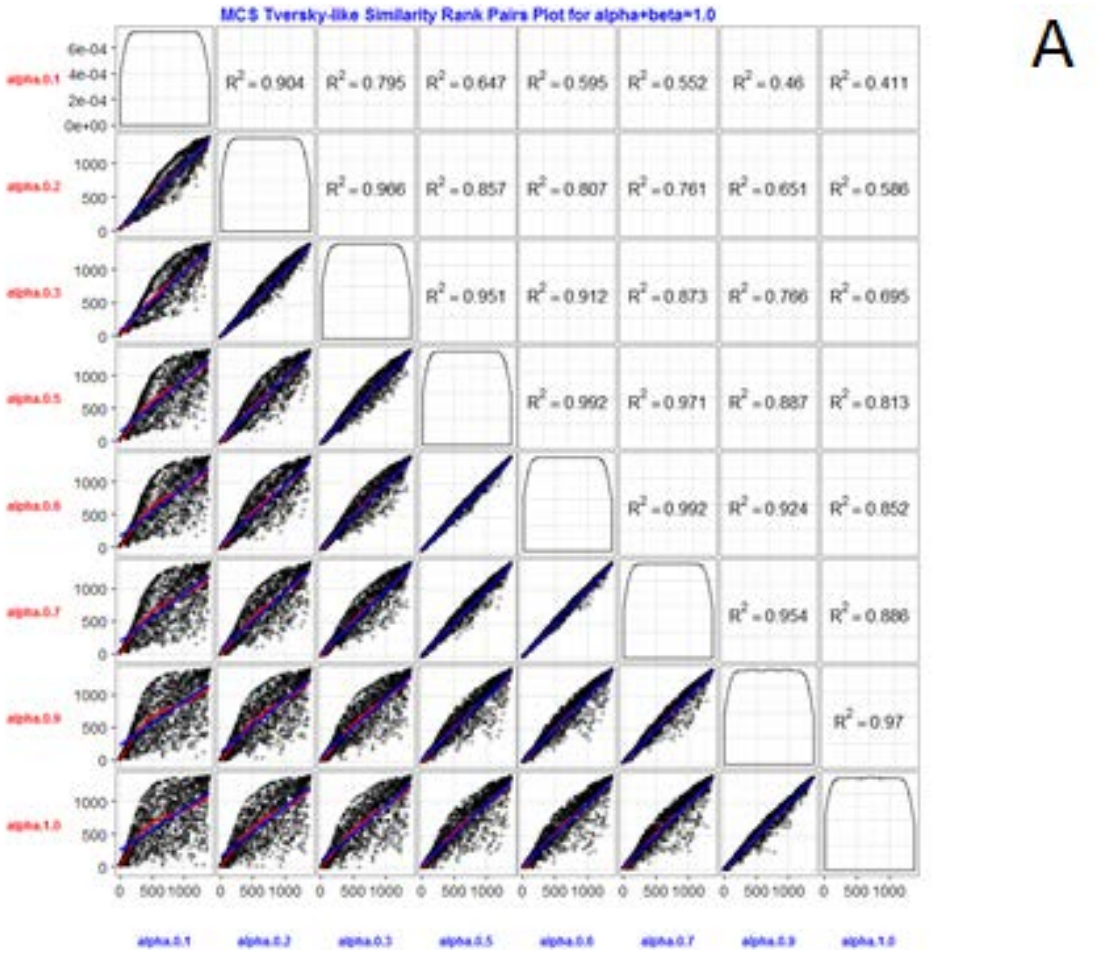
Figure 2. Relationship between the sum and the range of the rank order of the different encodings. A. Overview of the shape of the plot. B. names of marketed drugs that are reliably most or least like an endogenite whatever the encoding. C. Names of drugs for which there is at least one reasonably high rank order but for which mostly they are not encoded as that endogenite-like. D. Names of top 138 drugs (for which TS ≥ 0.8) judged by MACCS similarity in rank order. E. Names of metabolites most similar to the most metabolite-similar 138 drugs (for which TS ≥ 0.8) as judged by MACCS similarity. F. Average value of TS for multiple encodings vs MACCS-encoded Tanimoto similarity. G. Heatmap of similarities of drugs vs endogenites using the TYPICAL encoding. H. Cumulative plot of heatmap data of G.

Choosing the closest encoding for each comparison

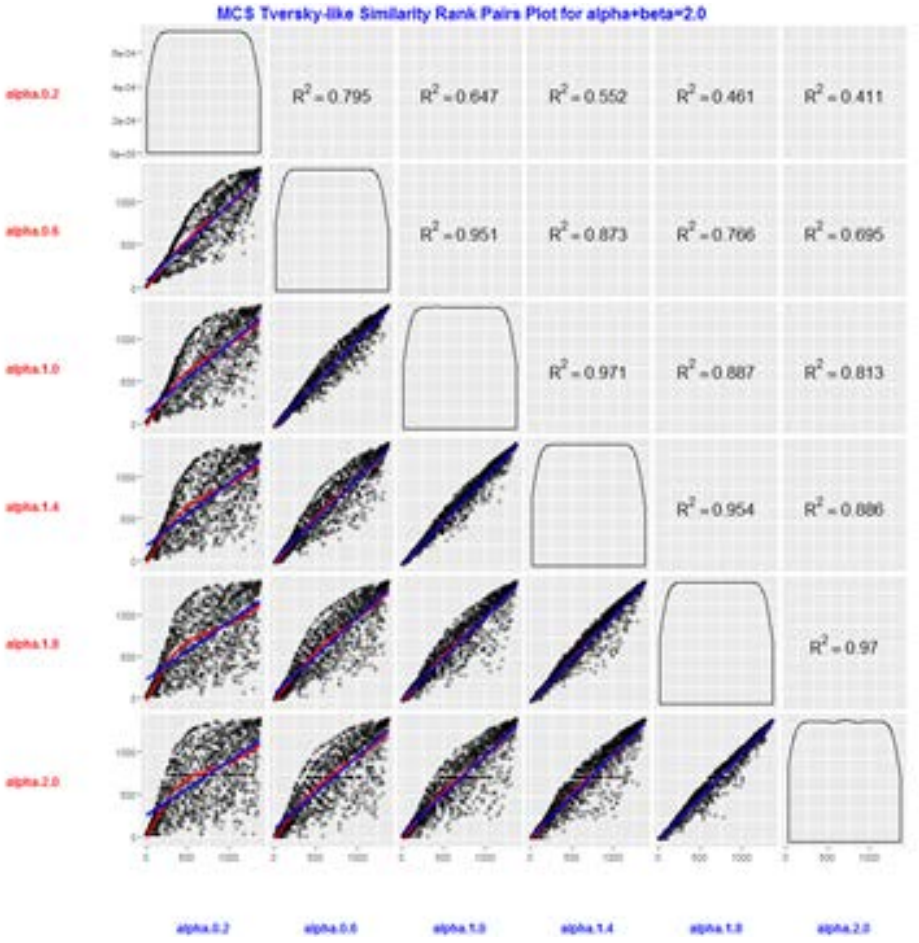
Inspection of figures 2A-2C shows a very considerable range for the majority of molecules, implying that for each molecule there is at least one encoding that is seen as having an especially close value of the Tanimoto similarity for a particular drug-endogenite pair. This best or largest value is here referred to as the Take Your Pick Improved Cheminformatic Analytical Likeness or TYPICAL encoding/similarity. Fig 2G shows a heatmap of the similarities of drugs and metabolites using the TYPICAL encoding (four molecules are dropped because of a curiosity with the Torsion encoding). Under these circumstances, the percentages of drugs having a TYPICAL similarity to an endogenite of 0.8, 0.75 and 0.7 are, respectively, 45%, 66% and 88%, as may also be observed in the cumulative plot of Fig 2G shown in Fig 2H.

Use of the maximum common substructure

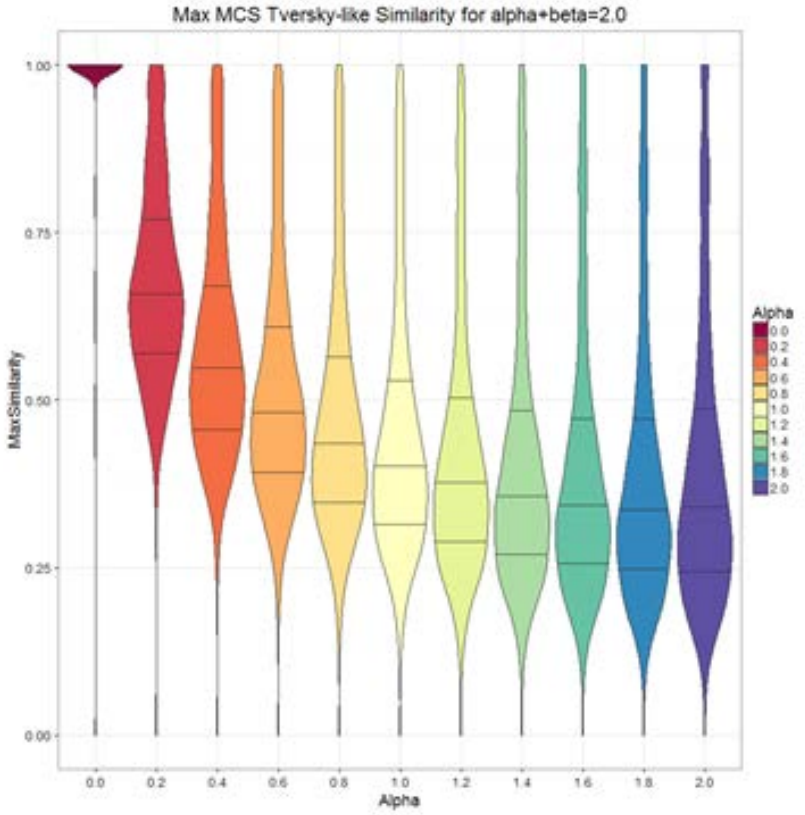
Another means of comparing structural similarities (and hence rank orders), and one that does not depend nearly as much (but see [130]) on the fingerprint encoding used, is according to the size of their maximum common substructure (MCS). As before [24], we have here done this using a series of values of the Tversky similarity, varying the Tversky similarity parameters (α and β) such that their sum was either 1 (Fig 3A) or 2 (Fig 3B). Since the encoding is the same, the correlations between the rank orders for different values of α and β are much higher than for the different encodings, with a clear trend of similarities being visible in the violin plot of Figure 3C. Finally, here, we illustrate a comparison of the MCS with a Dice coefficient ($\alpha = \beta = 0.5$) and the MACCS_Tanimoto; again for the drugs with the highest values of TS to a metabolite (we illustrate those over 0.85 this time) there is a clear consistency of the metabolite-likeness of their fingerprint-based MACCS encoding and their MCS with a Tversky similarity.



B



C



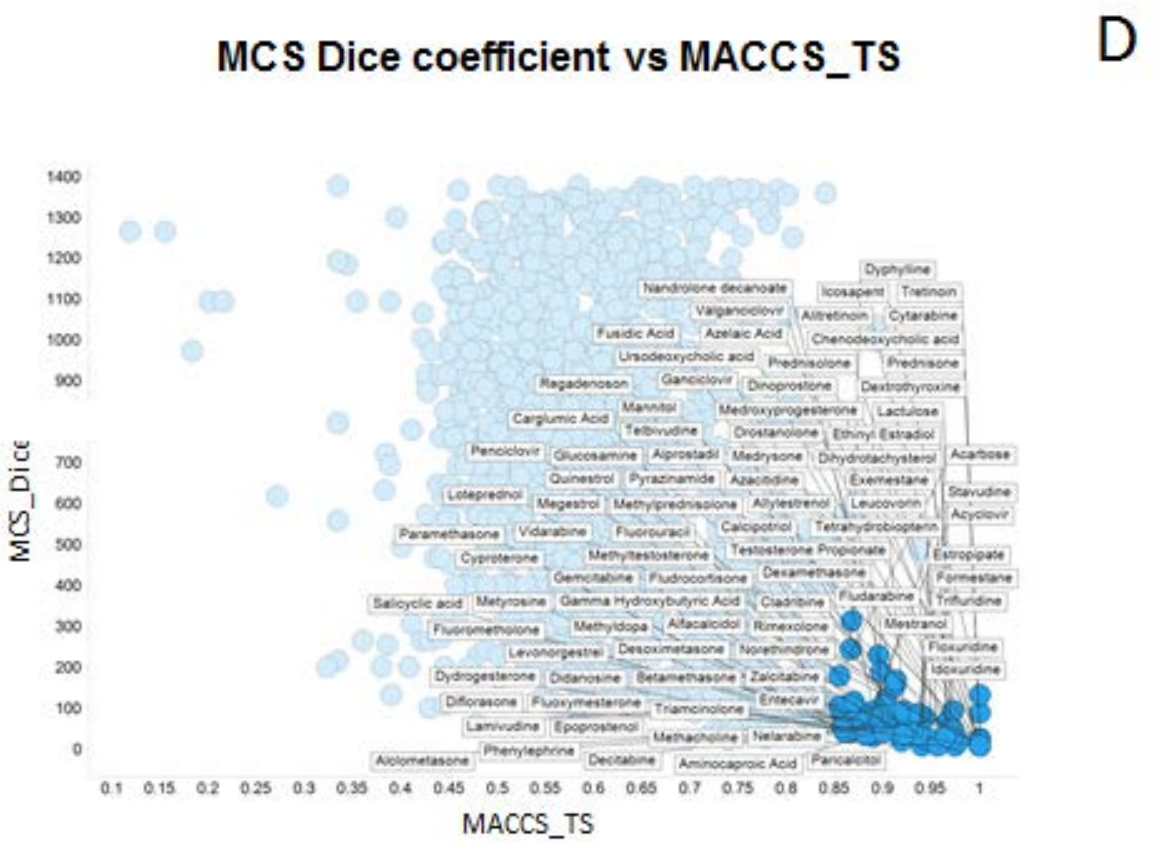
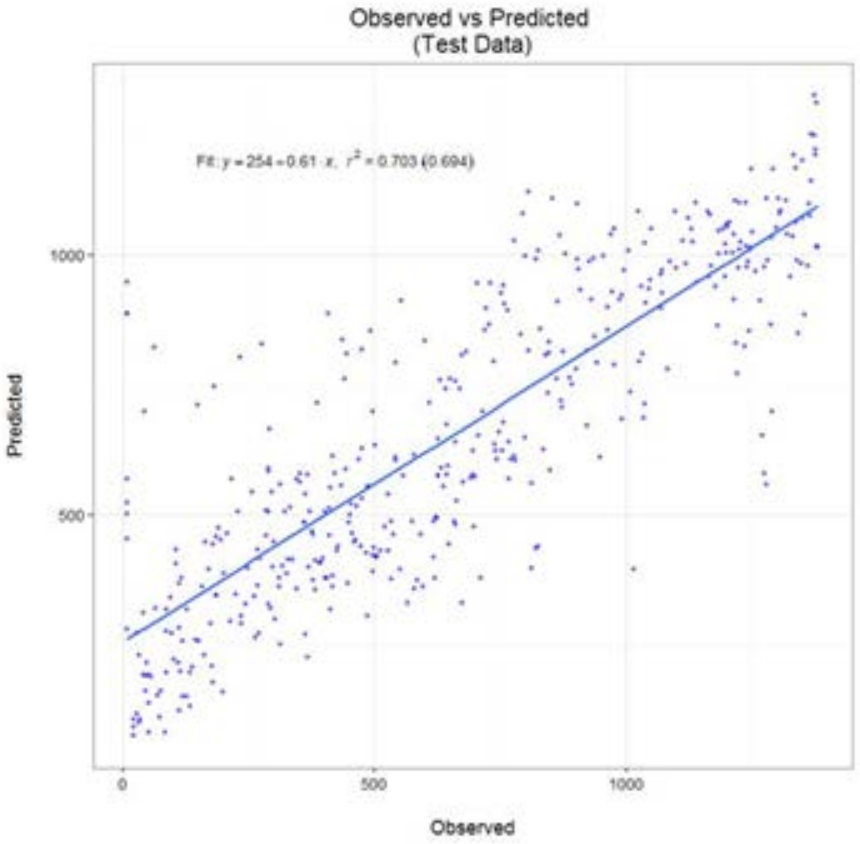


Figure 3. Rank order of drug-endogenite similarities as judged by the size of their maximum common substructures (MCS), for varying values of the Tversky similarities (α and β). A. Sum of α and β = 1. B. Sum of α and β = 2. C. distribution of values of the MCS for varying values of Sum of α and β when their sum is 2. D. MCS Dice coefficient against MACCS Tanimoto similarity.

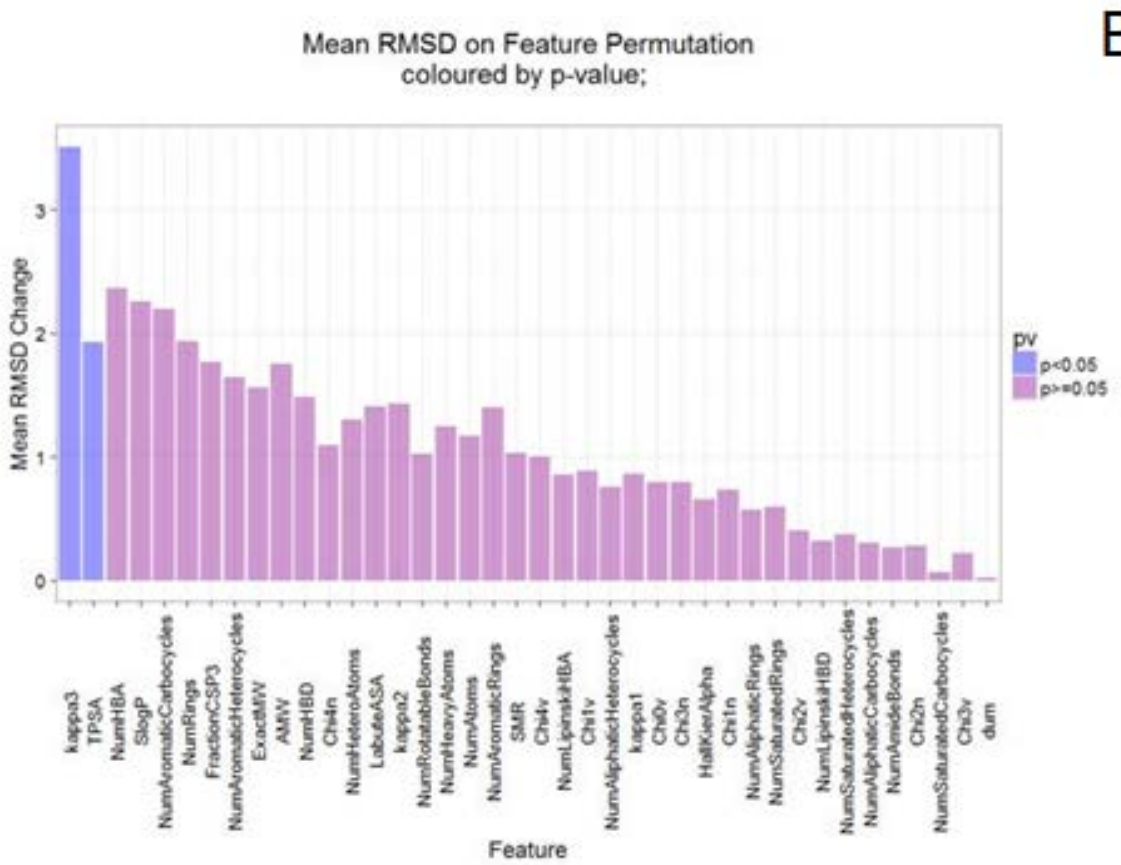
Assessment of contribution of descriptors to rank orders using random forest regression

In order to understand the structural bases for some of the rank orders, we set up a random forest regression (see e.g. [131; 132]) to assess whether we can indeed predict the rank of a particular drug molecule in terms of the Tanimoto similarity of its closest endogenous metabolite. As this is a supervised method, we trained on a subset of examples to see if we can predict an out-of-the-box set. The results are shown in Fig 4A, indicating a reasonable degree of success. To ensure that this was not due to any kind of overtraining, we performed target permutation i.e. we randomised statistically the values of the target column one thousand times (data not shown). This served to break any true correlations between the features and the targets, showing that the observed correlations were indeed real. Fig 4B shows an equivalent permutation on the features (i.e. the RDKit descriptors), to assess those which most contributed to the observed correlations. Finally, Fig 4C shows the improvement in correlation that was observed (in out-of-the-box data) as the number of features was increased; evidently ten features were sufficient to achieve the maximum correlation observed.

A



B



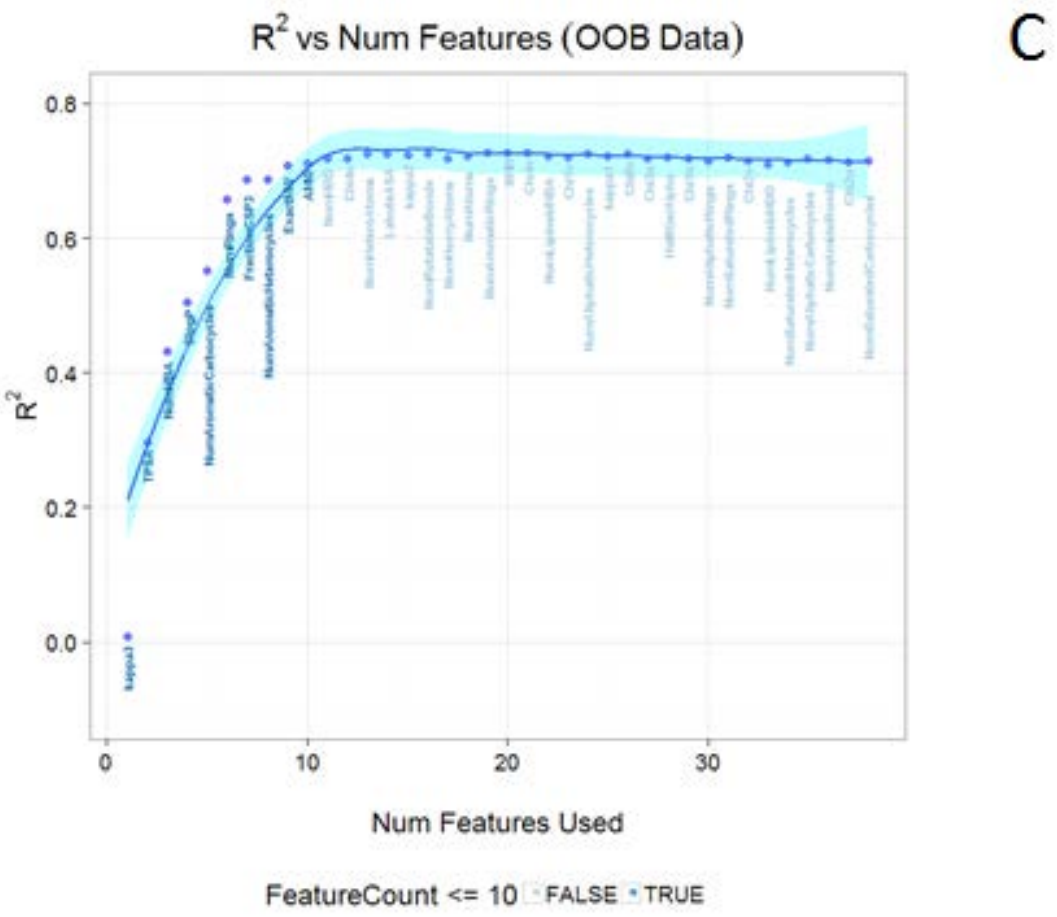


Figure 4. Random forest regression of RDKit features on the rank order prediction of the Tanimoto similarity of a given drug molecule, using the RDKit fingerprint encoding (a graph-based method that should not of itself be related to physicochemical descriptors) and the Tanimoto similarity. A. Predicted against actual. B. Ranked order of relevant features. C. Stepwise improvement in regression as features are added in the order of those seen in B.

How similar are drugs to dietary and medicinal natural products from plants and microbes?

As mentioned in the introduction, and leaving aside molecules that are actually both drugs and metabolites, some drugs are clearly much more similar to one or more endogenous metabolites than are others, and this is true for a variety of fingerprint encodings [22]. One question that we have not previously asked is about how much better our 'similarities' might be if we also used dietary or bioactive molecules that are not in Recon2. Specifically, including for evolutionary reasons rehearsed in the introduction, the question thus arose as to whether these similarities could be increased, especially for the "less similar" drugs, when we began to include bacterial, plant, and fungus-derived secondary metabolites.

We recognise that we must, so far as is reasonable, compare like with like, and certainly it can always be claimed that there is a greater likelihood of finding a molecule with a greater similarity (in a given encoding) as the size of a database is increased *per se*. The normal way of dealing with this is simply to quantify the likelihood that a given similarity could be achieved from a (more or less random) distribution of chemicals taken from that database [34; 109-113]. This is not entirely logical from a biological point of view, however, since such samples are not (from) a random

distribution but are the products of evolutionary selection. Thus we prefer other arguments, based on comparing biologically relevant databases. We do, however, also recognise that the gross distribution of properties such as MW, logP, total polar surface area (TPSA) and so on differs between the different databases, and we will need to ensure that this is not a trivial cause of any differences observed. Thus in some cases we used the MatchIt algorithm [133; 134] (and its attendant R code) to select subsets from the various databases with the same distributions of properties as those of endogenites.

The "Universal Natural Products Database" (UNPD) (http://pkuxxj.pku.edu.cn/UNPD) is said [122] to be the largest noncommercial and freely available database for natural products. At the time of its original publication [122] UNPD comprised 197,201 natural products from plants, animals and microorganisms. Our first task here involved regularising or 'cleaning' the UNPD for our purposes. Cleaning was perfomed using a KNIME workflow and lowered the number of molecules included from the ca 229,000 initially logged when we downloaded it in December 2016 to 155,048. The main 'loss' was due to the loss of (what we could not deconvolve as) duplicates. Some of these may have been stereoisomers, but the 2D connection table provided contained no stereochemistry. Figure 5 shows the distributions of four properties between the endogenous metabolites of Recon2 and the contents [122] of the 'cleaned' UNPD natural products database. Although there are clear differences, they are in fact surprisingly similar (see also [22; 122]), and as noted above, individual descriptors had only a minor influence on the random forest model.

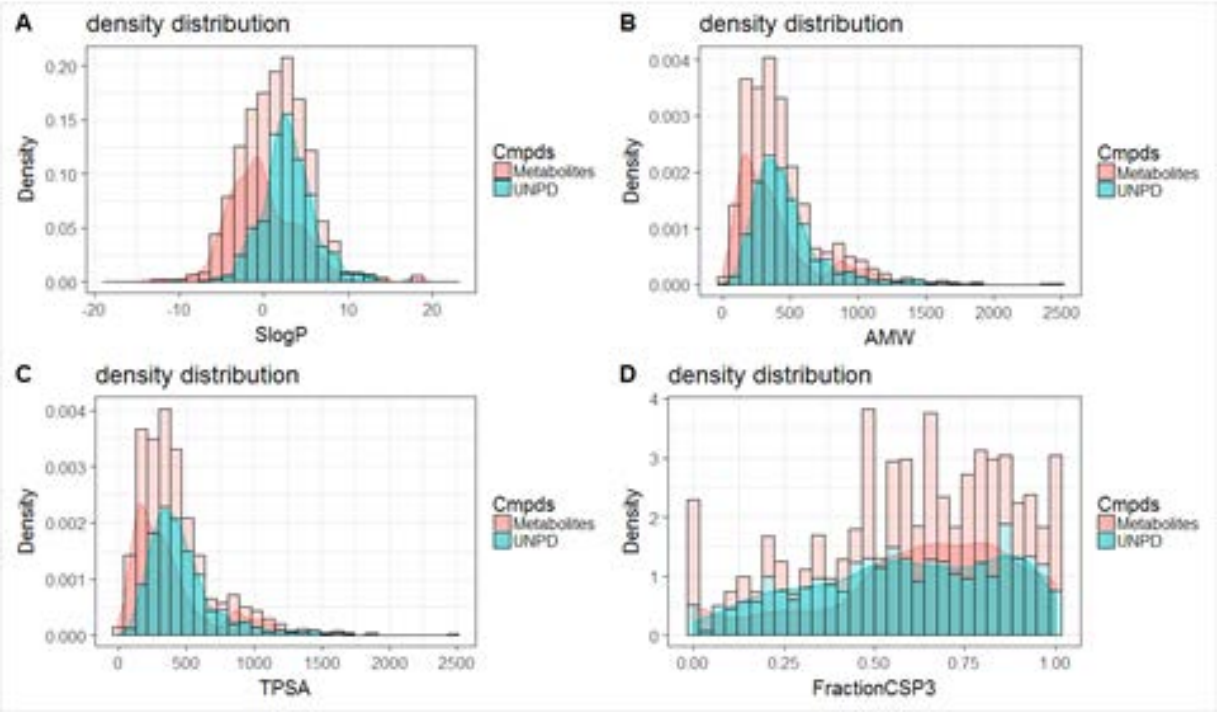
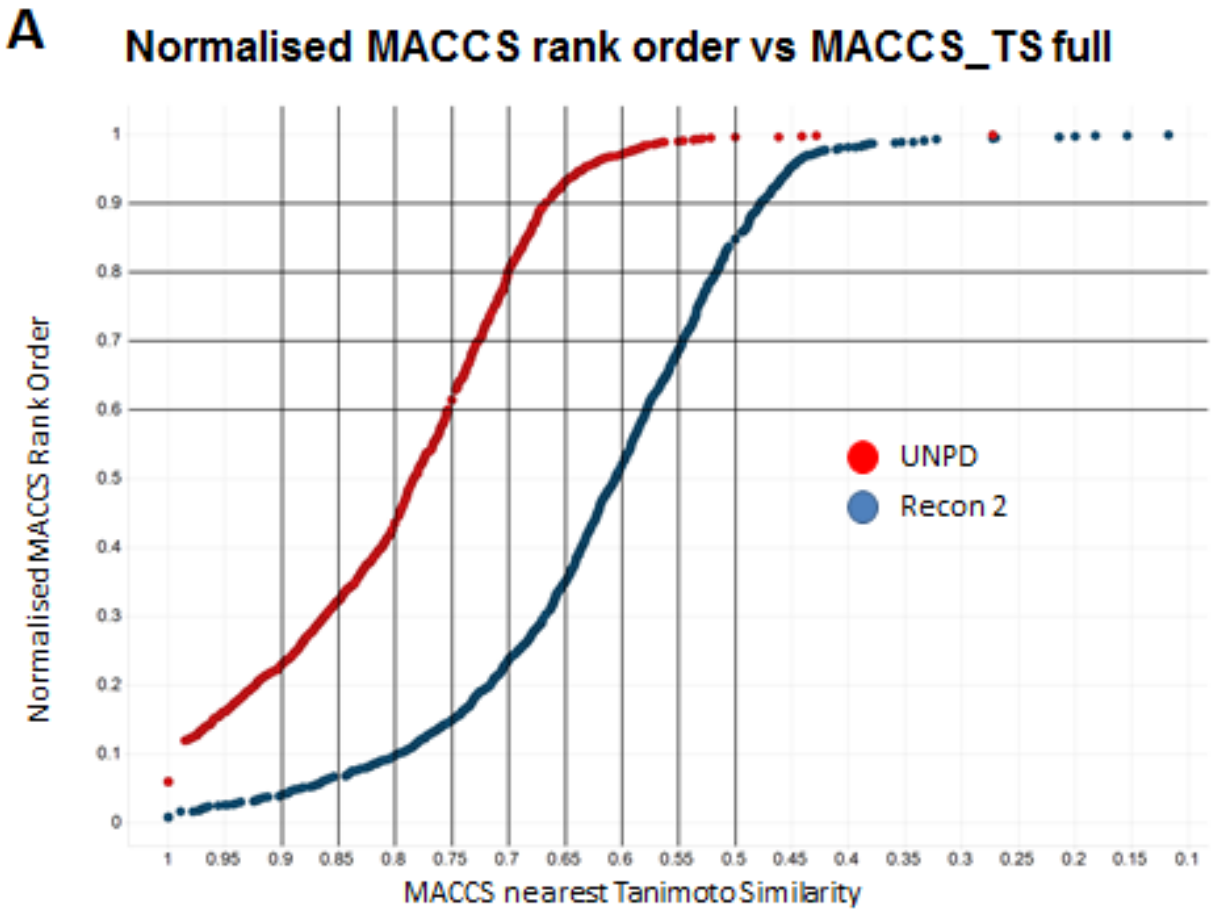
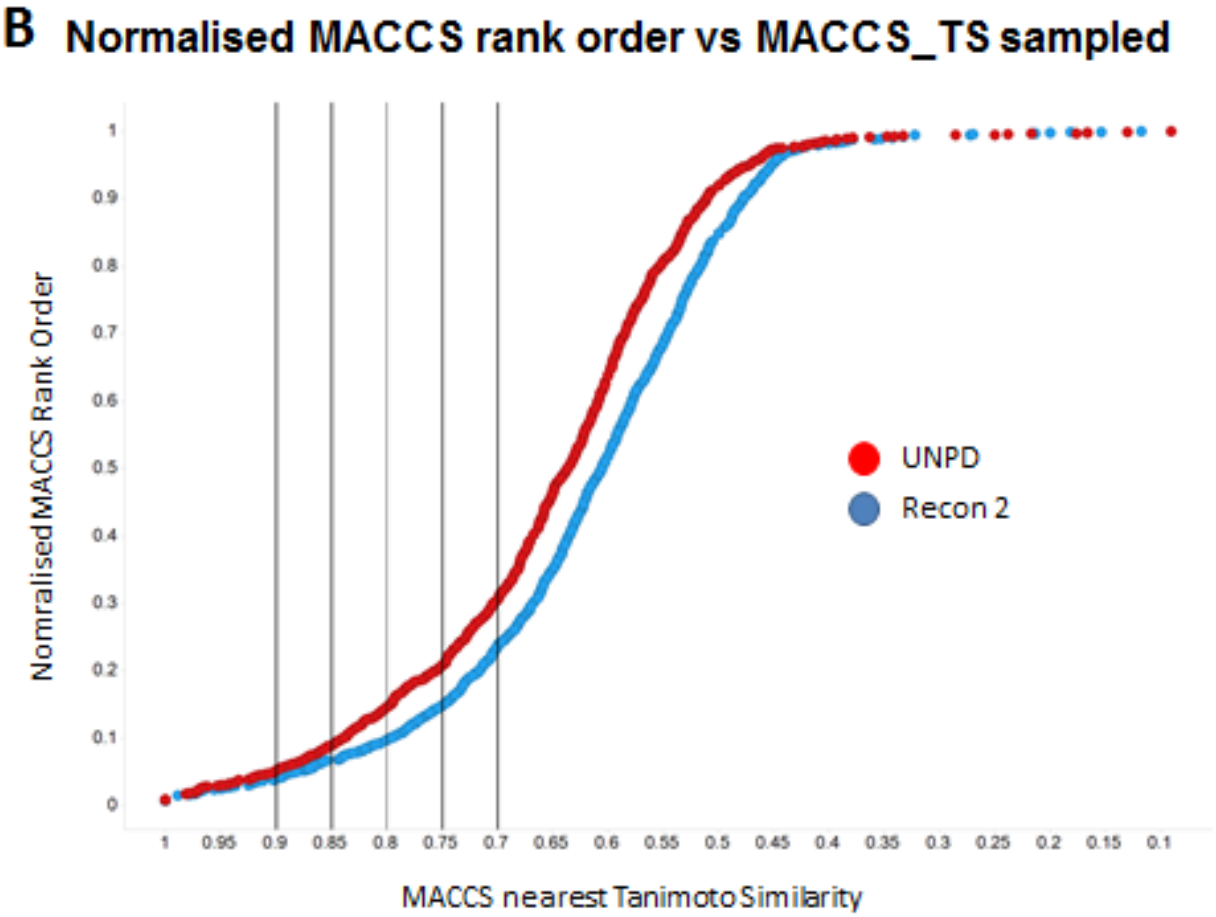


Figure 5. Distribution of four properties between the endogenous metabolites of Recon 2 [135] and the cleaned version of UNPD [122]. The original UNPD file as downloaded contained 229,358 molecules. 'Cleaning' removed duplicates as well as molecules that were in either Recon2 or in the list of marketed drugs, both of which were precisely as described and used previously [22-24]. The resulting spreadsheet retained resulted in a total of 155,048 molecules. The smoothed version is the probability density as derived from the R-encoded kernel density estimator at https://www.rdocumentation.org/packages/stats/versions/3.3.2/topics/density.

We next (Figure 6A) compared the ordered results of the Tanimoto similarity of the various marketed drugs to those of the nearest representative in our 'cleaned' version of UNPD. The results are absolutely striking; while 90% of marketed drugs had an endogenite with a TS > 0.5, the corresponding value for UNPD of 90% was a TS of 0.7. Table 1 shows the %age of drugs with a closest molecule with a TS exceeding various values (MACCS encoding) for endogenites and UNPD library members. Fairly obviously, the chance of finding a close homologue is massively greater (often four-fold or more) for the latter, especially for TS values greater than about 0.7.



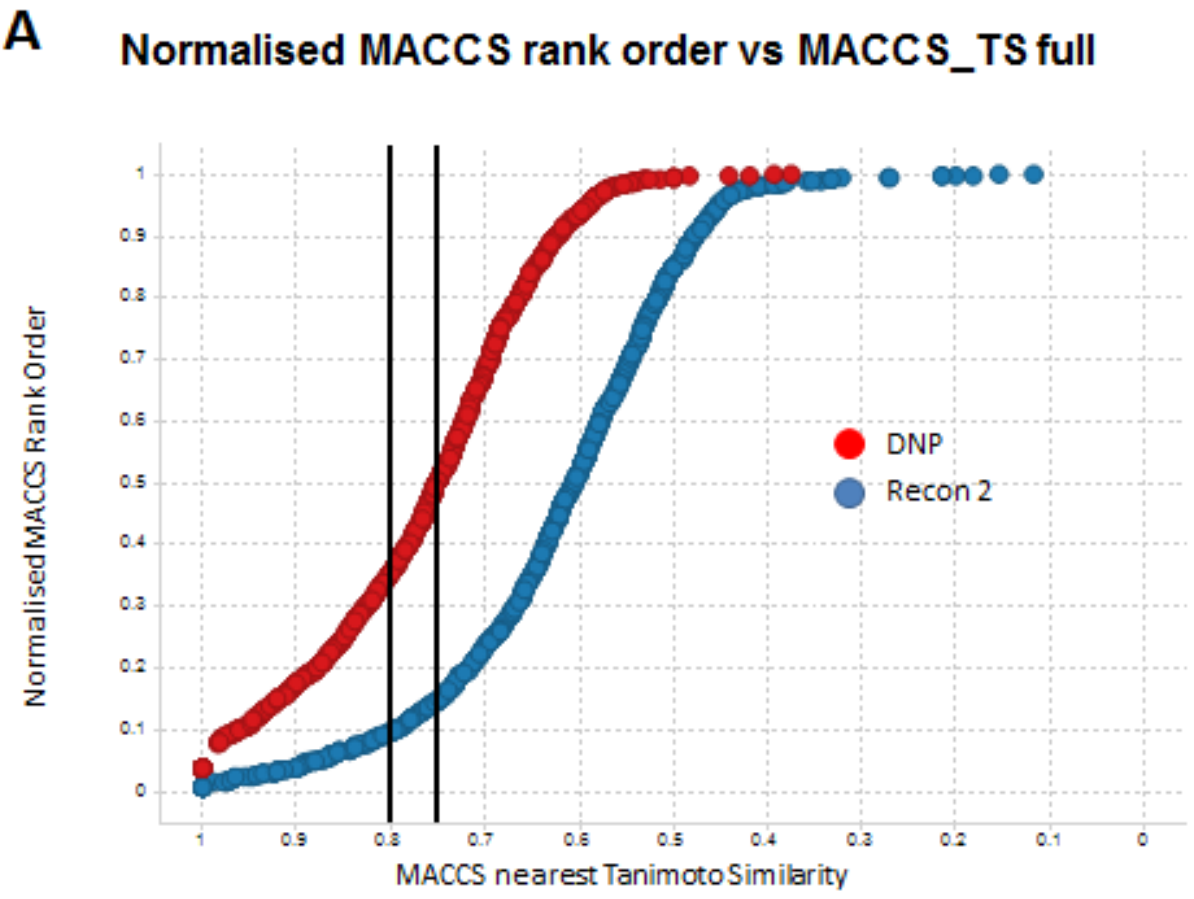


igure 6. Relationship between the normalised rank order of the nearest database molecule to marketed drugs for (red) the cleaned version of UNPD [122] and (blue) Recon2. 'Cleaning' removed duplicates as well as molecules that were in either Recon2 or in the list of marketed drugs, both of which were precisely as described and used previously [22-24]. ***A****. Full, cleaned version of UNPD.* ***B****. A sampled version of UNPD using the same number of molecules as those in Recon2 sampled as per the distributions in Fig 5.*

Table 1. *Tabulation of data from Fig 6A.*

TS > at least	Drugs/endogenites (%drugs)	Drugs/UNPD (% drugs)
0.5	1185 (85.8%)	1375 (99.6%)
0.55	941 (68.1%)	1368 (99.1%)
0.6	708 (51.3%)	1339 (97.0%)
0.65	486 (35.2%)	1289 (93.3%)
0.7	322 (23.3%)	1113 (80.6%)
0.75	201 (14.6%)	830 (60.1%)
0.8	138 (10.0%)	614 (44.5%)
0.85	93 (6.7%)	447 (32.4%)
0.9	53 (3.8%)	314 (22.7%)

One obvious point is that the number of molecules in the cleaned UNPD is roughly 100x greater than the number of those in Recon2, so it could be argued that this alone means statistically that there is simply a greater likelihood of finding a 'closer' molecule. While true, this ignores the biology (and the fact is that we did find massively more structurally close natural products than endogenites for a given drug), but we report both analyses. Thus, Fig 6B shows the same comparison as that of Fig 6A save that the UNPD molecules are sampled so as to be numerically equal to those of Recon2, and to share its distribution of the four molecular properties shown in Fig 1. In this case, the 'advantage' of UNPD is clearly diminished, albeit still substantial, with 70, 124, 197, 282 and 417 molecules with a TS > 0.9, 0.85, 0.8, 0.75 and 0.7 for UNPD, but only 57, 93, 130, 209 and 329 equivalently for Recon2. Thus for some values of TS, UNPD can enjoy a 50% advantage over Recon2 even when comparisons are strictly scaled to numbers, whatever the biology. We also ran the sampled version multiple times, to look at the 'range', but the numbers involved were great enough that this made negligible difference. Note that Recon2 does not contain the thousands of permutations of triglycerides and the like [136], that would increase its size substantially but not provide significantly better hits (i.e. the '100x' figure above is rather a substantial overestimate of the differences in true size), and we also know of many more endogenites that are not yet in Recon2 (see e.g. [9]).



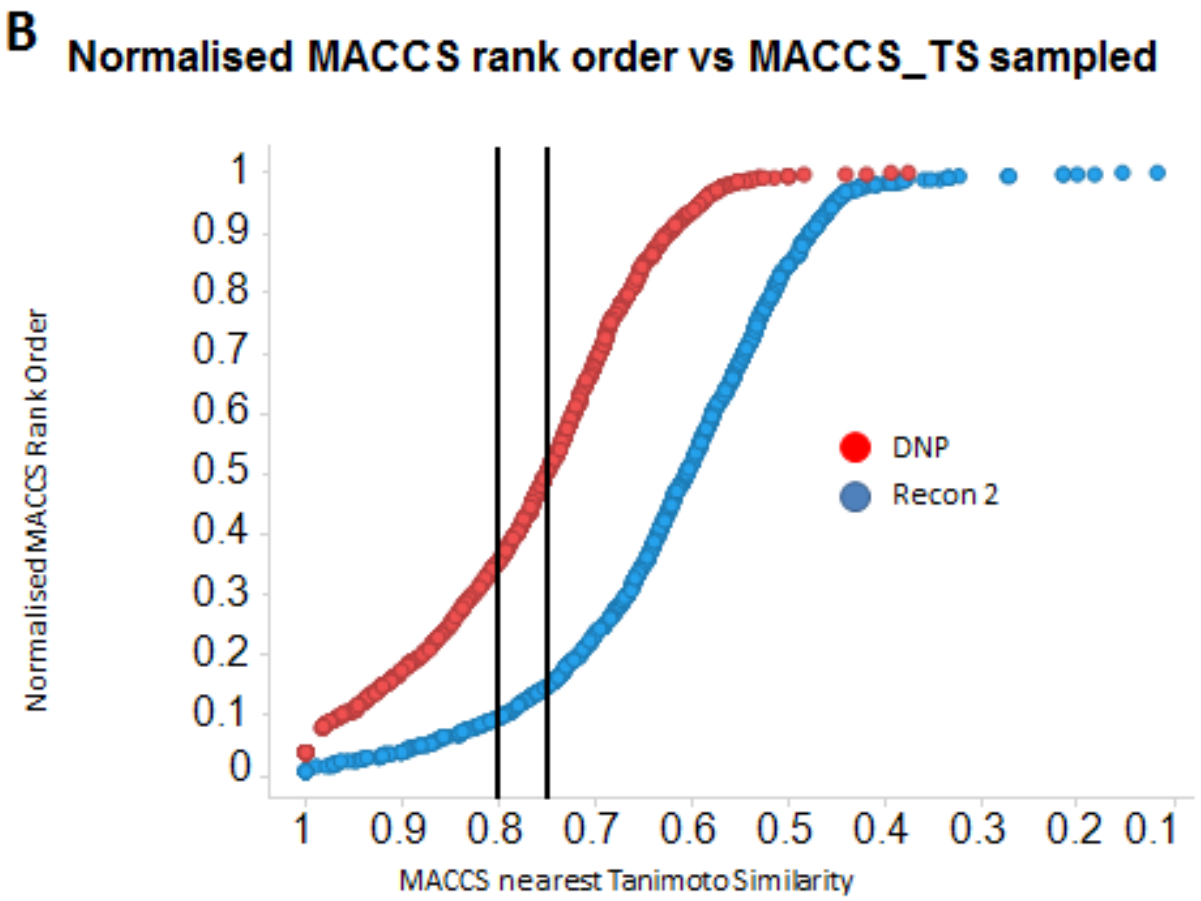


Figure 7. Relationship between the normalised rank order of the nearest database molecule to marketed drugs for (red) the cleaned version of the Dictionary of Natural Products (DNP) [125] and (blue) Recon2. 'Cleaning' removed duplicates as well as molecules that were in either Recon2 or in the list of marketed drugs, both of which were precisely as described and used previously [22-24]. ***A****. Full, cleaned version of DNP.* ***B****. A sampled version of DNP using the same number of molecules as those in Recon2 sampled as per the distribution of the properties in Fig 5.*

Fig 7A shows a similar comparison for the natural products in a cleaned-up version (see Materials and Methods) of the Dictionary of Natural Products (DNP) [125], seen as a fair comparison [137], and again using the MACCS encoding. Our cleaned DNP (with marketed drugs, endogenites and duplicates removed) contains 72,442 molecules, including 32,390 that are already in UNPD (implying 41,228 that are 'new', but also implying 123,443 that are in our UNPD but not in our DNP). Here there are at least 37% of drugs that have a TS greater than 0.8 to the nearest database member, and 50% have a TS greater than 0.75, contrasting with values for Recon2 of just 10% and 15%, respectively. These findings are roughly similar to (but the similarities slightly lower than) those from UNPD, indicating that at least some 'winners' are unique to UNPD and some to DNP. In a similar vein, Figure 7B shows the sampled version, with little impact.

Fig 8 shows the effects of cleaning and the degree of overlap of molecules in our cleaned versions of UNPD and DNP.

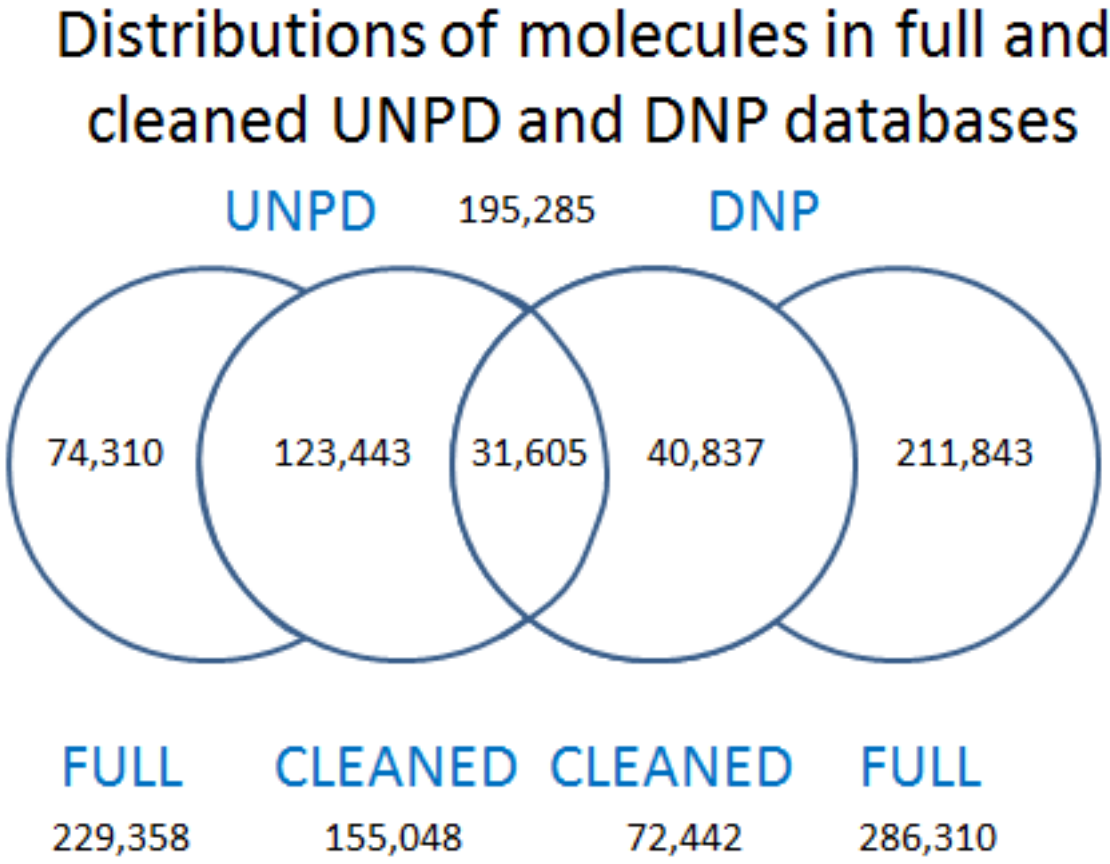


Figure 8. Overlaps between UNPD and DNP databases. 'Cleaning' removed duplicates as well as molecules that were in either Recon2 or in the list of marketed drugs, both of which were precisely as described and used previously [22-24].

The ZINC database [123] includes a very large number (ca 16M) and variety of synthetic molecules. As usual, we cleaned it to remove any molecules that were marketed drugs or Recon 2 metabolites, and ran it (and recon 2) against drugs as above. This time we ran it as 50 subsets, each of some 148,000, to show the range of curves that we could get. Thus the percentage of ZINC samples that had a member that was within a TS of 0.8 to drugs is between 35 and 45% depending on the sample, while that for a TS of 0.75 or greater varied from 0.59 to 0.74. This implies a considerably greater variation than that for the natural products.

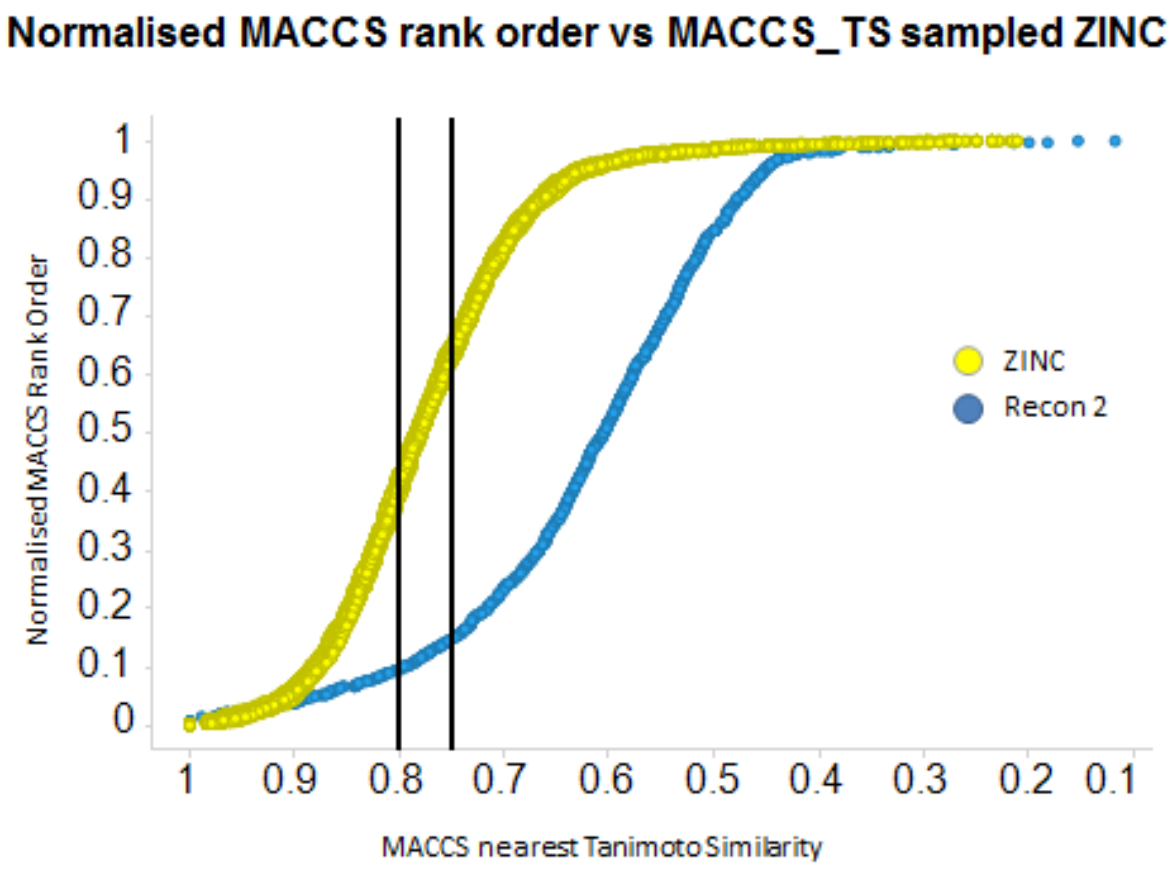
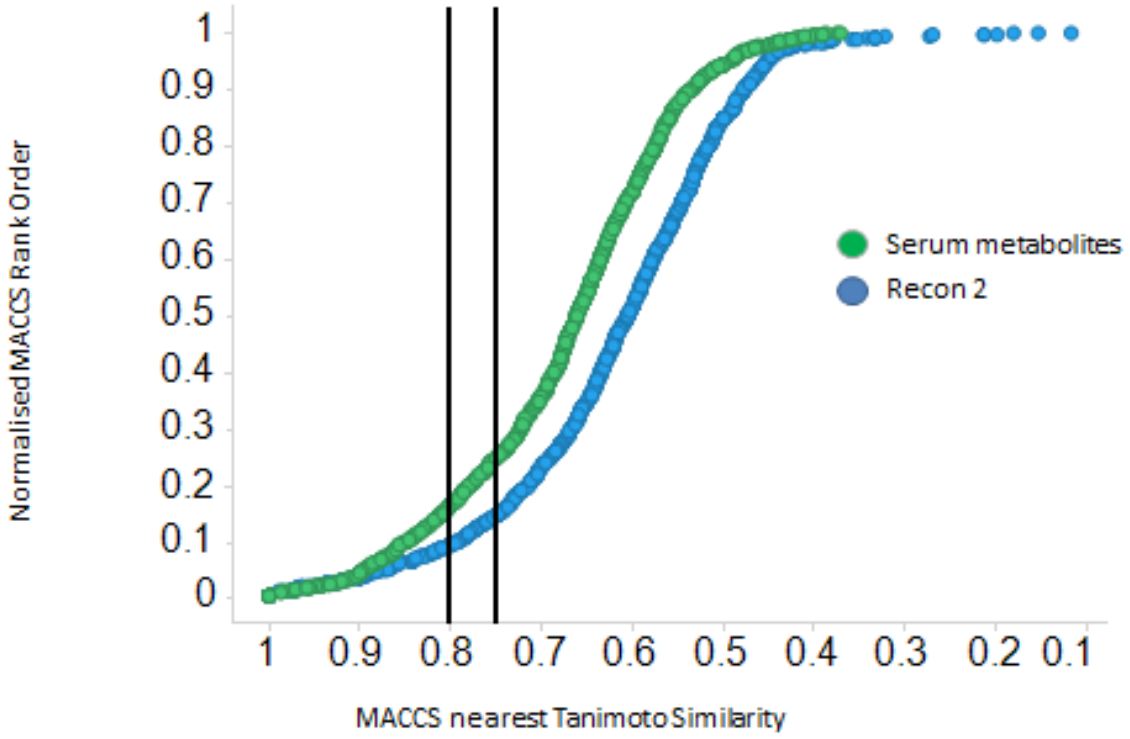


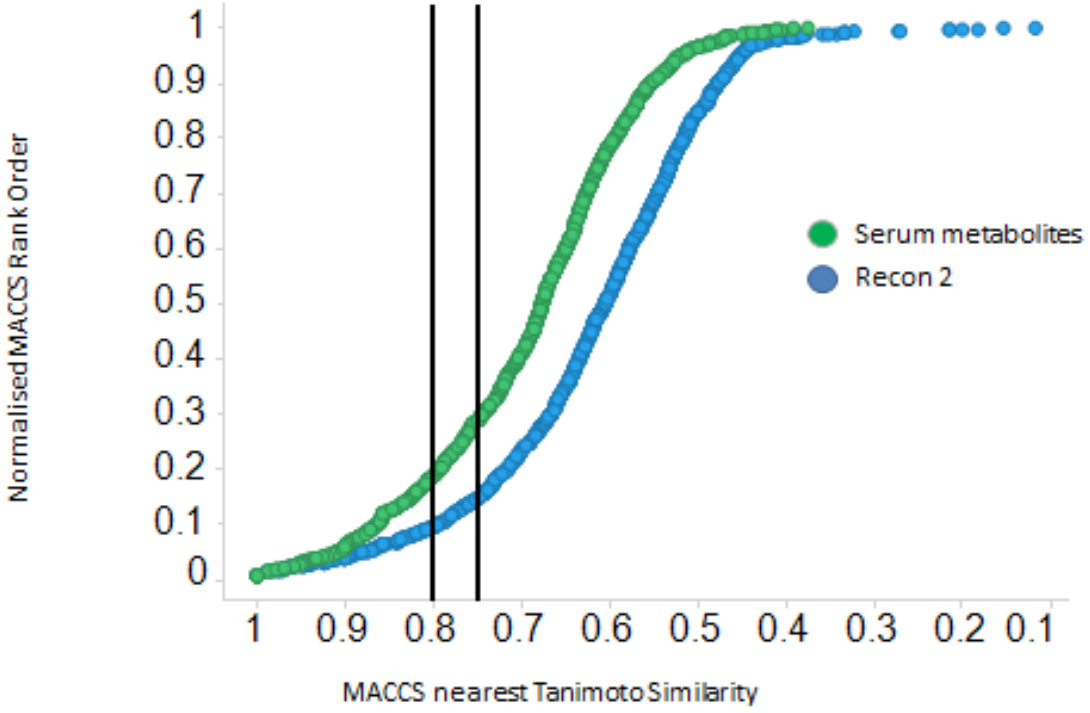
Figure 9. ZINC database. Relationship between the normalised rank order of the nearest database molecule to marketed drugs for the ZINC database (ZINC) [123] (red) and Recon 2 (blue). The ZINC database was 'cleaned' to remove molecules that were in either Recon2 or in the list of marketed drugs.

Another source of candidate transporter substrates was the list of molecules observed in serum as catalogued at http://www.serummetabolome.ca/, on the grounds that if they had reached the bloodstream they must have been transported there. We produced a version of this that again lacked all marketed drugs and recon2 metabolites, amounting to some 1480 molecules. Inspection of these indicated that they were mainly nutrients and their metabolites, along with the metabolites of various medicines. Of course what is in serum largely reflects what was recently ingested, and so it can hardly be expected to include all the natural products listed in UNPD and DNP. The curves are shown for both the subset normalized to the size of recon 2 (Fig 10A) and the full set (Fig 1B). Clearly, again, there are a significant number of 'serum' molecules that are not in Recon2 yet are structurally closer to drugs. A detailed analysis beyond this is not particularly pointful, since clearly what is in serum reflects recent ingestion only, and this is only a small subset of the contents of UNPD and DNP (Fig 10C).

A



B



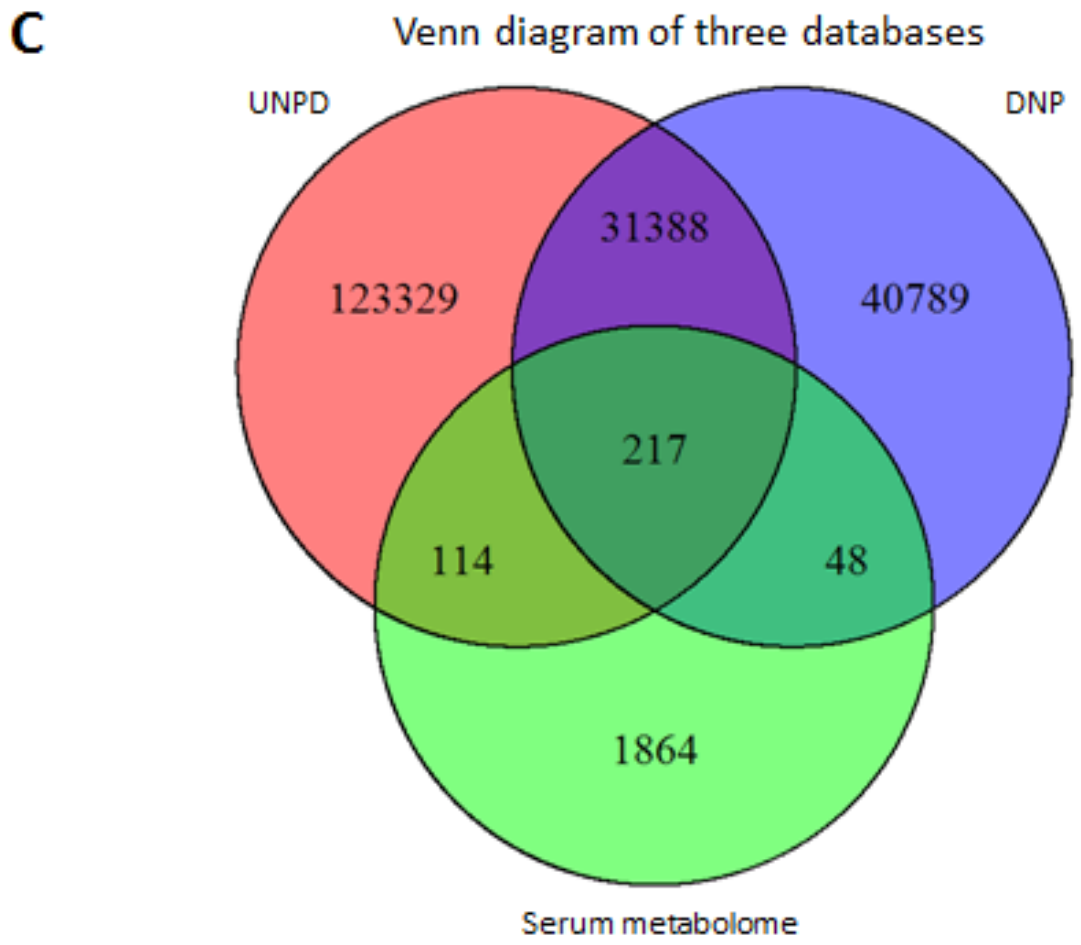
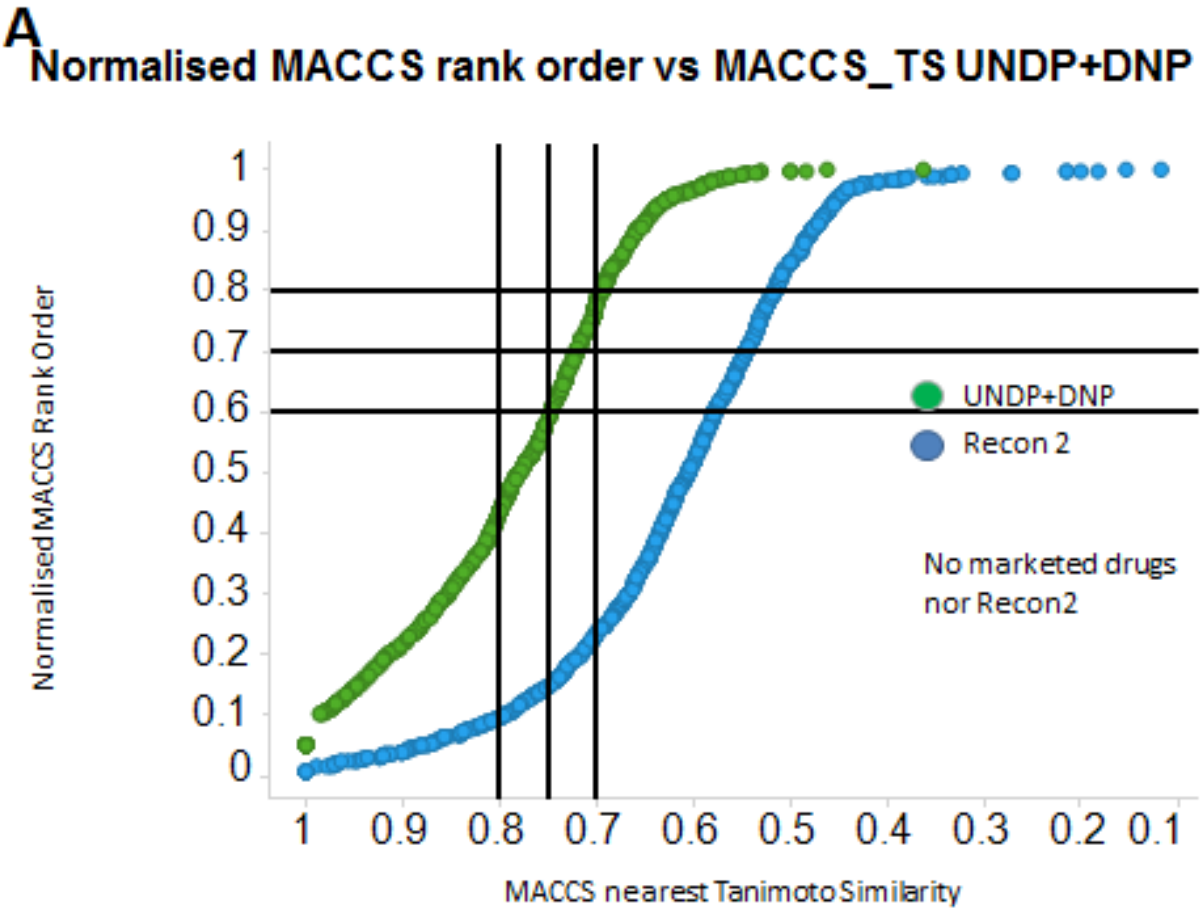


Figure 10. Human serum metabolome. Relationship between the normalised rank order of the nearest molecule to marketed drugs for the 'human serum metabolome' database **[121]** *(green) and Recon 2 (blue). The human serum metabolome database was 'cleaned' to remove molecules that were in either Recon2 or in the list of marketed drugs.* ***A.*** *Sampled subset to be the same size and property distribution as that of recon2.* ***B.*** *Full set of 'human serum metabolome' molecules after cleaning.* ***C.*** *Venn diagram of the co-distributions of molecules in the cleaned versions of UNPD, DNP and the human serum metabolome databases.*

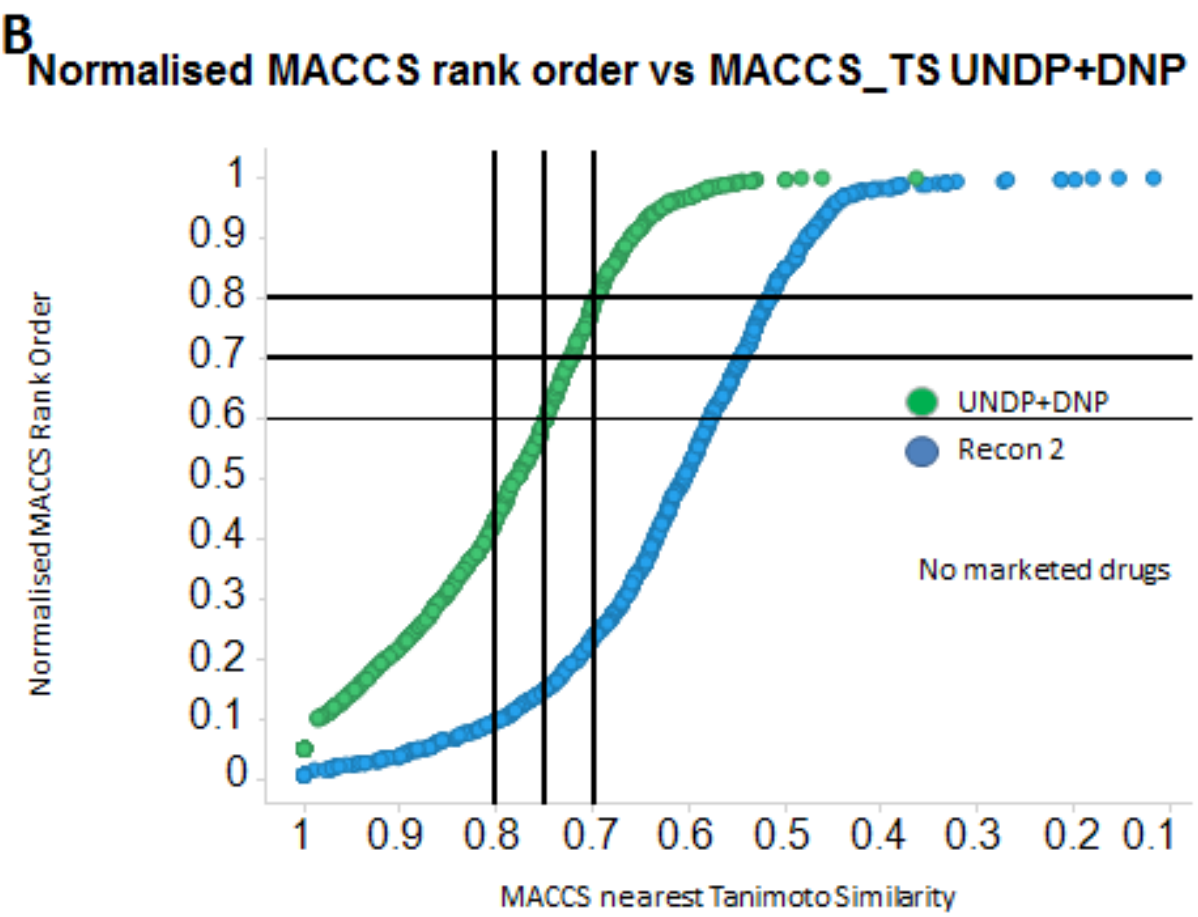
The union of the UNPD and DNP databases

Since there was (surprisingly) little overlap in the contents of the 'cleaned' versions of UNPD and DNP (Figures 8, 10C), it was of especial interest to run the analysis on their union, a set of some 195,285 molecules. The results are shown in Figure 11A for the full set for the MACCS rank order and Tanimoto similarities, and for each of the standard RDKit encodings in Fig 11B for a 148k subset. The results (Fig 11A,B) are absolutely striking: for the MACCS encoding, 45%, 60% and 80% of marketed drugs are within a TS of 0.8, 0.75 and 0.7 to at least one inhabitant of the union of the UNPD and DNP databases, regardless of the inclusion of recon2 metabolites. Fig 11C shows the data for the multiple encodings. On this basis, 80% of all drugs are within a TS of 0.8 of a natural product for the Patterned encoding, and almost all the 'missing' molecules with similarities above say 0.75 are natural products. This becomes even clearer in Fig 11D, where we chose the TYPICAL encoding, i.e. that which maximises the TS between a drug and a comparator molecule, regardless of the encoding, and performed this on the full combined natural proucts dataset of ~196,000 molecules. The result was that 92%, 98% and 99.5% are within a TS of a natural product of 0.9, 0.85 and 0.8, respectively. Fig 11E shows the rather widespread distribution of 'winning' similarities between the different encodings. Each is represented at least once, and, interestingly (as is also clear from Fig 11C), 'patterned' is the most common. This is a more recent addition to the RDKit stable, and

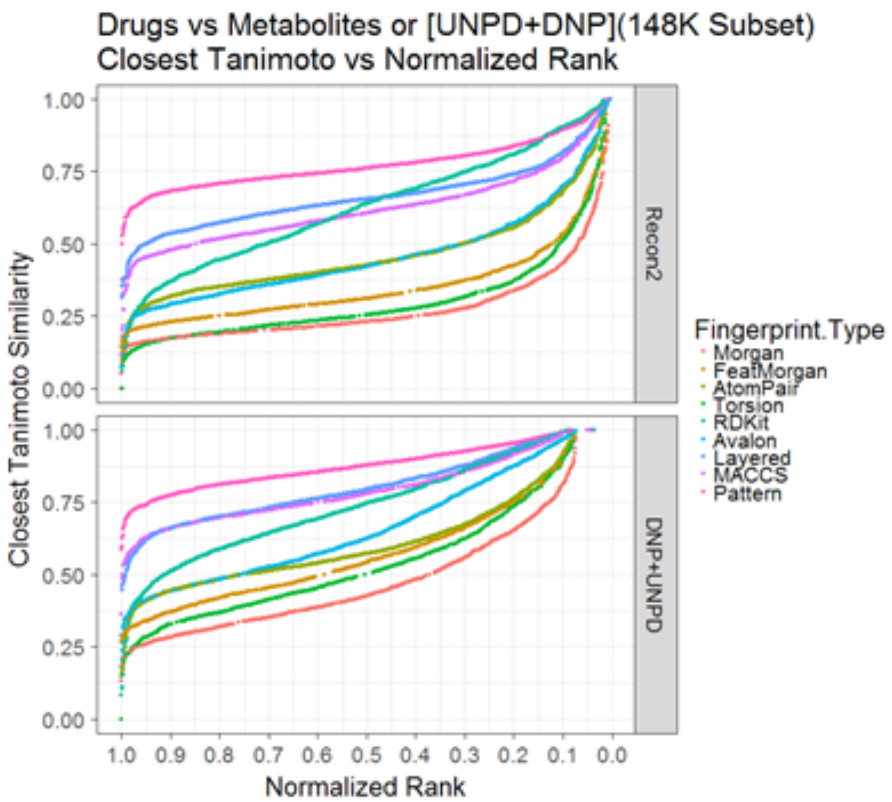
was not available when the comparison in Fig 1 was done [22]. However, the next most used are RDKit, MACCS, Layered, and Morgan; with the exception of the latter, the same may also be inferred from the endogenite-only data in Fig 1. There were exactly 500 occasions on which at least one endogenite was the closest in at least one encoding. However, when the TYPICAL encoding used, each of the 1381 drugs was closer to (or equal with) at least one exogenous natural product that is in either or both of UNPD and DNP than it is to an endogenite (data not shown). Finally, in previous work [15], we had compared a meta-analysis of 680 Caco-2 cell permeabilities of 187 marketed drugs with their endogenite-likeness (finding none). Neither did we find any relationship between Caco-2 permeability and any analysis of closeness to the union of the UNPD and DNP databases; as an illustration, Fig 11F compares the same permeabilities with the maximum pattern TS. Clearly any causal relationship that may exist is overwhelmed by the unknown variance [14] in k_{cat}, promiscuity, and transporter expression levels.



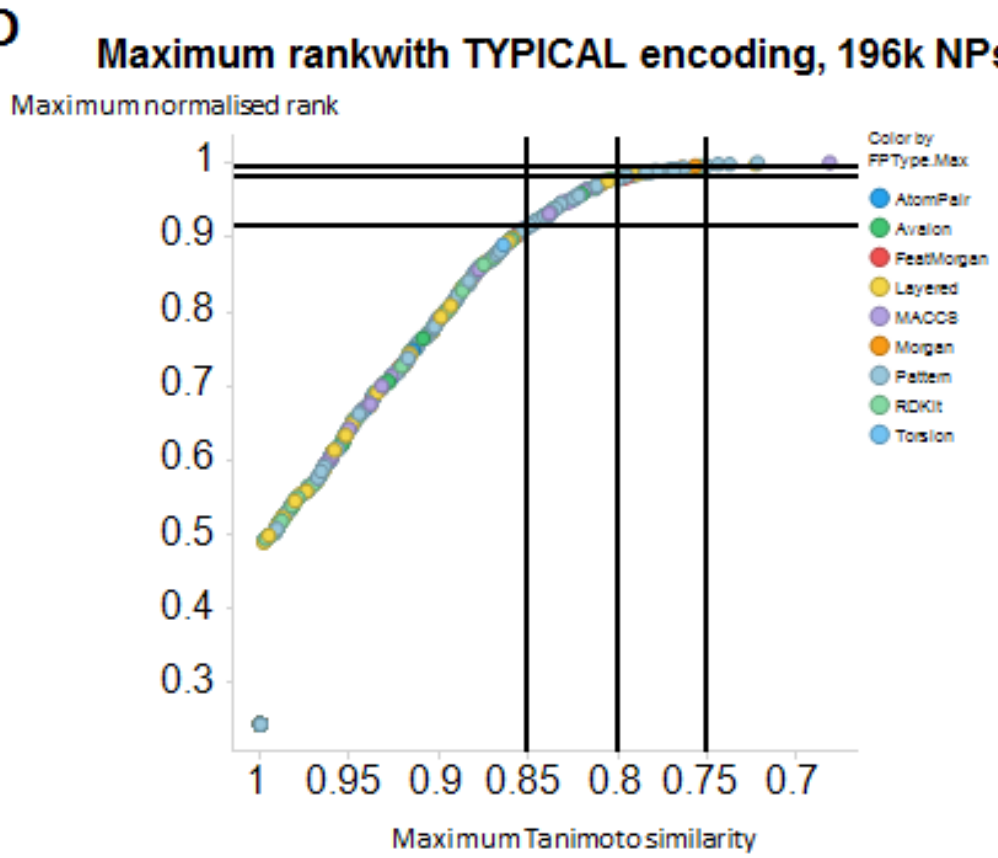
B



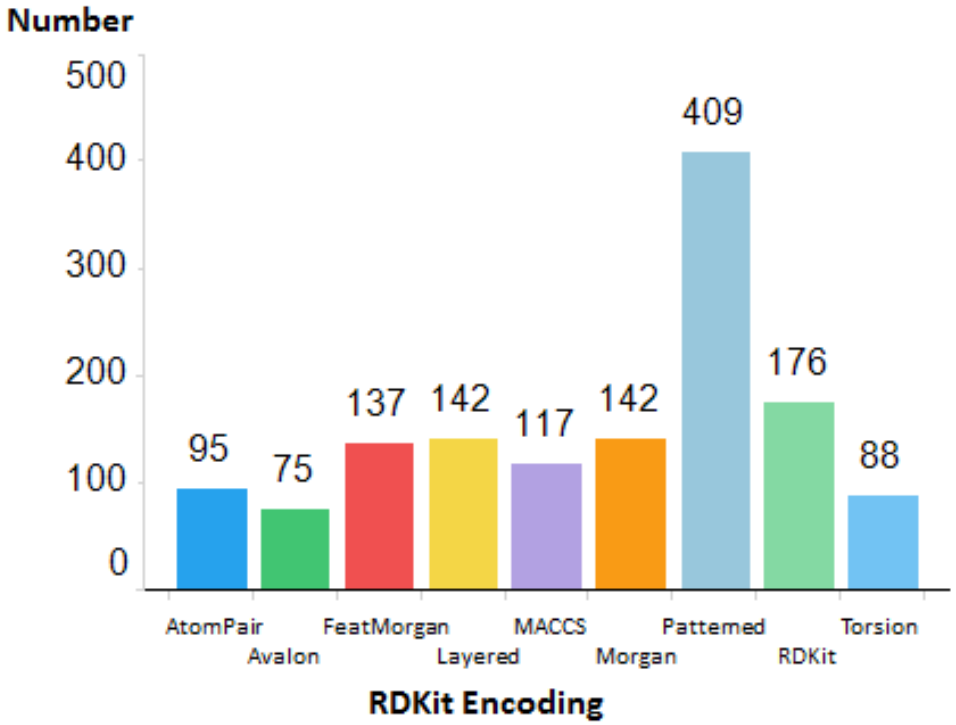
C



D



E



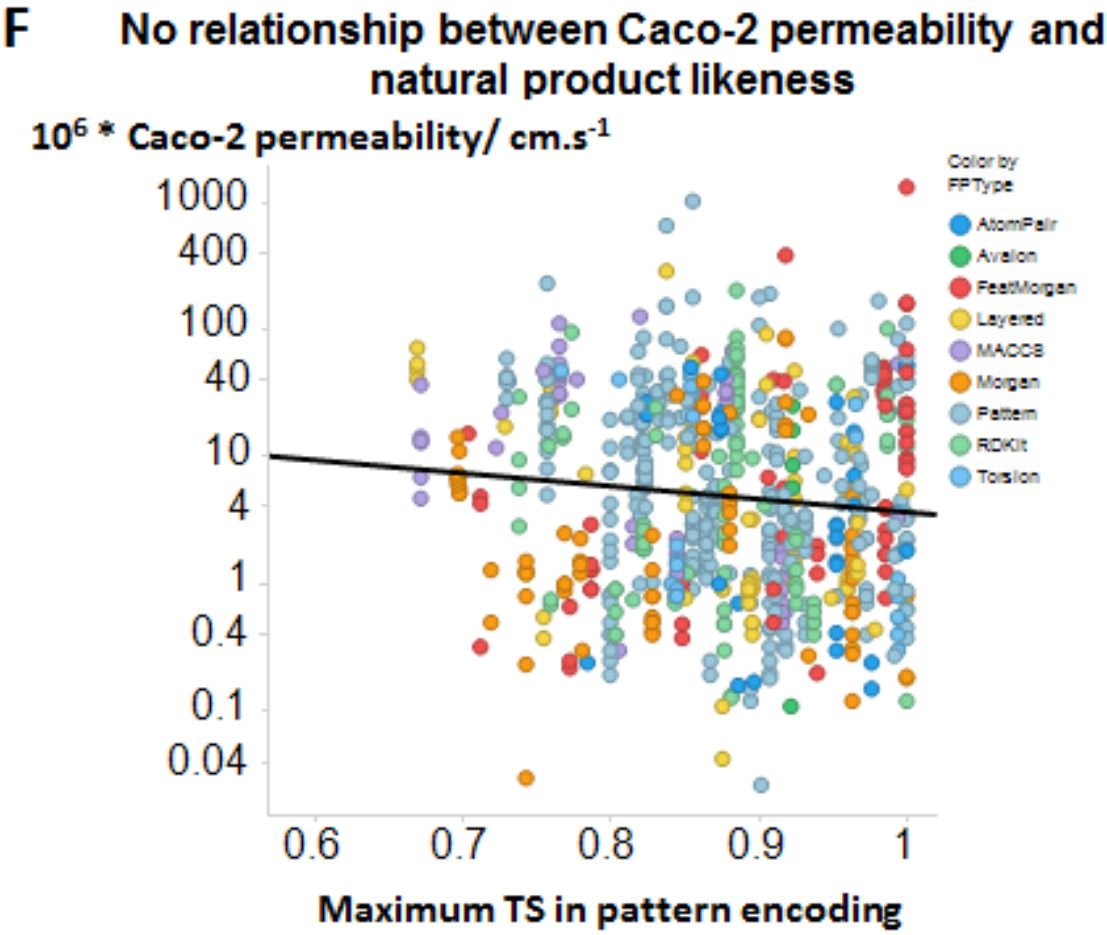


Figure 11. ***A****. Relationship between the normalised rank order of the nearest molecule to marketed drugs for the union of the UNPD and DNP natural products databases (green) and Recon 2 (blue). The UNPD and DNP databases were 'cleaned' to remove molecules that were in either Recon2 or in the list of marketed drugs.* ***B.*** *Similar data plotted for a 148k normalised subset, also lacking marketed drugs.* ***C****. The same, for each of the standard RDKit encodings.* ***D****. The same for the TYPICAL encoding.* ***E****. Distribution of encodings used in the winning molecules that contributed to the TYPICAL encoding.* ***F****. Lack of relationship between Caco-2 permeability and maximum TS of the union of the UNPD and DNP databases using the pattern encoding (slope = -1.02, r^2 = 0.011. Drugs are coloured according to the encoding with the largest TS.*

Discussion

Different similarities from different encodings

As we continue to analyse the structural 'similarities' of drugs and endogenites in different ways [14; 15; 22-24; 114], it is becoming increasingly clear that a given drug-endogenite pair can have a highly variable numerical similarity depending on which fingerprint encoding or metric of similarity is used. In the present work, we extend this recognition to the fact that – apart from a very small subset of 'reliably' endogenite-similar drugs – the degree of similarity and its rank order can be dominated by the encoding used. This was largely not the case in the analysis of Riniker and Landrum [112], who compared the similarity of fingerprints of larger and very different datasets of library compounds. We also noted that we could predict the rank order using random forest regression, so it was, as expected, a deterministic property.

Willett and colleagues have suggested that 'fusing' the results of different fingerprint encodings may give more robust analyses [138-144]. Our strategy is somewhat similar in that we recognise the highly variable rank orders (and Tanimoto similarities) that result from the different encodings, such that their variance tends to increase with their mean rank order. Summing (equivalently, averaging) the rank orders (see also [138; 144]) was a particularly convenient means of combining the data. When this was done, there was a clear trend to the effect that there was much less variance among those molecules with the most reliably high rank order (numerically small values), leading to a conclusion that for Tanimoto similarity values over ~0.75 or 0.8 the similarities are fairly robust to the specific encoding used, and on that basis may reasonably be considered 'reliable' or 'significant'. This said, there was often at least one encoding for which the TS between a given drug and at least one endogenite exceeded 0.75, such that taking the maximum of these regardless of the encoding did increase the number of 'similar' endogenites. Unfortunately, with occasional exceptions [145; 146], our knowledge of the substrate specificities of individual transporters is inadequate to the task of assessing whether the 'nearest' (or a nearby) metabolite is actually the 'natural' or endogenous substrate [12]; by and large, that will have to await further experimentation [12].

Natural products that are nutrients or bioactive drugs must necessarily be transported

As was recognized from its inception [147], the 'rule of 5' [147-153] is taken not to apply to large, natural products, and also does not apply if transporters are involved in the uptake. Natural products have been and remain a major source of successful (marketed) pharmaceutical drugs [154-161], Indeed, about one half of new drugs are based closely on natural products [156; 157], and many transporters exist for them [162-165].

It is to be assumed that anything that is of eventual benefit to (the reproductive fitness of) an organism is likely to be a subject of natural selection and adaptive evolution, even in the laboratory [166], *in vitro* [167], and *in silico* [168]. Thus, if the eating by a mammal of say a plant or fungus has beneficial properties in terms of improving the mammal's reproductive longevity, selection will act to enhance the uptake of the bioactive principles, at least to a non-toxic level. Certainly, as mentioned, it is well established that natural products themselves contribute importantly to the development of successful drugs (e.g. [154-156; 158-161; 169-172]). Consequently it is clear that a or the 'natural' substrate of at least some transporters is in fact likely to be an exogenous molecule that imparts health benefits, and ergothioneine and its uptake by SLC22A4 seem to provide a very clear example [49; 50; 58; 64; 65].

The acquisition of the ability to maintain lactase into adulthood (hence to tolerate lactose and dairy products) is highly heterogeneous and of recent evolutionary origin (~5000y BP [173-177]). Similarly, the actual human selection and prescription of plants as medications is of similar vintage [88; 90-92], albeit hominids and their evolutionary predecessors have been eating plants and fungi for many more millennia (angiosperms appear 50-100My ago). Hence, while it is not possible to replay the evolutionary tape, it is entirely reasonable that many of the several hundred human uptake transporters [68] were in fact selected, at least in part, precisely to transport exogenous secondary metabolites. Indeed, mammalian transporters are well known for their ability to transport many exogenous natural product drugs, e.g., SLCO family members for penicillins [178;

179], cephalosporins [180; 181], tetracycline [182], caffeine, theobromine and theophylline [183] and digoxin [184], SLC22 for berberine [185] and protoberberines [186], morphine [187], erythromycin [188] and theophylline [188], SLC15 for penicillins and cephalosporins [189; 190], SLC6 family (norepinephrine transporters) [191] for ephedrine derivatives [192], and SLC36 for arecaidine (an active constituent of the *Areca* nut, often wrongly referred to as the betel nut) [193].

Of necessity, there are transporters for exogenous natural products that serve as vitamins, such as ascorbate (SLC23 family [194; 195]), folate (SLC19 and SLC46 [196-198]), biotin and pantothenate (SLC5 [199]), nicotinate [200], thiamine (SLC19 and SLCO [201-204]), and riboflavin (SLC52 [205-209]).

In other cases the role of human protein transporters of natural products is well established, but their molecular nature (i.e. identity) has not yet been determined, e.g. those for psychoactive alkaloids such as cocaine [210] and nicotine [211-215] and opioids [216; 217]. Of course many transporters are known in the producer plants themselves [162-164].

Following this logic, the prediction, as tested here, is that at least some successful marketed drugs should be much closer to these plant and microbial molecules in structural terms than are the intermediary metabolites that are part of Recon2. **The prediction was amply demonstrated**, and serves to account for the otherwise anomalous finding that only a rather small fraction of intermediary human metabolites are reliably 'similar' (using the MACCS/TS metric) to marketed drugs at the level of 0.75 or 0.8. However, by contrast, we find that as many as 80% of natural products show such a similarity when surveyed extensively. This is consistent with the earlier, pilot findings (Fig 5C of [22]), and with the fact that Caco-2 permeability was poorly correlated with (the MACCS encoding of) endogenite-likeness [15].

Evolutionary aspects of 'secondary metabolites' and other natural products

The question of what roles might be played by secondary metabolites in evolutionary terms is an old one, and almost certainly does not have a unitary answer. Note that the original definition of 'secondary metabolites' was to the effect that only a small number of organisms made a given such molecule [218]. However, most of this literature on secondary metabolites, taken as virtually synonymous with 'natural products', focuses on the benefits to be gained by the producer organisms themselves. To this end, there is abundant evidence that at least some natural products are used as signals by (and towards) other individuals of the same species, and are thus pheromones [219]. Necessarily, evolutionarily early variants of natural products may lack potency at the concentrations expressed [220] and the fact that the wider the number produced, the greater the likelihood of their selection [221] can explain why the selection pressure, in terms of benefitting the producer, may often be quite modest. However, our focus here is on the benefits to consumer organisms.

It now seems clear that our earlier focus [22-24; 114] on transporters just of human-encoded intermediary metabolites as the potential source of the 'natural' substrates of the transporters on which pharmaceutical drugs hitchhike was somewhat misplaced. This is because humans, and at least their vertebrate (and indeed invertebrate) evolutionary predecessors, have

been exposed for millions of years to plant- and microbe-based dietary substances that had bioactivities of various kinds, many of which must have been beneficial and thus conferred a selective advantage, however small [222; 223], on the host. The origins of this remain uncertain, but the early ones (ca 1.8Gy BP) may have involved simple engulfment [224], with major eukaryote diversity set in place by 800My BP [225]. Even if one considers only angiosperms as potential sources of nutrients, these begin to arise ca 133 My BP [226] (with Solanaceae at around half that period BP [227]). Thus, the need (and ability) to take up plant and microbial metabolites is likely a trait of rather ancient origin.

Transporter phylogenetics

Transporter phylogenetics is an area that is still highly under-researched (as are transporters in general [12]), and indeed nearly 100 new families are introduced into the Transporter Classification Database (TCDB) every year [228]. This is not the place to pursue that issue in detail, so a single example will suffice: a BLAST search of the sequence for human SLC22A4, the ergothioneine transporter, reveals (data not shown) that it is widespread among modern mammals, but obvious homologues are not to be found in reptiles, fishes, or lower taxa.

Drugs and natural products

It is, of course, well known that many natural products can serve as medicines [87; 229], and that many purified substances derived therefrom are the basis of a significant fraction – probably 35-60% depending on how one counts – of marketed drugs (see above, and [157; 230-232]). We think that the present work highlights even more clearly how important natural products and their derivatives are likely to be in terms of producing novel, safe and efficacious drugs.

Conclusions

The present analysis takes forward our continuing analysis of the structural similarities between marketed drugs and naturally occurring substances in two major ways. First, by looking at rank orders of similarities between encodings, we find very major differences, such that the metabolite with the closest TS to a drug in one encoding may be very different in both nature and TS value from that when compared with another encoding. There is no encoding that seems to us to have any special intellectual privileges, and as stressed by Everitt [233] unsupervised analyses should anyway best be judged simply on their utility. On this basis, we consider it entirely legitimate to pick and choose encodings to maximise apparent similarities, as a guide to testing, for instance, which other substances are competing substrates for the transport of a particular drug.

Previously, we focussed solely on human endogenites and the contents of Recon2 when making these comparisons. However, not least because of the discovery by Gründemann and colleagues [64; 65] that SLC22A4 is in fact an ergothioneine transporter, we now recognise that we should include all kinds of plant- and microbial (and any other) natural products to which humans might have been exposed in evolution, and transport of whose bioactive principles might have been selected adaptively on the basis of their nutritional or medicinal activities and benefits. Clearly, natural evolution may be expected to have selected for the ability to transport molecules that in kind and amount were of benefit to the host. When we include such natural products, we find that the closeness of at least one of them, using one or more encodings, to marketed drugs is increased massively. This at once hereby points us at substances that might be the

'natural' substrates of a given drug transporter, suggests molecules for QSAR studies thereon, and potentially provides novel scaffolds for pharmaceutical drug discovery.

***Acknowledgements**: We thank the BBSRC for financial support (grants BB/K019783/1 and BB/M017702/1), and Professor David Wishart and colleagues for providing their serum metabolome database in a particularly convenient format.*

References

[1] Dobson, P. D. & Kell, D. B. (2008). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat Rev Drug Disc* **7,** 205-220.

[2] Dobson, P. D., Patel, Y. & Kell, D. B. (2009). "Metabolite-likeness" as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Disc Today* **14,** 31-40.

[3] Dobson, P., Lanthaler, K., Oliver, S. G. & Kell, D. B. (2009). Implications of the dominant role of cellular transporters in drug uptake. *Curr Top Med Chem* **9,** 163-184.

[4] Giacomini, K. M., Huang, S. M., Tweedie, D. J., Benet, L. Z., Brouwer, K. L., Chu, X., Dahlin, A., Evers, R., Fischer, V., Hillgren, K. M., Hoffmaster, K. A., Ishikawa, T., Keppler, D., Kim, R. B., Lee, C. A., Niemi, M., Polli, J. W., Sugiyama, Y., Swaan, P. W., Ware, J. A., Wright, S. H., Wah Yee, S., Zamek-Gliszczynski, M. J. & Zhang, L. (2010). Membrane transporters in drug development. *Nat Rev Drug Discov* **9,** 215-236.

[5] Kell, D. B., Dobson, P. D. & Oliver, S. G. (2011). Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only. *Drug Disc Today* **16,** 704-714.

[6] Kell, D. B., Dobson, P. D., Bilsland, E. & Oliver, S. G. (2013). The promiscuous binding of pharmaceutical drugs and their transporter-mediated uptake into cells: what we (need to) know and how we can do so. *Drug Disc Today* **18,** 218-239.

[7] Kell, D. B. (2013). Finding novel pharmaceuticals in the systems biology era using multiple effective drug targets, phenotypic screening, and knowledge of transporters: where drug discovery went wrong and how to fix it. *FEBS J* **280,** 5957-5980.

[8] Sugiyama, Y. & Steffansen, B. (2013). Transporters in Drug Development: Discovery, Optimization, Clinical Study and Regulation. AAPS/Springer, New York.

[9] Kell, D. B. & Goodacre, R. (2014). Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. *Drug Disc Today* **19,** 171-182.

[10] Kell, D. B. & Oliver, S. G. (2014). How drugs get into cells: tested and testable predictions to help discriminate between transporter-mediated uptake and lipoidal bilayer diffusion. *Front Pharmacol* **5,** 231.

[11] Winter, G. E., Radic, B., Mayor-Ruiz, C., Blomen, V. A., Trefzer, C., Kandasamy, R. K., Huber, K. V. M., Gridling, M., Chen, D., Klampfl, T., Kralovics, R., Kubicek, S., Fernandez-Capetillo, O., Brummelkamp, T. R. & Superti-Furga, G. (2014). The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity. *Nat Chem Biol* **10,** 768-773.

[12] César-Razquin, A., Snijder, B., Frappier-Brinton, T., Isserlin, R., Gyimesi, G., Bai, X., Reithmeier, R. A., Hepworth, D., Hediger, M. A., Edwards, A. M. & Superti-Furga, G. (2015). A call for systematic research on solute carriers. *Cell* **162,** 478-87.

[13] Kell, D. B. (2015). What would be the observable consequences if phospholipid bilayer diffusion of drugs into cells is negligible? *Trends Pharmacol Sci* **36,** 15-21.

[14] Mendes, P., Oliver, S. G. & Kell, D. B. (2015). Fitting transporter activities to cellular drug concentrations and fluxes: why the bumblebee can fly. *Trends Pharmacol Sci* **36,** 710-723.

[15] O'Hagan, S. & Kell, D. B. (2015). The apparent permeabilities of Caco-2 cells to marketed drugs: magnitude, and independence from both biophysical properties and endogenite similarities *PeerJ* **3,** e1405.

[16] Kell, D. B. (2016). Implications of endogenous roles of transporters for drug discovery: hitchhiking and metabolite-likeness. *Nat Rev Drug Disc* **15,** 143-144.

[17] Kell, D. B. (2016). How drugs pass through biological cell membranes – a paradigm shift in our understanding? *Beilstein Magazine* **2,** http://www.beilstein-institut.de/download/628/09_kell.pdf.

[18] Mooij, M. G., Nies, A. T., Knibbe, C. A. J., Schaeffeler, E., Tibboel, D., Schwab, M. & de Wildt, S. N. (2016). Development of Human Membrane Transporters: Drug Disposition and Pharmacogenetics. *Clin Pharmacokinet* **55,** 507-24.

[19] Govindarajan, R. & Sparreboom, A. (2016). Drug Transporters: Advances and Opportunities. *Clin Pharmacol Ther* **100,** 398-403.

[20] Grixti, J., Day, P. J. & Kell, D. B. (2017). Enhancing drug efficacy and therapeutic index through cheminformatics-based selection of small molecule binary weapons that improve transporter-mediated targeting: a cytotoxicity system based on gemcitabine. *Front Pharmacol***,** in press. http://review.frontiersin.org/review/248388/16/121421#/tab/History.

[21] Kell, D. B. (2015). The transporter-mediated cellular uptake of pharmaceutical drugs is based on their metabolite-likeness and not on their bulk biophysical properties: Towards a systems pharmacology *Perspect Sci* **6,** 66-83.

[22] O'Hagan, S., Swainston, N., Handl, J. & Kell, D. B. (2015). A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics* **11,** 323-339.

[23] O'Hagan, S. & Kell, D. B. (2015). Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites. *Front Pharmacol* **6,** 105.

[24] O'Hagan, S. & Kell, D. B. (2016). MetMaxStruct: a Tversky-similarity-based strategy for analysing the (sub)structural similarities of drugs and endogenous metabolites. *Front Pharmacol* **7,** 266.

[25] Karakoc, E., Sahinalp, S. C. & Cherkasov, A. (2006). Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J Chem Inf Model* **46,** 2167-82.

[26] Gupta, S. & Aires-de-Sousa, J. (2007). Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. *Mol Divers* **11,** 23-36.

[27] Khanna, V. & Ranganathan, S. (2009). Physicochemical property space distribution among human metabolites, drugs and toxins. *BMC Bioinformatics* **10,** S10.

[28] Peironcely, J. E., Reijmers, T., Coulier, L., Bender, A. & Hankemeier, T. (2011). Understanding and classifying metabolite space and metabolite-likeness. *PLoS One* **6,** e28966.

[29] Hamdalla, M. A., Mandoiu, II, Hill, D. W., Rajasekaran, S. & Grant, D. F. (2013). BioSM: Metabolomics Tool for Identifying Endogenous Mammalian Biochemical Structures in Chemical Structure Space. *J Chem Inf Model* **53,** 601-12.

[30] Nigam, S. K. (2015). What do drug transporters really do? *Nat Rev Drug Discov* **14,** 29-44.

[31] Johnson, M. A. & Maggiora, G. M. (1990). Concepts and applications of molecular similarity. Wiley, New York.

[32] Kubinyi, H. (1998). Similarity and dissimilarity: A medicinal chemist's view. *Perspect Drug Discov Des* **9-11,** 225-252.

[33] Arif, S. M., Holliday, J. D. & Willett, P. (2013). Comparison of chemical similarity measures using different numbers of query structures. *J Inf Sci* **39,** 7-14.

[34] Baldi, P. & Nasr, R. (2010). When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *J Chem Inf Model* **50,** 1205-22.

[35] Sheridan, R. P. (2007). Chemical similarity searches: when is complexity justified? *Expert Opin Drug Discov* **2,** 423-30.

[36] Todeschini, R. & Consonni, V. (2009). *Molecular descriptors for cheminformatics*. WILEY-VCH Verlag GmbH, Weinheim.
[37] Willett, P., Barnard, J. M. & Downs, G. M. (1998). Chemical similarity searching. *J Chem Inf Comp Sci* **38,** 983-996.
[38] Stumpfe, D. & Bajorath, J. (2011). Similarity searching. *Wires Comput Mol Sci* **1,** 260-282.
[39] Willett, P. (2014). The calculation of molecular structural similarity: principles and practice. *Mol Inform* **33,** 403-413.
[40] Peltason, L., Iyer, P. & Bajorath, J. (2010). Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J Chem Inf Model* **50,** 1021-33.
[41] Wassermann, A. M., Wawer, M. & Bajorath, J. (2010). Activity Landscape Representations for Structure-Activity Relationship Analysis. *J Med Chem*.
[42] Cruz-Monteagudo, M., Medina-Franco, J. L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro, M. N. D. S. & Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov Today* **19,** 1069-80.
[43] Martin, Y. C., Kofron, J. L. & Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *J Med Chem* **45,** 4350-8.
[44] Albery, W. J. & Knowles, J. R. (1976). Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry* **15,** 5631-5640.
[45] Fersht, A. (1977). *Enzyme structure and mechanism, 2nd ed.* W.H. Freeman, San Francisco.
[46] Keleti, T. (1986). *Basic enzyme kinetics*. Akadémiai Kiadó, Budapest.
[47] Cornish-Bowden, A. (1995). Fundamentals of enzyme kinetics, 2nd ed. Portland Press, London.
[48] Paul, B. D. & Snyder, S. H. (2010). The unusual amino acid L-ergothioneine is a physiologic cytoprotectant. *Cell Death Differ* **17,** 1134-40.
[49] Cheah, I. K. & Halliwell, B. (2012). Ergothioneine; antioxidant potential, physiological function and role in disease. *Biochim Biophys Acta* **1822,** 784-93.
[50] Halliwell, B., Cheah, I. K. & Drum, C. L. (2016). Ergothioneine, an adaptive antioxidant for the protection of injured tissues? A hypothesis. *Biochem Biophys Res Commun* **470,** 245-50.
[51] den Hengst, C. D. & Buttner, M. J. (2008). Redox control in actinobacteria. *Biochim Biophys Acta* **1780,** 1201-16.
[52] Akanmu, D., Cecchini, R., Aruoma, O. I. & Halliwell, B. (1991). The antioxidant action of ergothioneine. *Arch Biochem Biophys* **288,** 10-6.
[53] Aruoma, O. I., Whiteman, M., England, T. G. & Halliwell, B. (1997). Antioxidant action of ergothioneine: assessment of its ability to scavenge peroxynitrite. *Biochem Biophys Res Commun* **231,** 389-91.
[54] Asahi, T., Wu, X., Shimoda, H., Hisaka, S., Harada, E., Kanno, T., Nakamura, Y., Kato, Y. & Osawa, T. (2016). A mushroom-derived amino acid, ergothioneine, is a potential inhibitor of inflammation-related DNA halogenation. *Biosci Biotechnol Biochem* **80,** 313-7.
[55] Weigand-Heller, A. J., Kris-Etherton, P. M. & Beelman, R. B. (2012). The bioavailability of ergothioneine from mushrooms (*Agaricus bisporus*) and the acute effects on antioxidant capacity and biomarkers of inflammation. *Prev Med* **54 Suppl,** S75-8.
[56] Cheah, I. K., Tang, R. M. Y., Yew, T. S., Lim, K. H. C. & Halliwell, B. (2016). Administration of Pure Ergothioneine to Healthy Human Subjects: Uptake, Metabolism, and Effects on Biomarkers of Oxidative Damage and Inflammation. *Antioxid Redox Signal*.
[57] D'Onofrio, N., Servillo, L., Giovane, A., Casale, R., Vitiello, M., Marfella, R., Paolisso, G. & Balestrieri, M. L. (2016). Ergothioneine oxidation in the protection against high-glucose induced endothelial senescence: Involvement of SIRT1 and SIRT6. *Free Radic Biol Med* **96,** 211-22.

[58] Cheah, I. K., Ong, R. L., Gruber, J., Yew, T. S., Ng, L. F., Chen, C. B. & Halliwell, B. (2013). Knockout of a putative ergothioneine transporter in *Caenorhabditis elegans* decreases lifespan and increases susceptibility to oxidative damage. *Free Radic Res* **47,** 1036-45.

[59] Aruoma, O. I., Spencer, J. P. E. & Mahmood, N. (1999). Protection against oxidative damage and cell death by the natural antioxidant ergothioneine. *Food Chem Toxicol* **37,** 1043-53.

[60] Jang, J. H., Aruoma, O. I., Jen, L. S., Chung, H. Y. & Surh, Y. J. (2004). Ergothioneine rescues PC12 cells from beta-amyloid-induced apoptotic death. *Free Radic Biol Med* **36,** 288-99.

[61] Repine, J. E. & Elkins, N. D. (2012). Effect of ergothioneine on acute lung injury and inflammation in cytokine insufflated rats. *Prev Med* **54 Suppl,** S79-82.

[62] Sheridan, K. J., Lechner, B. E., Keeffe, G. O., Keller, M. A., Werner, E. R., Lindner, H., Jones, G. W., Haas, H. & Doyle, S. (2016). Ergothioneine Biosynthesis and Functionality in the Opportunistic Fungal Pathogen, *Aspergillus fumigatus*. *Sci Rep* **6,** 35306.

[63] Alamgir, K. M., Masuda, S., Fujitani, Y., Fukuda, F. & Tani, A. (2015). Production of ergothioneine by *Methylobacterium* species. *Front Microbiol* **6,** 1185.

[64] Gründemann, D., Harlfinger, S., Golz, S., Geerts, A., Lazar, A., Berkels, R., Jung, N., Rubbert, A. & Schömig, E. (2005). Discovery of the ergothioneine transporter. *Proc Natl Acad Sci U S A* **102,** 5256-61.

[65] Gründemann, D. (2012). The ergothioneine transporter controls and indicates ergothioneine activity--a review. *Prev Med* **54 Suppl,** S71-4.

[66] Shimizu, T., Masuo, Y., Takahashi, S., Nakamichi, N. & Kato, Y. (2015). Organic cation transporter OCTN1-mediated uptake of food-derived antioxidant ergothioneine into infiltrating macrophages during intestinal inflammation in mice. *Drug Metab Pharmacokinet* **30,** 231-9.

[67] Nakamichi, N., Shima, H., Asano, S., Ishimoto, T., Sugiura, T., Matsubara, K., Kusuhara, H., Sugiyama, Y., Sai, Y., Miyamoto, K., Tsuji, A. & Kato, Y. (2013). Involvement of carnitine/organic cation transporter OCTN1/SLC22A4 in gastrointestinal absorption of metformin. *J Pharm Sci* **102,** 3407-17.

[68] Hediger, M. A., Clemencon, B., Burrier, R. E. & Bruford, E. A. (2013). The ABCs of membrane transporters in health and disease (SLC series): Introduction. *Mol Aspects Med* **34,** 95-107.

[69] Dai, J. Y., Yang, J. L. & Li, C. (2008). Transport and metabolism of flavonoids from Chinese herbal remedy Xiaochaihu- tang across human intestinal Caco-2 cell monolayers. *Acta Pharmacol Sin* **29,** 1086-93.

[70] Maestro, A., Terdoslavich, M., Vanzo, A., Kuku, A., Tramer, F., Nicolin, V., Micali, F., Decorti, G. & Passamonti, S. (2010). Expression of bilitranslocase in the vascular endothelium and its function as a flavonoid transporter. *Cardiovasc Res* **85,** 175-183.

[71] Ziberna, L., Fornasaro, S., Čvorović, J., Tramer, F. & Passamonti, S. (2014). Bioavailability of flavonoids: the role of cell membrane transporters. In *Polyphenols in human health and disease* (ed. R. R. Watson, V. R. Preedy and S. Zibadi), pp. 489-511. Elsevier, Amsterdam.

[72] Lies, B., Martens, S., Schmidt, S., Boll, M. & Wenzel, U. (2012). Flavone potently stimulates an apical transporter for flavonoids in human intestinal Caco-2 cells. *Mol Nutr Food Res* **56,** 1627-35.

[73] Braidot, E., Petrussa, E., Bertolini, A., Peresson, C., Ermacora, P., Loi, N., Terdoslavich, M., Passamonti, S., Macri, F. & Vianello, A. (2008). Evidence for a putative flavonoid translocator similar to mammalian bilitranslocase in grape berries (Vitis vinifera L.) during ripening. *Planta* **228,** 203-213.

[74] Karawajczyk, A., Drgan, V., Medic, N., Oboh, G., Passamonti, S. & Novič, M. (2007). Properties of flavonoids influencing the binding to bilitranslocase investigated by neural network modelling. *Biochem Pharmacol* **73,** 308-20.

[75] Passamonti, S., Vrhovsek, U. & Mattivi, F. (2002). The interaction of anthocyanins with bilitranslocase. *Biochem Biophys Res Commun* **296,** 631-6.

[76] Passamonti, S., Vanzo, A., Vrhovsek, U., Terdoslavich, M., Cocolo, A., Decorti, G. & Mattivi, F. (2005). Hepatic uptake of grape anthocyanins and the role of bilitranslocase. *Food Res Internat* **38,** 953-960.

[77] Vanzo, A., Terdoslavich, M., Brandoni, A., Torres, A. M., Vrhovsek, U. & Passamonti, S. (2008). Uptake of grape anthocyanins into the rat kidney and the involvement of bilitranslocase. *Mol Nutr Food Res* **52,** 1106-16.

[78] Passamonti, S., Terdoslavich, M., Franca, R., Vanzo, A., Tramer, F., Braidot, E., Petrussa, E. & Vianello, A. (2009). Bioavailability of flavonoids: a review of their membrane transport and the function of bilitranslocase in animal and plant organisms. *Curr Drug Metab* **10,** 369-94.

[79] Jackson, F. (1996). The coevolutionary relationship of humans and domesticated plants. *Yearbook of Physical Anthropology* **39,** 161-176.

[80] Johns, T. (1999). The chemical ecology of human ingestive behaviors. *Ann Rev Anthropol* **28,** 27-50.

[81] Sullivan, R. J., Hagen, E. H. & Hammerstein, P. (2008). Revealing the paradox of drug reward in human evolution. *Proceedings of the Royal Society B-Biological Sciences* **275,** 1231-1241.

[82] Arjamaa, O. & Vuorisalo, T. (2010). Gene-Culture Coevolution and Human Diet. *American Scientist* **98,** 140-147.

[83] Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R. & Gordon, J. I. (2008). Evolution of mammals and their gut microbes. *Science* **320,** 1647-51.

[84] Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., Henrissat, B., Knight, R. & Gordon, J. I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332,** 970-4.

[85] Richerson, P. J., Boyd, R. & Henrich, J. (2010). Gene-culture coevolution in the age of genomics. *Proc Natl Acad Sci U S A* **107 Suppl 2,** 8985-92.

[86] de Pasquale, A. (1984). Pharmacognosy: the oldest modern science. *J Ethnopharmacol* **11,** 1-16.

[87] Gurib-Fakim, A. (2006). Medicinal plants: traditions of yesterday and drugs of tomorrow. *Mol Aspects Med* **27,** 1-93.

[88] Narayanaswamy, V. (1981). Origin and development of ayurveda: (a brief history). *Anc Sci Life* **1,** 1-7.

[89] Mohd Fauzi, F., Koutsoukas, A., Lowe, R., Joshi, K., Fan, T. P., Glen, R. C. & Bender, A. (2013). Chemogenomics approaches to rationalizing the mode-of-action of traditional Chinese and Ayurvedic medicines. *J Chem Inf Model* **53,** 661-73.

[90] Joshi, V. K., Joshi, A. & Dhiman, K. S. (2016). The Ayurvedic Pharmacopoeia of India, development and perspectives. *J Ethnopharmacol*.

[91] Jaiswal, Y., Liang, Z. & Zhao, Z. (2016). Botanical Drugs in Ayurveda and Traditional Chinese Medicine. *J Ethnopharmacol*.

[92] Yang, G. L., Gu, W., Zhang, H. Q., Zhai, X. F., Li, X. Q. & Ling, C. Q. (2016). The application status of Chinese herbal medicine in military health service in China. *Chin J Integr Med* **22,** 555-60.

[93] Killgore, J., Smidt, C., Duich, L., Romero-Chapman, N., Tinker, D., Reiser, K., Melko, M., Hyde, D. & Rucker, R. B. (1989). Nutritional importance of pyrroloquinoline quinone. *Science* **245,** 850-2.

[94] Stites, T. E., Mitchell, A. E. & Rucker, R. B. (2000). Physiological importance of quinoenzymes and the O-quinone family of cofactors. *J Nutr* **130,** 719-27.

[95] Steinberg, F., Stites, T. E., Anderson, P., Storms, D., Chan, I., Eghbali, S. & Rucker, R. (2003). Pyrroloquinoline quinone improves growth and reproductive performance in mice fed chemically defined diets. *Exp Biol Med (Maywood)* **228,** 160-6.

[96] Tao, R., Karliner, J. S., Simonis, U., Zheng, J., Zhang, J., Honbo, N. & Alano, C. C. (2007). Pyrroloquinoline quinone preserves mitochondrial function and prevents oxidative injury in adult rat cardiac myocytes. *Biochem Biophys Res Commun* **363,** 257-62.

[97] Rucker, R., Chowanadisai, W. & Nakano, M. (2009). Potential physiological importance of pyrroloquinoline quinone. *Altern Med Rev* **14,** 268-77.

[98] Misra, H. S., Rajpurohit, Y. S. & Khairnar, N. P. (2012). Pyrroloquinoline-quinone and its versatile roles in biological processes. *J Biosci* **37,** 313-25.

[99] Harris, C. B., Chowanadisai, W., Mishchuk, D. O., Satre, M. A., Slupsky, C. M. & Rucker, R. B. (2013). Dietary pyrroloquinoline quinone (PQQ) alters indicators of inflammation and mitochondrial-related metabolism in human subjects. *J Nutr Biochem* **24,** 2076-84.

[100] Akagawa, M., Nakano, M. & Ikemoto, K. (2015). Recent progress in studies on the health benefits of pyrroloquinoline quinone. *Biosci Biotechnol Biochem* **80,** 13-22.

[101] Kumar, N. & Kar, A. (2015). Pyrroloquinoline quinone (PQQ) has potential to ameliorate streptozotocin-induced diabetes mellitus and oxidative stress in mice: A histopathological and biochemical study. *Chem Biol Interact* **240,** 278-90.

[102] Qin, J., Wu, M., Yu, S., Gao, X., Zhang, J., Dong, X., Ji, J., Zhang, Y., Zhou, L., Zhang, Q. & Ding, F. (2015). Pyrroloquinoline quinone-conferred neuroprotection in rotenone models of Parkinson's disease. *Toxicol Lett* **238,** 70-82.

[103] Jonscher, K. R., Stewart, M. S., Alfonso-Garcia, A., DeFelice, B. C., Wang, X. X., Luo, Y., Levi, M., Heerwagen, M. J., Janssen, R. C., de la Houssaye, B. A., Wiitala, E., Florey, G., Jonscher, R. L., Potma, E. O., Fiehn, O. & Friedman, J. E. (2016). Early PQQ supplementation has persistent long-term protective effects on developmental programming of hepatic lipotoxicity and inflammation in obese mice. *FASEB J*.

[104] Zhang, Q., Chen, S., Yu, S., Qin, J., Zhang, J., Cheng, Q., Ke, K. & Ding, F. (2016). Neuroprotective effects of pyrroloquinoline quinone against rotenone injury in primary cultured midbrain neurons and in a rat model of Parkinson's disease. *Neuropharmacology* **108,** 238-51.

[105] Wu, J. Z., Huang, J. H., Khanabdali, R., Kalionis, B., Xia, S. J. & Cai, W. J. (2016). Pyrroloquinoline quinone enhances the resistance to oxidative stress and extends lifespan upon DAF-16 and SKN-1 activities in *C. elegans*. *Exp Gerontol* **80,** 43-50.

[106] Salisbury, S. A., Forrest, H. S., Cruse, W. B. & Kennard, O. (1979). A novel coenzyme from bacterial primary alcohol dehydrogenases. *Nature* **280,** 843-4.

[107] Duine, J. A. (1989). PQQ and quinoprotein research--the first decade. *Biofactors* **2,** 87-94.

[108] Anthony, C. (2001). Pyrroloquinoline quinone (PQQ) and quinoprotein enzymes. *Antioxid Redox Signal* **3,** 757-74.

[109] Vogt, M. & Bajorath, J. (2011). Introduction of the conditional correlated Bernoulli model of similarity value distributions and its application to the prospective prediction of fingerprint search performance. *J Chem Inf Model* **51,** 2496-506.

[110] Vogt, M. & Bajorath, J. (2011). Predicting the performance of fingerprint similarity searching. *Methods Mol Biol* **672,** 159-73.

[111] Hähnke, V., Rupp, M., Hartmann, A. K. & Schneider, G. (2013). Pharmacophore Alignment Search Tool (PhAST): Significance Assessment of Chemical Similarity. *Mol Inform* **32,** 625-646.

[112] Riniker, S. & Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J Cheminform* **5,** 26.

[113] O'Boyle, N. M. & Sayle, R. A. (2016). Comparing structural fingerprints using a literature-based similarity benchmark. *J Cheminform* **8,** 36.
[114] O'Hagan, S. & Kell, D. B. (2017). Analysis of drug-endogenous human metabolite similarities in terms of their maximum common substructures. *J Cheminform***,** in press.
[115] Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. (2015). The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov* **14,** 111-29.
[116] Füllbeck, M., Michalsky, E., Dunkel, M. & Preissner, R. (2006). Natural products: sources and databases. *Nat Prod Rep* **23,** 347-56.
[117] Johnson, S. R. & Lange, B. M. (2015). Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front Bioeng Biotechnol* **3,** 22.
[118] Tung, C. W. (2014). Public databases of plant natural products for computational drug discovery. *Curr Comput Aided Drug Des* **10,** 191-6.
[119] Yongye, A. B., Waddell, J. & Medina-Franco, J. L. (2012). Molecular scaffold analysis of natural products databases in the public domain. *Chem Biol Drug Des* **80,** 717-24.
[120] Medina-Franco, J. L. (2015). Discovery and Development of Lead Compounds from Natural Sources Using Computational Approaches. In *Evidence-Based Validation of Herbal Medicine* (ed. P. K. Mukherjee), pp. 455-475. Elsevier, Amsterdam.
[121] Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E., Huang, P., Hollander, Z., Pedersen, T. L., Smith, S. R., Bamforth, F., Greiner, R., McManus, B., Newman, J. W., Goodfriend, T. & Wishart, D. S. (2011). The human serum metabolome. *PLoS One* **6,** e16957.
[122] Gu, J. Y., Gui, Y. S., Chen, L. R., Yuan, G., Lu, H. Z. & Xu, X. J. (2013). Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PloS one* **8,** e62839.
[123] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **52,** 1757-68.
[124] Sterling, T. & Irwin, J. J. (2015). ZINC 15 - Ligand Discovery for Everyone. *J Chem Inf Model* **55,** 2324-2337.
[125] Hill, R. A. (2016). *Dictionary of natural products*. CRC Press, Boca Raton.
[126] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K. & Wiswedel, B. (2008). KNIME: the Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications* (ed. C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker), pp. 319-326. Springer, Berlin.
[127] Mazanetz, M. P., Marmon, R. J., Reisser, C. B. T. & Morao, I. (2012). Drug discovery applications for KNIME: an open source data mining platform. *Curr Top Med Chem* **12,** 1965-79.
[128] O'Hagan, S. & Kell, D. B. (2015). Software review: The KNIME workflow environment and its applications in Genetic Programming and machine learning. *Genetic Progr Evol Mach* **16,** 387-391.
[129] Landrum, G. A., Penzotti, J. E. & Putta, S. (2006). Feature-map vectors: a new class of informative descriptors for computational drug discovery. *J Comput Aided Mol Des* **20,** 751-62.
[130] Kawabata, T. (2011). Build-up algorithm for atomic correspondence between chemical structures. *J Chem Inf Model* **51,** 1775-87.
[131] Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5-32.
[132] Knight, C. G., Platt, M., Rowe, W., Wedge, D. C., Khan, F., Day, P., McShea, A., Knowles, J. & Kell, D. B. (2009). Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. *Nucleic Acids Res* **37,** e6.

[133] Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15,** 199-236.

[134] Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *J Stat Software* **42**.

[135] Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottír, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bölling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Goryanin, I., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jones, J., Jonsson, J. J., Juty, N., Keating, S., Nookaew, I., Le Novère, N., Malys, N., Mazein, A., Papin, J. A., Patel, Y., Price, N. D., Selkov Sr., E., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., Beek, H. V., Weichart, D., Nielsen, J. B., Westerhoff, H. V., Kell, D. B., Mendes, P. & Palsson, B. Ø. (2013). A community-driven global reconstruction of human metabolism. *Nat Biotechnol.* **31,** 419-425.

[136] Fahy, E., Subramaniam, S., Murphy, R. C., Nishijima, M., Raetz, C. R., Shimizu, T., Spener, F., van Meer, G., Wakelam, M. J. & Dennis, E. A. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res* **50 Suppl,** S9-14.

[137] Whittle, M., Willett, P., Klaffke, W. & van Noort, P. (2003). Evaluation of similarity measures for searching the dictionary of natural products database. *J Chem Inf Comput Sci* **43,** 449-57.

[138] Chen, B. N., Mueller, C. & Willett, P. (2010). Combination Rules for Group Fusion in Similarity-Based Virtual Screening. *Mol Inform* **29,** 533-541.

[139] Duesbury, E., Holliday, J. & Willett, P. (2015). Maximum common substructure-based data fusion in similarity searching. *J Chem Inf Model* **55,** 222-30.

[140] Ginn, C. M. R., Willett, P. & Bradshaw, J. (2000). Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design* **20,** 1-16.

[141] Salim, N., Holliday, J. & Willett, P. (2003). Combination of fingerprint-based similarity coefficients using data fusion. *J Chem Inf Comp Sci* **43,** 435-442.

[142] Whittle, M., Gillet, V. J., Willett, P. & Loesel, J. (2006). Analysis of data fusion methods in virtual screening: theoretical model. *J Chem Inf Model* **46,** 2193-205.

[143] Willett, P. (2013). Combination of Similarity Rankings Using Data Fusion. *J Chem Inf Model* **53,** 1-10.

[144] Gillet, V. J., Holliday, J. D. & Willett, P. (2015). Chemoinformatics at the University of Sheffield 2002-2014. *Mol Inform* **34,** 598-607.

[145] Liu, H. C., Jamshidi, N., Chen, Y., Eraly, S. A., Cho, S. Y., Bhatnagar, V., Wu, W., Bush, K. T., Abagyan, R., Palsson, B. O. & Nigam, S. K. (2016). An Organic Anion Transporter 1 (OAT1)-centered Metabolic Network. *J Biol Chem* **291,** 19474-86.

[146] Samsudin, F., Parker, J. L., Sansom, M. S. P., Newstead, S. & Fowler, P. W. (2016). Accurate Prediction of Ligand Affinities for a Proton-Dependent Oligopeptide Transporter. *Cell Chem Biol* **23,** 299-309.

[147] Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **23,** 3-25.

[148] Clardy, J. & Walsh, C. (2004). Lessons from natural molecules. *Nature* **432,** 829-37.

[149] Abad-Zapatero, C. (2007). A Sorcerer's apprentice and The Rule of Five: from rule-of-thumb to commandment and beyond. *Drug Discov Today* **12,** 995-7.

[150] Singh, S. B. & Pelaez, F. (2008). Biodiversity, chemical diversity and drug discovery. *Prog Drug Res* **65,** 141, 143-74.

[151] Doak, B. C., Over, B., Giordanetto, F. & Kihlberg, J. (2014). Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chem Biol* **21,** 1115-42.
[152] Petit, J., Meurice, N., Kaiser, C. & Maggiora, G. (2012). Softening the Rule of Five-where to draw the line? *Bioorg Med Chem* **20,** 5343-5351.
[153] Leeson, P. D. (2016). Molecular inflation, attrition and the rule of five. *Adv Drug Deliv Rev* **101,** 22-33.
[154] Gozalbes, R. & Pineda-Lucena, A. (2011). Small molecule databases and chemical descriptors useful in chemoinformatics: an overview. *Comb Chem High Throughput Screen* **14,** 548-558.
[155] Holdgate, G. A. (2007). Thermodynamics of binding interactions in the rational drug design process. *Expert opinion on drug discovery* **2,** 1103-1114.
[156] Newman, D. J. & Cragg, G. M. (2012). Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. *J Nat Prod* **75,** 311-335.
[157] Newman, D. J. & Cragg, G. M. (2016). Natural Products as Sources of New Drugs from 1981 to 2014. *J Nat Prod* **79,** 629-61.
[158] Oprea, T. I., Davis, A. M., Teague, S. J. & Leeson, P. D. (2001). Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comp Sci* **41,** 1308-1315.
[159] Oprea, T. I., Allu, T. K., Fara, D. C., Rad, R. F., Ostopovici, L. & Bologa, C. G. (2007). Lead-like, drug-like or "Pub-like": how different are they? *J Comput Aided Mol Des* **21,** 113-9.
[160] van Deursen, R., Blum, L. C. & Reymond, J. L. (2011). Visualisation of the chemical space of fragments, lead-like and drug-like molecules in PubChem. *J Comput Aided Mol Des* **25,** 649-662.
[161] Wunberg, T., Hendrix, M., Hillisch, A., Lobell, M., Meier, H., Schmeck, C., Wild, H. & Hinzen, B. (2006). Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov Today* **11,** 175-80.
[162] Schulz, B. & Kolukisaoglu, H. Ü. (2006). Genomics of plant ABC transporters: the alphabet of photosynthetic life forms or just holes in membranes? *FEBS Lett* **580,** 1010-6.
[163] Shitan, N. & Yazaki, K. (2013). New insights into the transport mechanisms in plant vacuoles. *Int Rev Cell Mol Biol* **305,** 383-433.
[164] Yazaki, K. (2006). ABC transporters involved in the transport of plant secondary metabolites. *FEBS Lett* **580,** 1183-91.
[165] Mousa, J. J. & Bruner, S. D. (2016). Structural and mechanistic diversity of multidrug transporters. *Nat Prod Rep* **33,** 1255-1267.
[166] Lenski, R. E., Ofria, C., Pennock, R. T. & Adami, C. (2003). The evolutionary origin of complex features. *Nature* **423,** 139-144.
[167] Currin, A., Swainston, N., Day, P. J. & Kell, D. B. (2015). Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev* **44,** 1172-1239.
[168] Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D. & Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*.
[169] Ertl, P. & Schuffenhauer, A. (2008). Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs. *Prog Drug Res* **66,** 217, 219-35.
[170] Ertl, P., Roggo, S. & Schuffenhauer, A. (2008). Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* **48,** 68-74.
[171] Newman, D. J. & Cragg, G. M. (2016). Natural Product Scaffolds of Value in Medicinal Chemistry In *Privileged Scaffolds in Medicinal Chemistry: Design, Synthesis, Evaluation* (ed. S. Bräse), pp. 348-378. RSC, London.
[172] Schmidt, B. M., Ribnicky, D. M., Lipsky, P. E. & Raskin, I. (2007). Revisiting the ancient concept of botanical therapeutics. *Nat Chem Biol* **3,** 360-6.

[173] Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E. & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74,** 1111-20.

[174] Harris, E. E. & Meyer, D. (2006). The molecular signature of selection underlying human adaptations. *Am J Phys Anthropol* **Suppl 43,** 89-130.

[175] Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., Swallow, D. M. & Thomas, M. G. (2011). Evolution of lactase persistence: an example of human niche construction. *Philos Trans R Soc Lond B Biol Sci* **366,** 863-77.

[176] Walter, J. & Ley, R. (2011). The human gut microbiome: ecology and recent evolutionary changes. *Annu Rev Microbiol* **65,** 411-29.

[177] Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I. & Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science* **354,** 760-764.

[178] Jariyawat, S., Sekine, T., Takeda, M., Apiwattanakul, N., Kanai, Y., Sophasan, S. & Endou, H. (1999). The interaction and transport of beta-lactam antibiotics with the cloned rat renal organic anion transporter 1. *J Pharmacol Exp Ther* **290,** 672-7.

[179] VanWert, A. L., Bailey, R. M. & Sweet, D. H. (2007). Organic anion transporter 3 (Oat3/Slc22a8) knockout mice exhibit altered clearance and distribution of penicillin G. *Am J Physiol Renal Physiol* **293,** F1332-41.

[180] Khamdang, S., Takeda, M., Babu, E., Noshiro, R., Onozato, M. L., Tojo, A., Enomoto, A., Huang, X. L., Narikawa, S., Anzai, N., Piyachaturawat, P. & Endou, H. (2003). Interaction of human and rat organic anion transporter 2 with various cephalosporin antibiotics. *Eur J Pharmacol* **465,** 1-7.

[181] Ueo, H., Motohashi, H., Katsura, T. & Inui, K. (2005). Human organic anion transporter hOAT3 is a potent transporter of cephalosporin antibiotics, in comparison with hOAT1. *Biochem Pharmacol* **70,** 1104-13.

[182] Babu, E., Takeda, M., Narikawa, S., Kobayashi, Y., Yamamoto, T., Cha, S. H., Sekine, T., Sakthisekaran, D. & Endou, H. (2002). Human organic anion transporters mediate the transport of tetracycline. *Jpn J Pharmacol* **88,** 69-76.

[183] Sugawara, M., Mochizuki, T., Takekuma, Y. & Miyazaki, K. (2005). Structure-affinity relationship in the interactions of human organic anion transporter 1 with caffeine, theophylline, theobromine and their metabolites. *Biochim Biophys Acta* **1714,** 85-92.

[184] Mikkaichi, T., Suzuki, T., Onogawa, T., Tanemoto, M., Mizutamari, H., Okada, M., Chaki, T., Masuda, S., Tokui, T., Eto, N., Abe, M., Satoh, F., Unno, M., Hishinuma, T., Inui, K., Ito, S., Goto, J. & Abe, T. (2004). Isolation and characterization of a digoxin transporter and its rat homologue expressed in the kidney. *Proc Natl Acad Sci U S A* **101,** 3569-74.

[185] Nies, A. T., Herrmann, E., Brom, M. & Keppler, D. (2008). Vectorial transport of the plant alkaloid berberine by double-transfected cells expressing the human organic cation transporter 1 (OCT1, SLC22A1) and the efflux pump MDR1 P-glycoprotein (ABCB1). *Naunyn Schmiedebergs Arch Pharmacol* **376,** 449-61.

[186] Li, L., Sun, S., Weng, Y., Song, F., Zhou, S., Bai, M., Zhou, H., Zeng, S. & Jiang, H. (2016). Interaction of six protoberberine alkaloids with human organic cation transporters 1, 2 and 3. *Xenobiotica* **46,** 175-83.

[187] Tzvetkov, M. V., Pereira, J. N. D., Meineke, I., Saadatmand, A. R., Stingl, J. C. & Brockmöller, J. (2013). Morphine is a substrate of the organic cation transporter OCT1 and polymorphisms in OCT1 gene affect morphine pharmacokinetics after codeine administration. *Biochem Pharmacol* **86,** 666-678.

[188] Kobayashi, Y., Sakai, R., Ohshiro, N., Ohbayashi, M., Kohyama, N. & Yamamoto, T. (2005). Possible involvement of organic anion transporter 2 on the interaction of theophylline with erythromycin in the human liver. *Drug Metab Dispos* **33,** 619-22.

[189] Li, M., Anderson, G. D., Phillips, B. R., Kong, W., Shen, D. D. & Wang, J. (2006). Interactions of amoxicillin and cefaclor with human renal organic anion and peptide transporters. *Drug Metab Dispos* **34,** 547-55.

[190] Sala-Rabanal, M., Loo, D. D., Hirayama, B. A., Turk, E. & Wright, E. M. (2006). Molecular interactions between dipeptides, drugs and the human intestinal H+ -oligopeptide cotransporter hPEPT1. *J Physiol* **574,** 149-66.

[191] Bröer, S. & Gether, U. (2012). The solute carrier 6 family of transporters. *Br J Pharmacol* **167,** 256-78.

[192] Raffel, D. M., Chen, W., Jung, Y. W., Jang, K. S., Gu, G. & Cozzi, N. V. (2013). Radiotracers for cardiac sympathetic innervation: transport kinetics and binding affinities for the human norepinephrine transporter. *Nucl Med Biol* **40,** 331-7.

[193] Voigt, V., Laug, L., Zebisch, K., Thondorf, I., Markwardt, F. & Brandsch, M. (2013). Transport of the areca nut alkaloid arecaidine by the human proton-coupled amino acid transporter 1 (hPAT1). *J Pharm Pharmacol* **65,** 582-90.

[194] Tsukaguchi, H., Tokui, T., Mackenzie, B., Berger, U. V., Chen, X. Z., Wang, Y., Brubaker, R. F. & Hediger, M. A. (1999). A family of mammalian Na^+-dependent L-ascorbic acid transporters. *Nature* **399,** 70-5.

[195] May, J. M. (2011). The SLC23 family of ascorbate transporters: ensuring that you get and keep your daily dose of vitamin C. *Br J Pharmacol* **164,** 1793-801.

[196] Hou, Z. & Matherly, L. H. (2014). Biology of the major facilitative folate transporters SLC19A1 and SLC46A1. *Curr Top Membr* **73,** 175-204.

[197] Laftah, A. H., Latunde-Dada, G. O., Fakih, S., Hider, R. C., Simpson, R. J. & McKie, A. T. (2009). Haem and folate transport by proton-coupled folate transporter/haem carrier protein 1 (SLC46A1). *Br J Nutr* **101,** 1150-6.

[198] Matherly, L. H., Wilson, M. R. & Hou, Z. (2014). The major facilitative folate transporters solute carrier 19A1 and solute carrier 46A1: biology and role in antifolate chemotherapy of cancer. *Drug Metab Dispos* **42,** 632-49.

[199] Uchida, Y., Ito, K., Ohtsuki, S., Kubo, Y., Suzuki, T. & Terasaki, T. (2015). Major involvement of Na^+ -dependent multivitamin transporter (SLC5A6/SMVT) in uptake of biotin and pantothenic acid by human brain capillary endothelial cells. *J Neurochem* **134,** 97-112.

[200] Jeanguenin, L., Lara-Nunez, A., Rodionov, D. A., Osterman, A. L., Komarova, N. Y., Rentsch, D., Gregory, J. F., 3rd & Hanson, A. D. (2012). Comparative genomics and functional analysis of the NiaP family uncover nicotinate transporters from bacteria, plants, and mammals. *Funct Integr Genomics* **12,** 25-34.

[201] Chen, L., Shu, Y., Liang, X., Chen, E. C., Yee, S. W., Zur, A. A., Li, S., Xu, L., Keshari, K. R., Lin, M. J., Chien, H. C., Zhang, Y., Morrissey, K. M., Liu, J., Ostrem, J., Younger, N. S., Kurhanewicz, J., Shokat, K. M., Ashrafi, K. & Giacomini, K. M. (2014). OCT1 is a high-capacity thiamine transporter that regulates hepatic steatosis and is a target of metformin. *Proc Natl Acad Sci U S A* **111,** 9983-8.

[202] Ganapathy, V., Smith, S. B. & Prasad, P. D. (2004). SLC19: the folate/thiamine transporter family. *Pflugers Arch* **447,** 641-6.

[203] Zhao, R. & Goldman, I. D. (2013). Folate and thiamine transporters mediated by facilitative carriers (SLC19A1-3 and SLC46A1) and folate receptors. *Mol Aspects Med* **34,** 373-85.

[204] Ortigoza-Escobar, J. D., Molero-Luis, M., Arias, A., Oyarzabal, A., Darin, N., Serrano, M., Garcia-Cazorla, A., Tondo, M., Hernández, M., Garcia-Villoria, J., Casado, M., Gort, L., Mayr, J. A., Rodríguez-Pombo, P., Ribes, A., Artuch, R. & Pérez-Dueñas, B. (2016). Free-thiamine is

a potential biomarker of thiamine transporter-2 deficiency: a treatable cause of Leigh syndrome. *Brain* **139,** 31-8.

[205] Fujimura, M., Yamamoto, S., Murata, T., Yasujima, T., Inoue, K., Ohta, K. Y. & Yuasa, H. (2010). Functional characteristics of the human ortholog of riboflavin transporter 2 and riboflavin-responsive expression of its rat ortholog in the small intestine indicate its involvement in riboflavin absorption. *J Nutr* **140,** 1722-7.

[206] Moriyama, Y. (2011). Riboflavin transporter is finally identified. *J Biochem* **150,** 341-3.

[207] Yonezawa, A. & Inui, K. (2013). Novel riboflavin transporter family RFVT/SLC52: identification, nomenclature, functional characterization and genetic diseases of RFVT/SLC52. *Mol Aspects Med* **34,** 693-701.

[208] Sabui, S., Ghosal, A. & Said, H. M. (2014). Identification and characterization of 5'-flanking region of the human riboflavin transporter 1 gene (SLC52A1). *Gene* **553,** 49-56.

[209] Ghosal, A., Sabui, S. & Said, H. M. (2015). Identification and characterization of the minimal 5'-regulatory region of the human riboflavin transporter-3 (SLC52A3) in intestinal epithelial cells. *Am J Physiol Cell Physiol* **308,** C189-96.

[210] Chapy, H., Smirnova, M., Andre, P., Schlatter, J., Chiadmi, F., Couraud, P. O., Scherrmann, J. M., Decleves, X. & Cisternino, S. (2015). Carrier-Mediated Cocaine Transport at the Blood-Brain Barrier as a Putative Mechanism in Addiction Liability. *Int J Neuropsychopharmacol* **18**.

[211] Fukada, A., Saito, H. & Inui, K. (2002). Transport mechanisms of nicotine across the human intestinal epithelial cell line Caco-2. *J Pharmacol Exp Ther* **302,** 532-8.

[212] Tega, Y., Kubo, Y., Yuzurihara, C., Akanuma, S. & Hosoya, K. (2015). Carrier-Mediated Transport of Nicotine Across the Inner Blood-Retinal Barrier: Involvement of a Novel Organic Cation Transporter Driven by an Outward H^+ Gradient. *J Pharm Sci* **104,** 3069-75.

[213] Tega, Y., Akanuma, S., Kubo, Y. & Hosoya, K. (2015). Involvement of the H^+/organic cation antiporter in nicotine transport in rat liver. *Drug Metab Dispos* **43,** 89-92.

[214] Takano, M., Nagahiro, M. & Yumoto, R. (2016). Transport Mechanism of Nicotine in Primary Cultured Alveolar Epithelial Cells. *J Pharm Sci* **105,** 982-8.

[215] Tega, Y., Yuzurihara, C., Kubo, Y., Akanuma, S., Ehrhardt, C. & Hosoya, K. (2016). Functional expression of nicotine influx transporter in A549 human alveolar epithelial cells. *Drug Metab Pharmacokinet* **31,** 99-101.

[216] Sadiq, M. W., Bostrom, E., Keizer, R., Bjorkman, S. & Hammarlund-Udenaes, M. (2013). Oxymorphone active uptake at the blood-brain barrier and population modeling of its pharmacokinetic-pharmacodynamic relationship. *J Pharm Sci* **102,** 3320-31.

[217] Gharavi, R., Hedrich, W., Wang, H. & Hassan, H. E. (2015). Transporter-Mediated Disposition of Opioids: Implications for Clinical Drug Interactions. *Pharm Res* **32,** 2477-2502.

[218] Bu'lock, J. D. (1961). Intermediary metabolism and antibiotic synthesis. *Adv. Microbial Physiol.* **3,** 293-333.

[219] Kell, D. B., Kaprelyants, A. S. & Grafen, A. (1995). On pheromones, social behaviour and the functions of secondary metabolism in bacteria. *Trends Ecol. Evolution* **10,** 126-129.

[220] Yim, G., Wang, H. H. & Davies, J. (2006). The truth about antibiotics. *Int J Med Microbiol* **296,** 163-70.

[221] Firn, R. D. & Jones, C. G. (2003). Natural products--a simple model to explain chemical diversity. *Nat Prod Rep* **20,** 382-91.

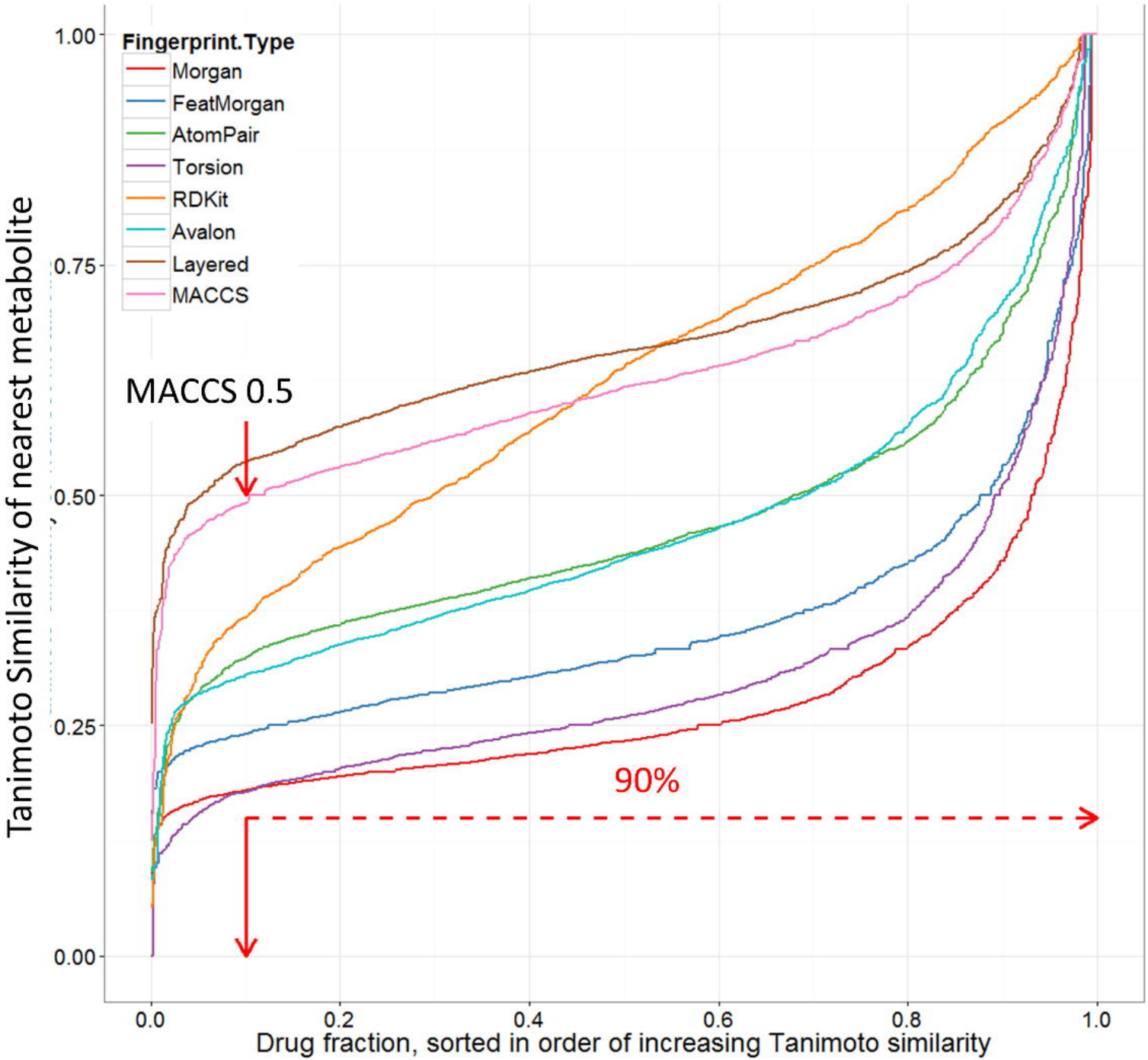
[222] Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

[223] Dawkins, R. (2006). *The selfish gene: 30th anniversary edition*. Oxford University Press, Oxford.

[224] Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. (2006). Eukaryotic organisms in Proterozoic oceans. *Philos Trans R Soc Lond B Biol Sci* **361,** 1023-38.
[225] Knoll, A. H. (2014). Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb Perspect Biol* **6**.
[226] Parfrey, L. W., Lahr, D. J., Knoll, A. H. & Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci U S A* **108,** 13624-9.
[227] Wilf, P., Carvalho, M. R., Gandolfo, M. A. & Cúneo, N. R. (2017). Eocene lantern fruits from Gondwanan Patagonia and the early origins of Solanaceae. *Science* **355,** 71-75.
[228] Saier, M. H., Jr., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C. & Moreno-Hagelsieb, G. (2016). The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res* **44,** D372-9.
[229] Lahlou, M. (2013). The Success of Natural Products in Drug Discovery. *Pharmacol Pharm* **4,** 17-31.
[230] Li, J. W.-H. & Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science* **325,** 161-5.
[231] Molinari, G. (2009). Natural products in drug discovery: present status and perspectives. *Adv Exp Med Biol* **655,** 13-27.
[232] van Herwerden, E. F. & Süssmuth, R. D. (2016). Sources for Leads: Natural Products and Libraries. *Handb Exp Pharmacol* **232,** 91-123.
[233] Everitt, B. S. (1993). *Cluster Analysis*. Edward Arnold, London.

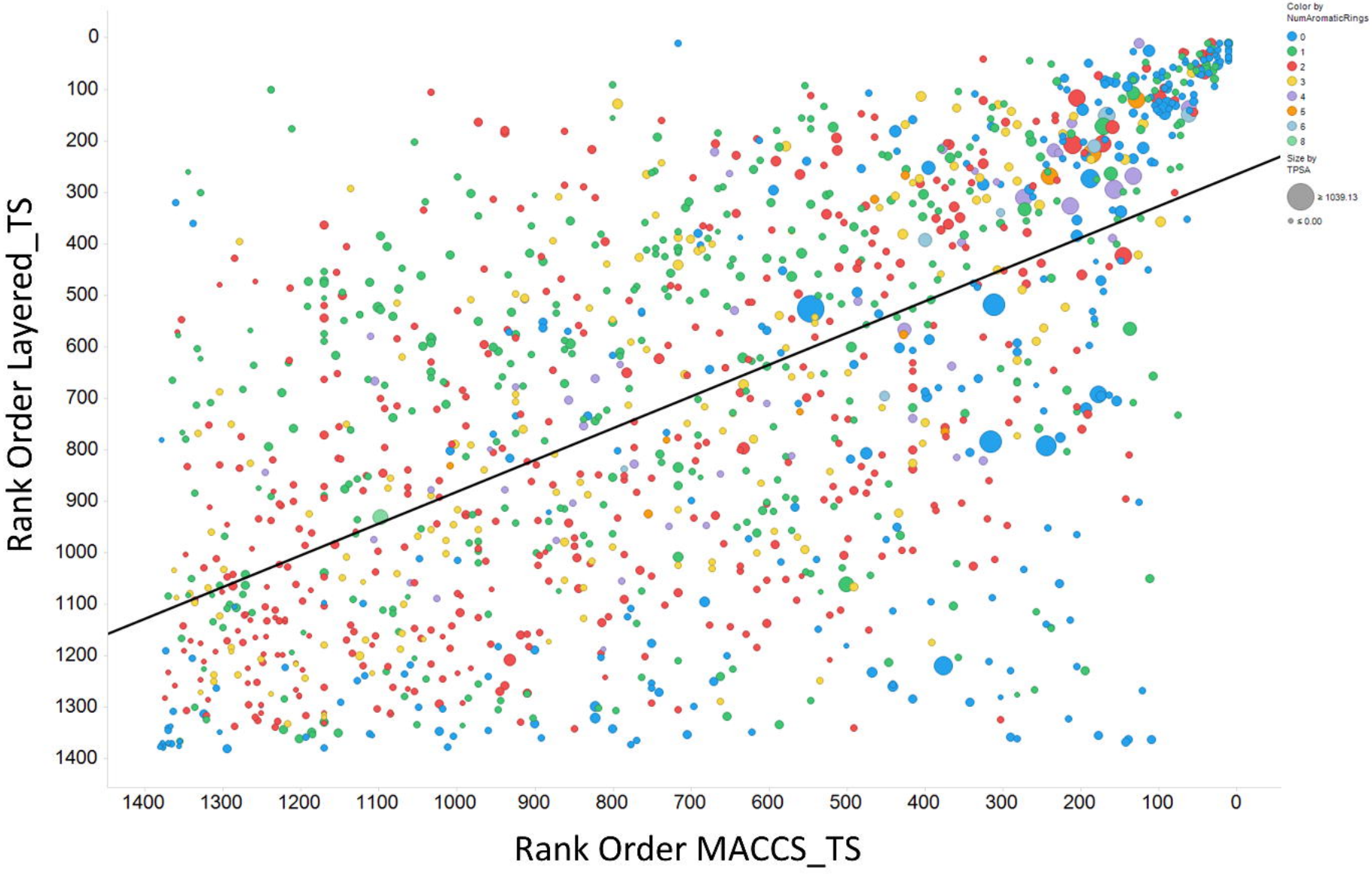
Cumulative Closest Tanimoto distance for different fingerprints

A



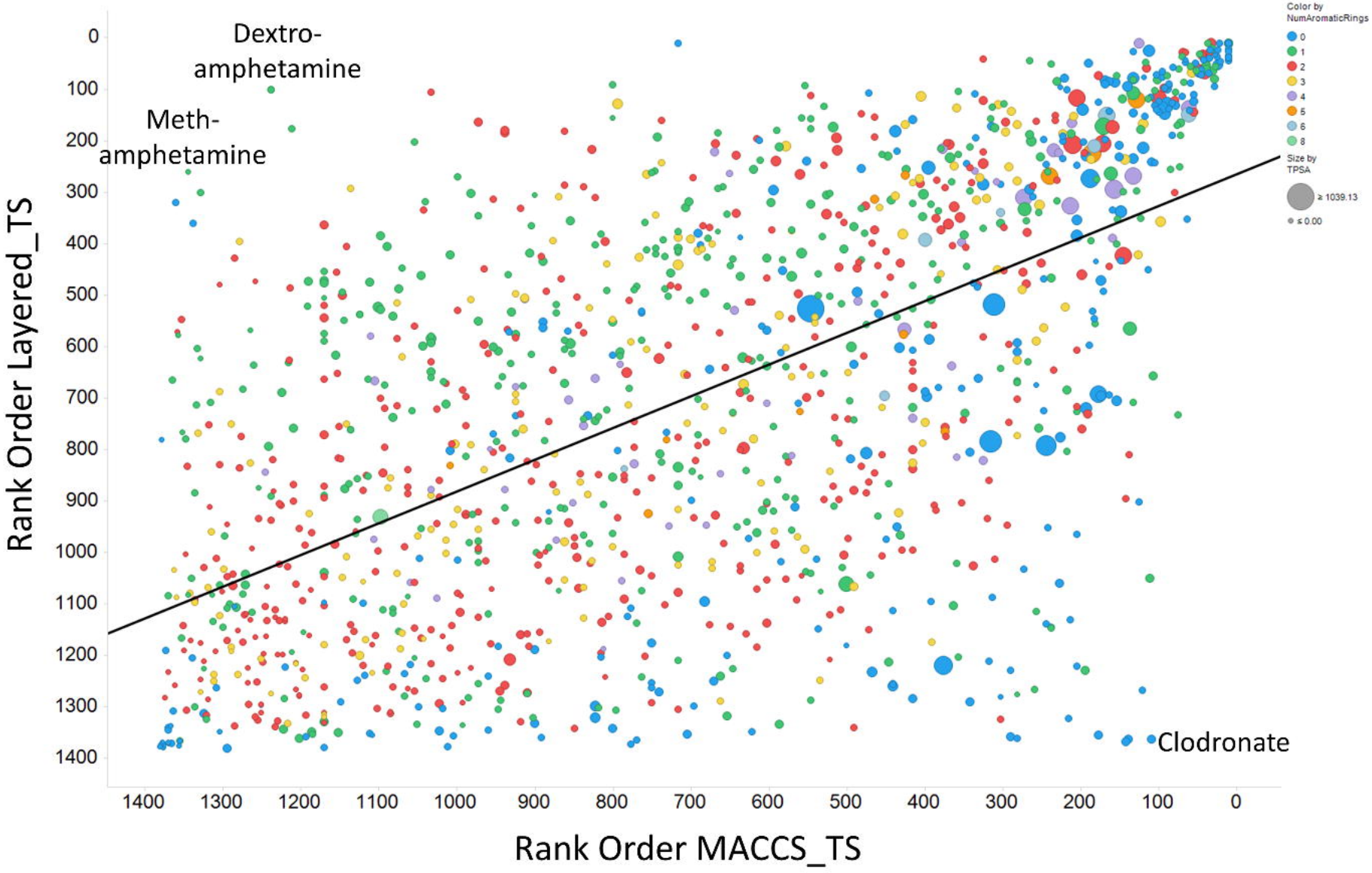
Rank Order of Layered_TS vs MACCS_TS

B

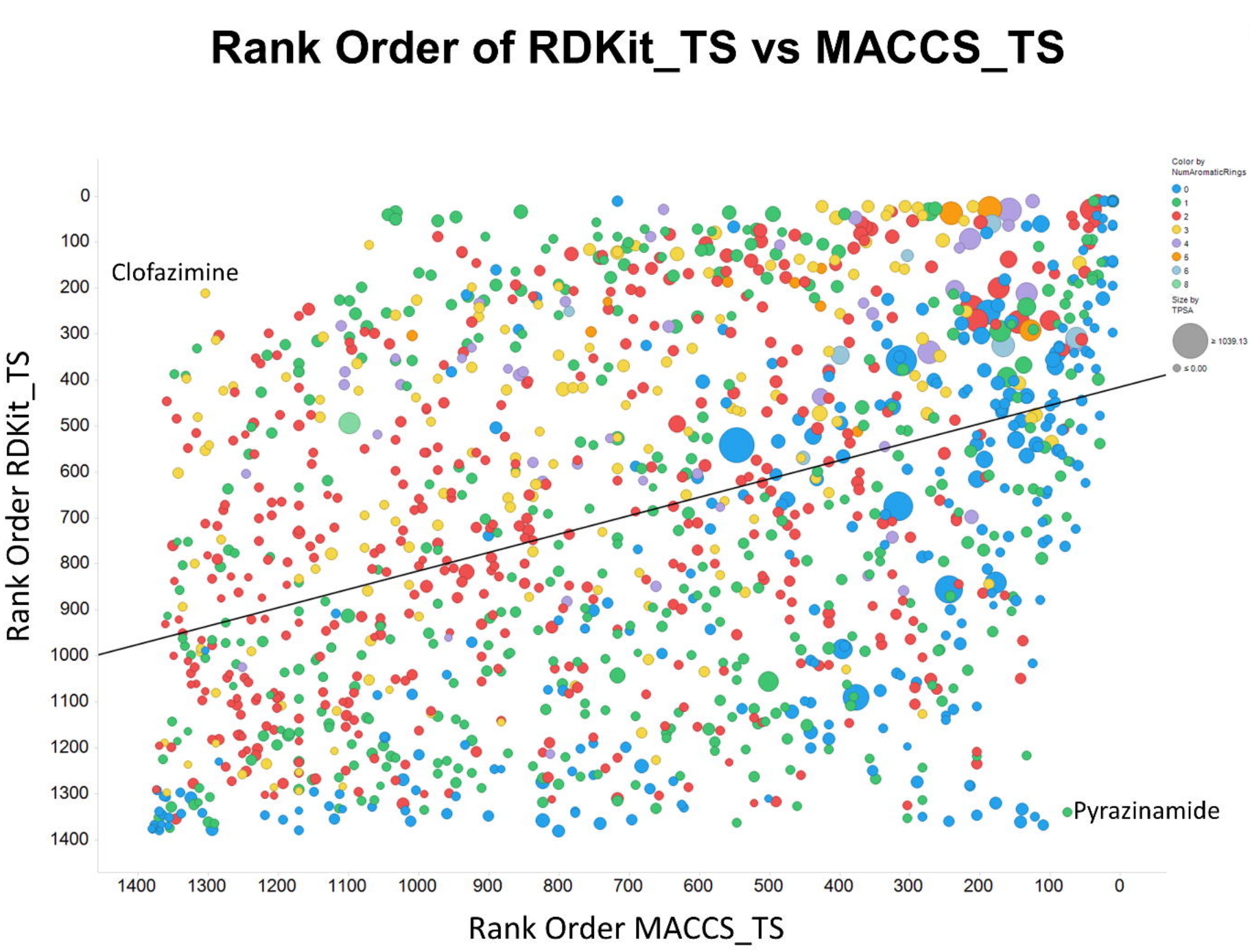


Rank Order of Layered_TS vs MACCS_TS

B

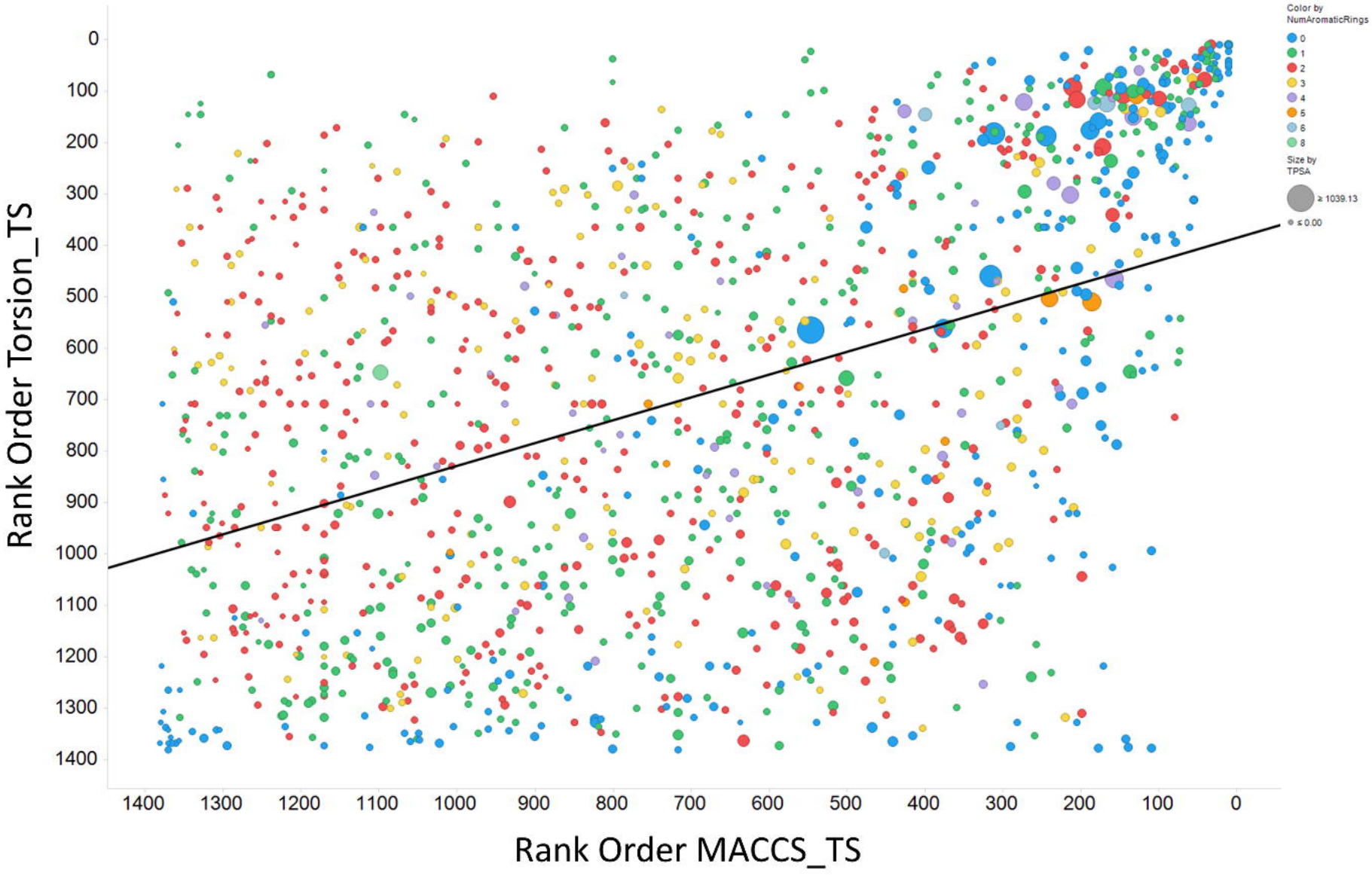


C


Rank Order of RDKit_TS vs MACCS_TS
Clofazimine
Pyrazinamide
Rank Order RDKit_TS
Rank Order MACCS_TS
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
Color by NumAromaticRings
0
1
2
3
4
5
6
8
Size by TPSA
≥ 1039.13
≤ 0.00

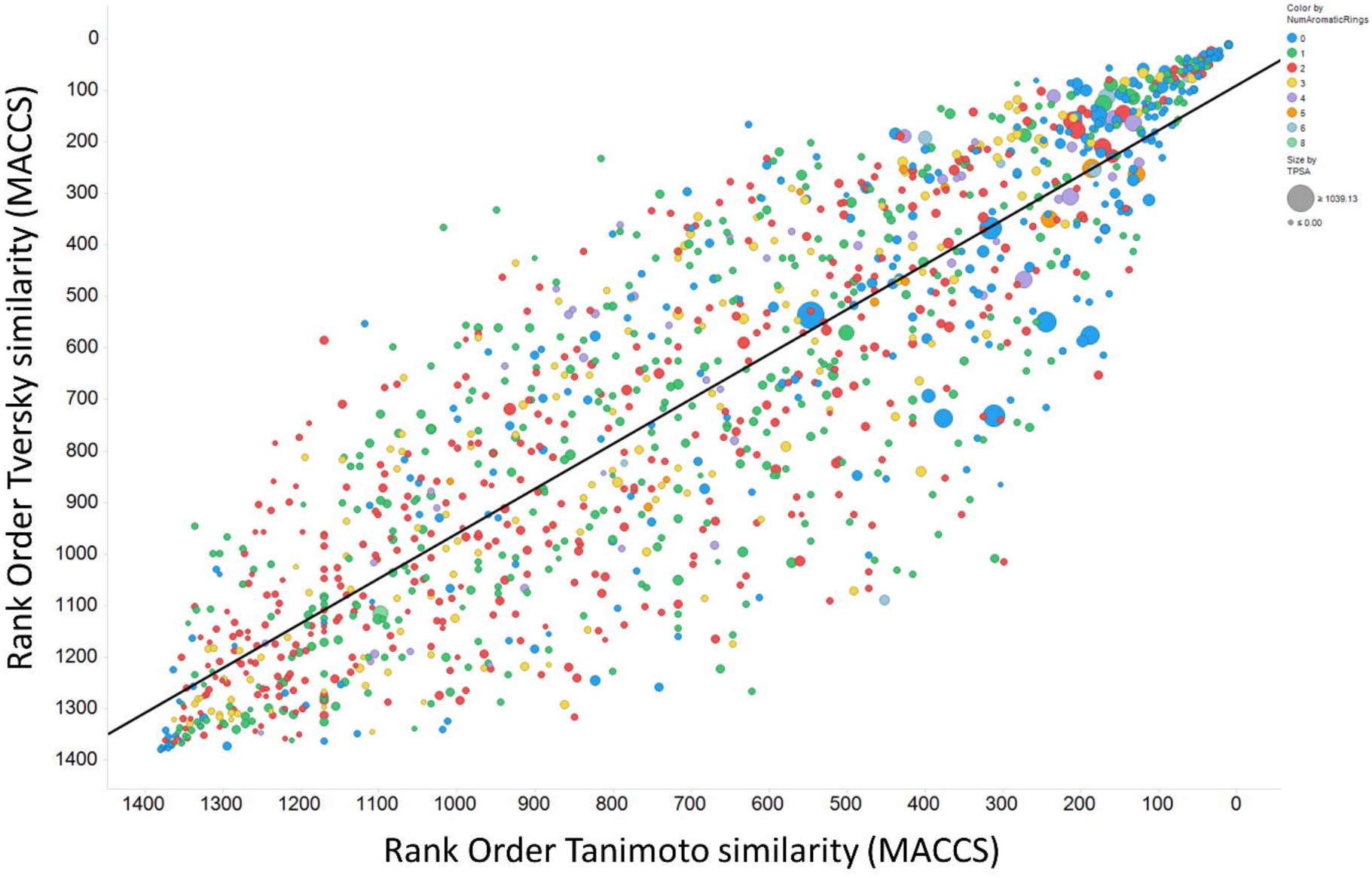
D

Rank Order of Torsion_TS vs MACCS_TS

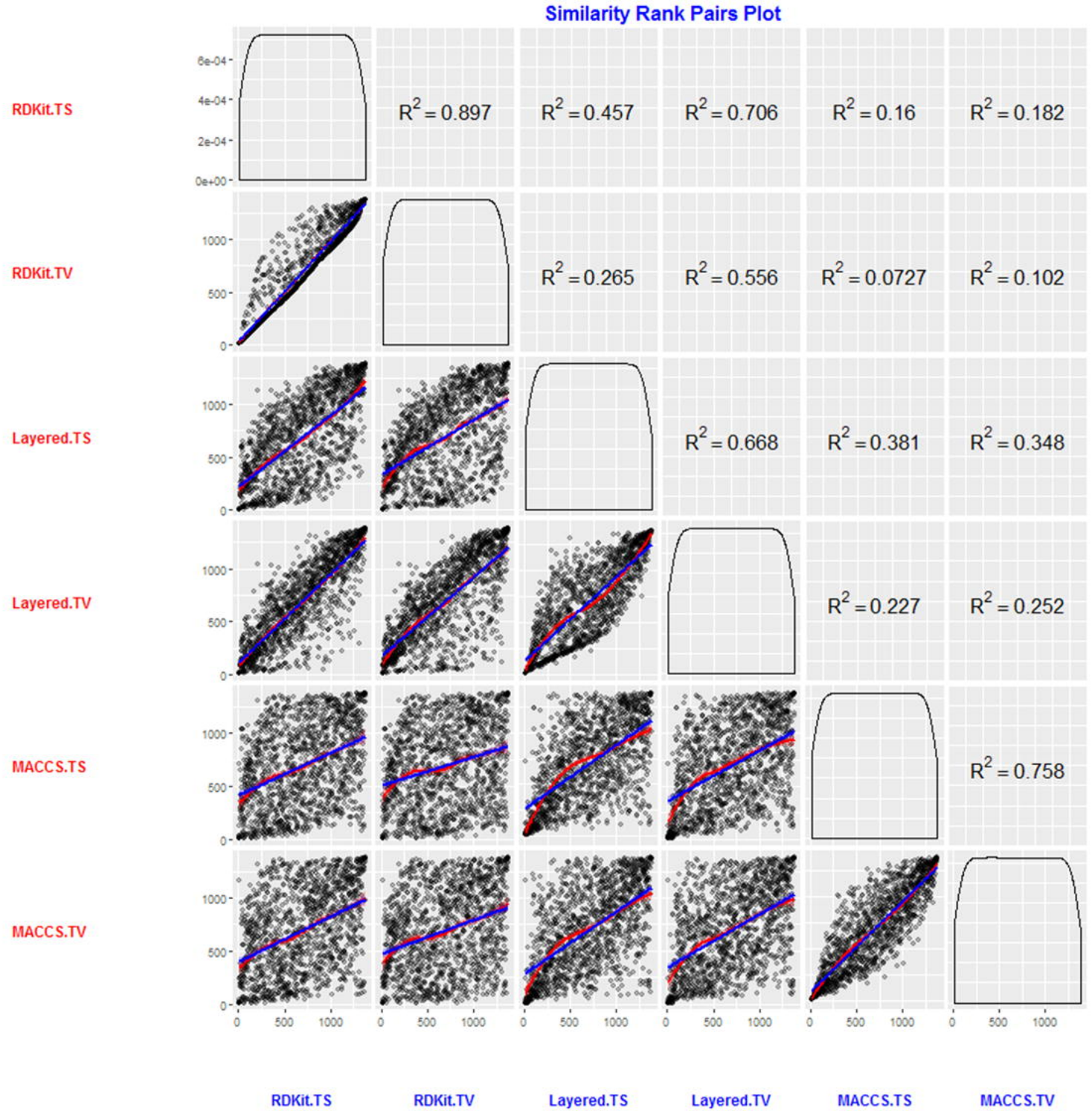


Rank Order of Tversky vs Tanimoto similarity (MACCS)

E

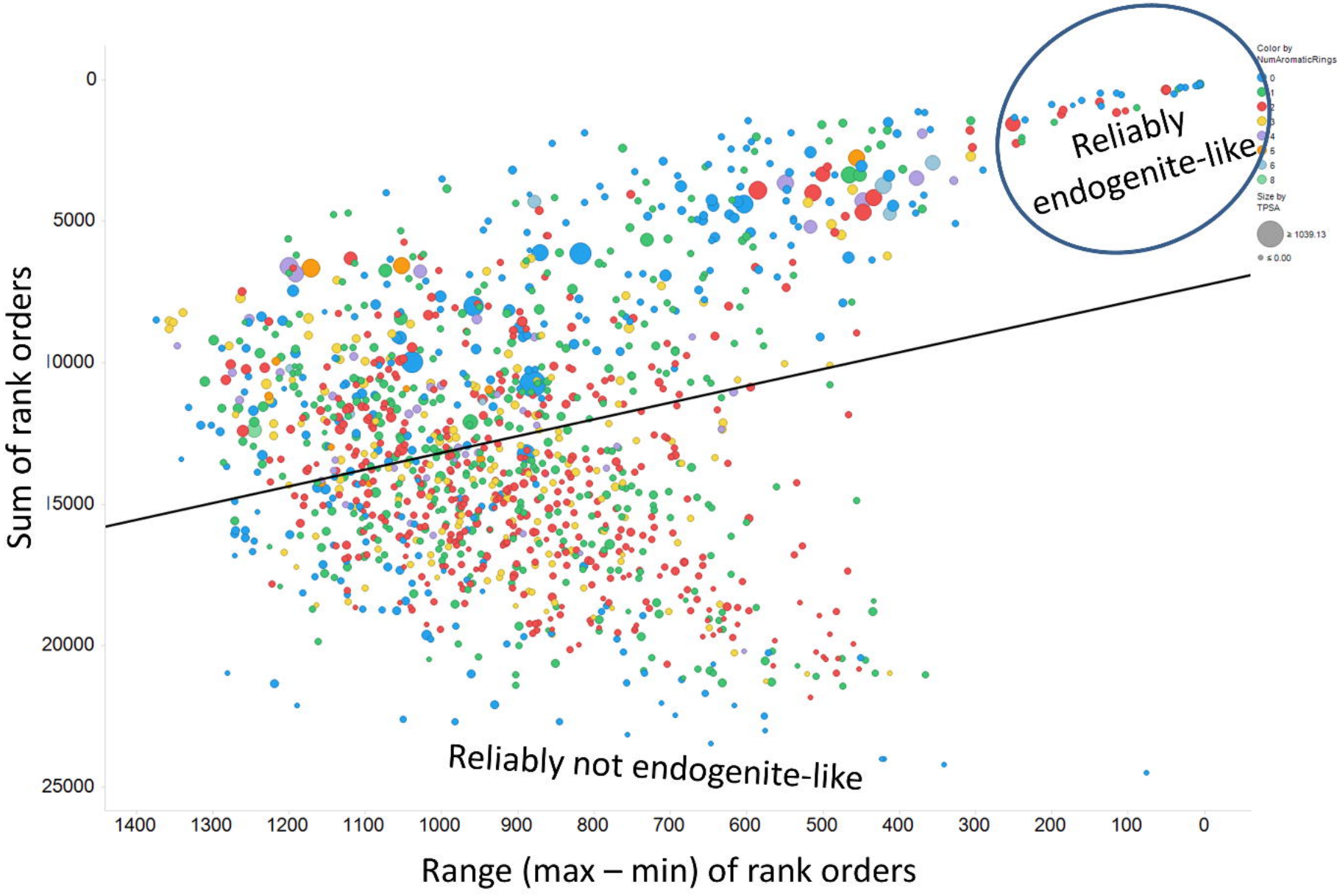


F


Similarity Rank Pairs Plot
RDKit.TS
RDKit.TV
Layered.TS
Layered.TV
MACCS.TS
MACCS.TV
$R^2 = 0.897$
$R^2 = 0.457$
$R^2 = 0.706$
$R^2 = 0.16$
$R^2 = 0.182$
$R^2 = 0.265$
$R^2 = 0.556$
$R^2 = 0.0727$
$R^2 = 0.102$
$R^2 = 0.668$
$R^2 = 0.381$
$R^2 = 0.348$
$R^2 = 0.227$
$R^2 = 0.252$
$R^2 = 0.758$
6e-04
4e-04
2e-04
0e+00
1000
500
0
0
500
1000

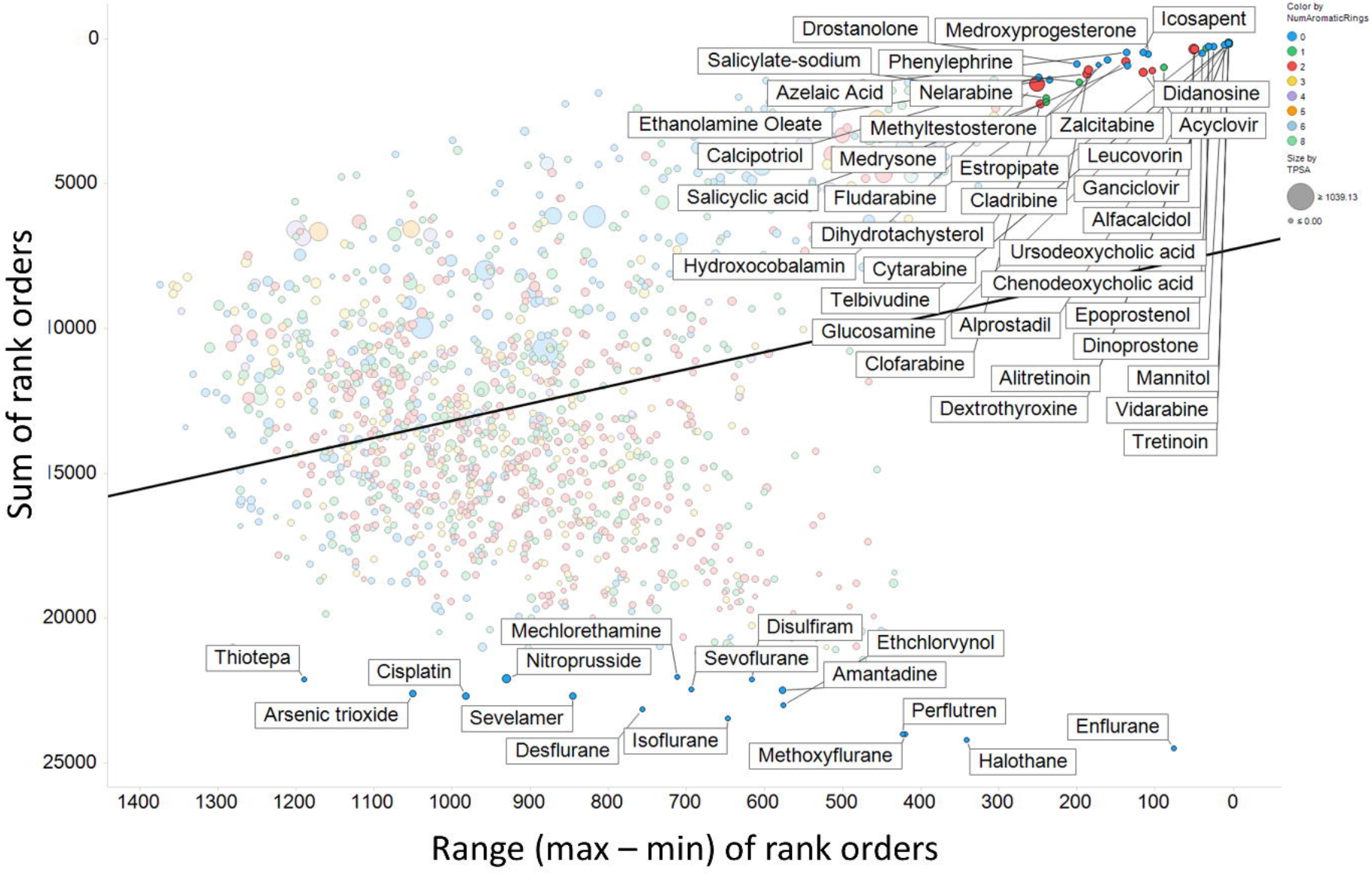
A

Sum of rank orders vs range



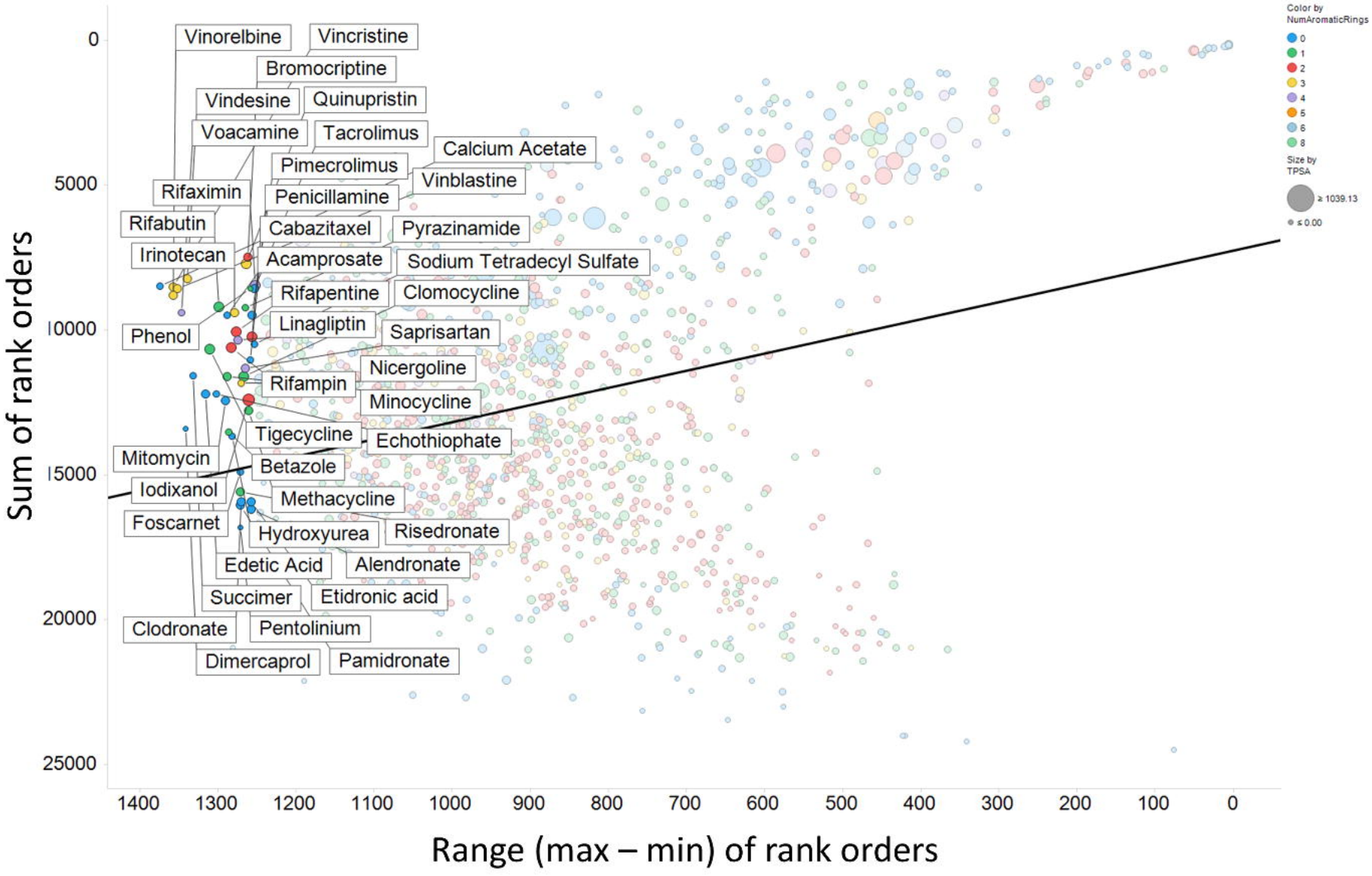
B

Sum of rank orders vs range



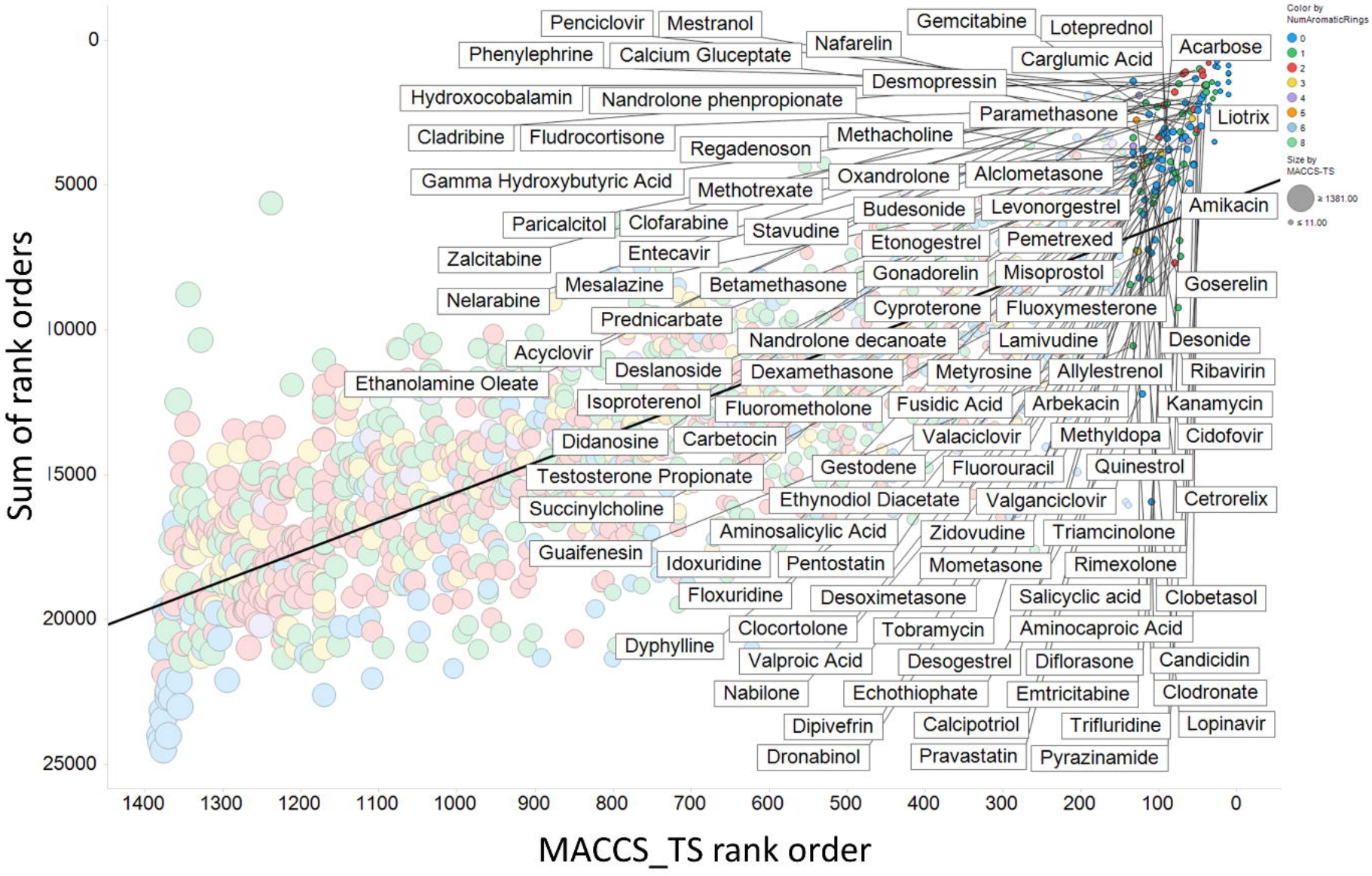
C

Sum of rank orders vs range



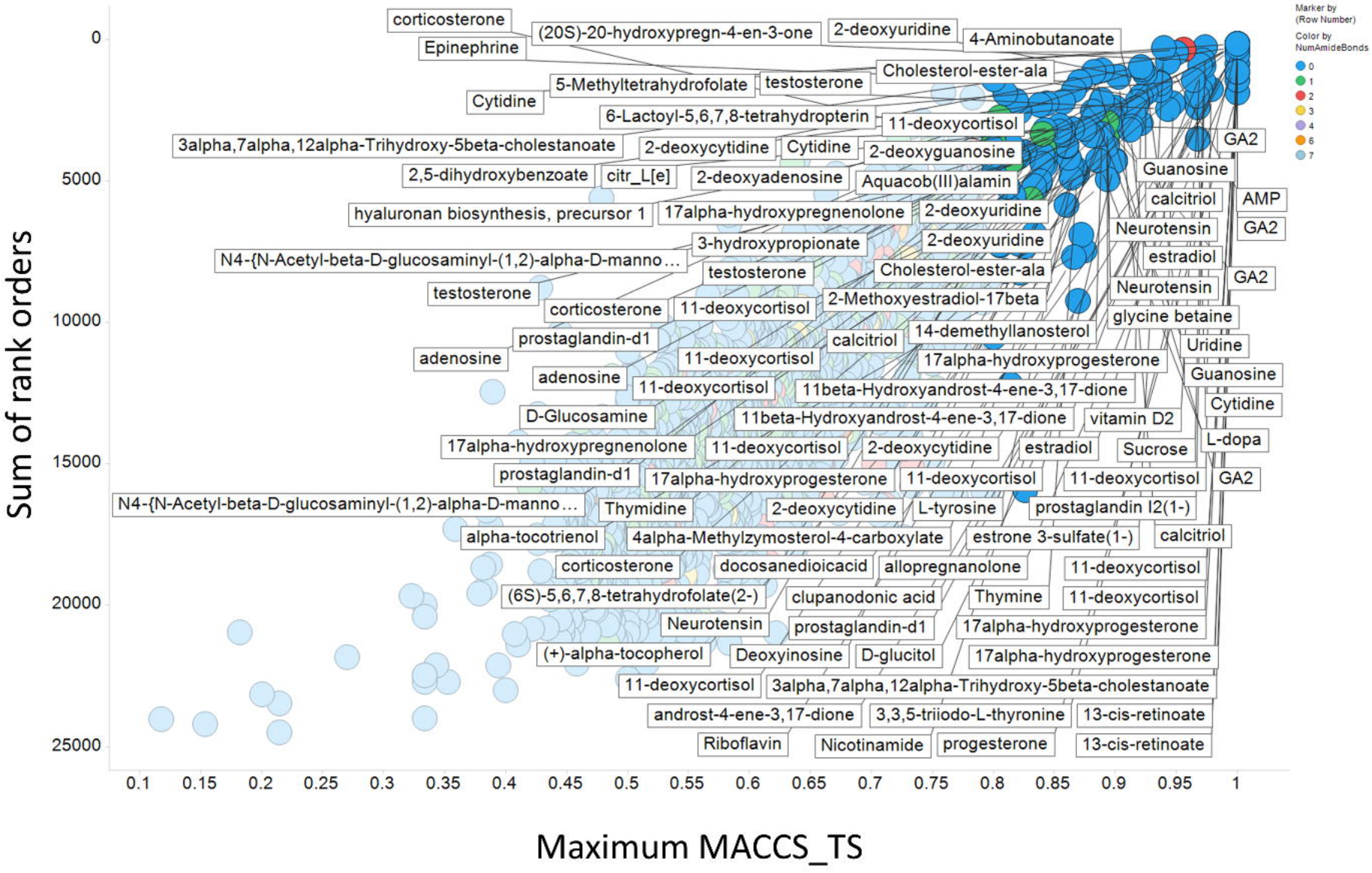
D

Sum of rank orders MACCS_TS vs rank order



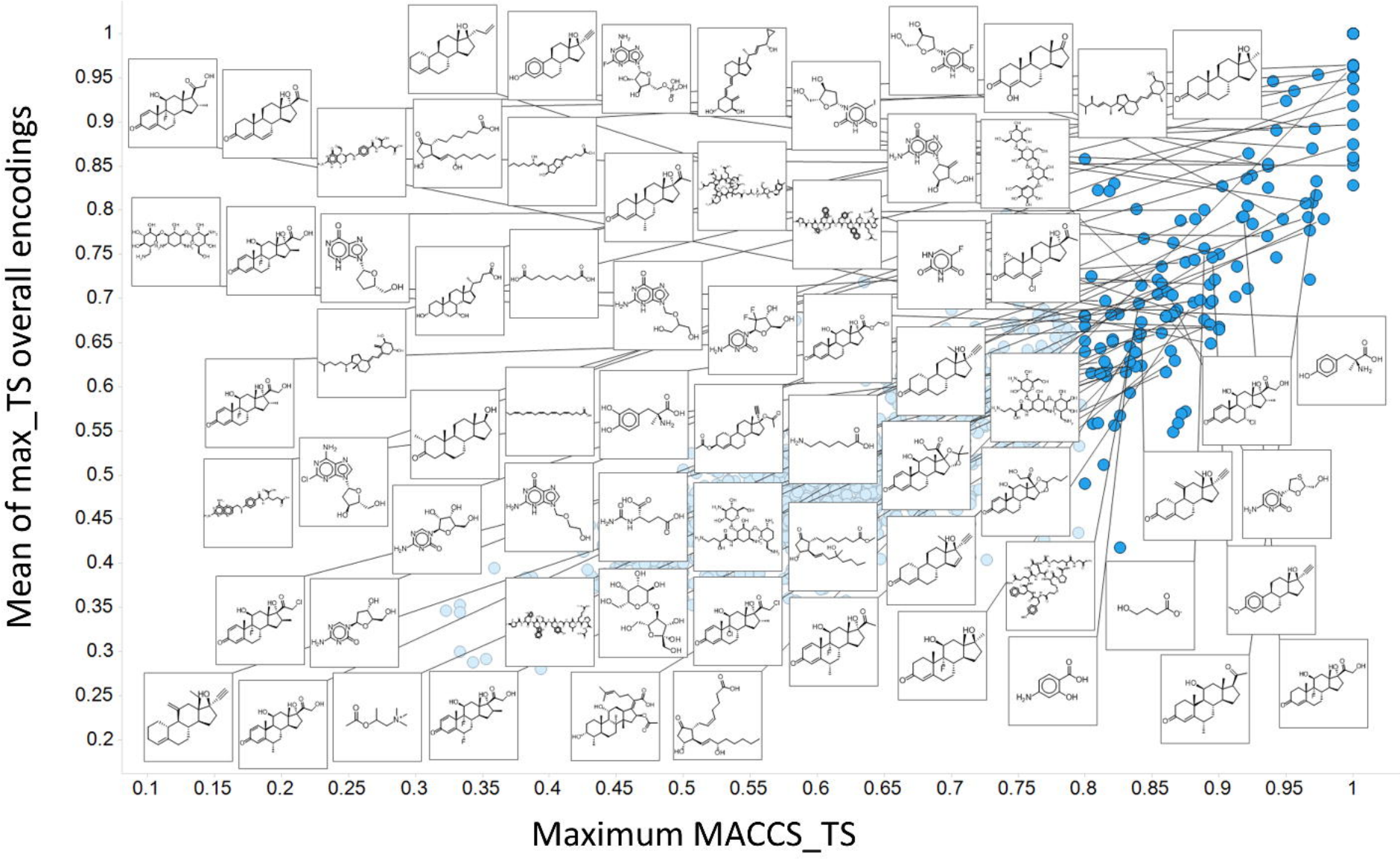
E

Sum of rank orders vs Maximum MACCS_TS

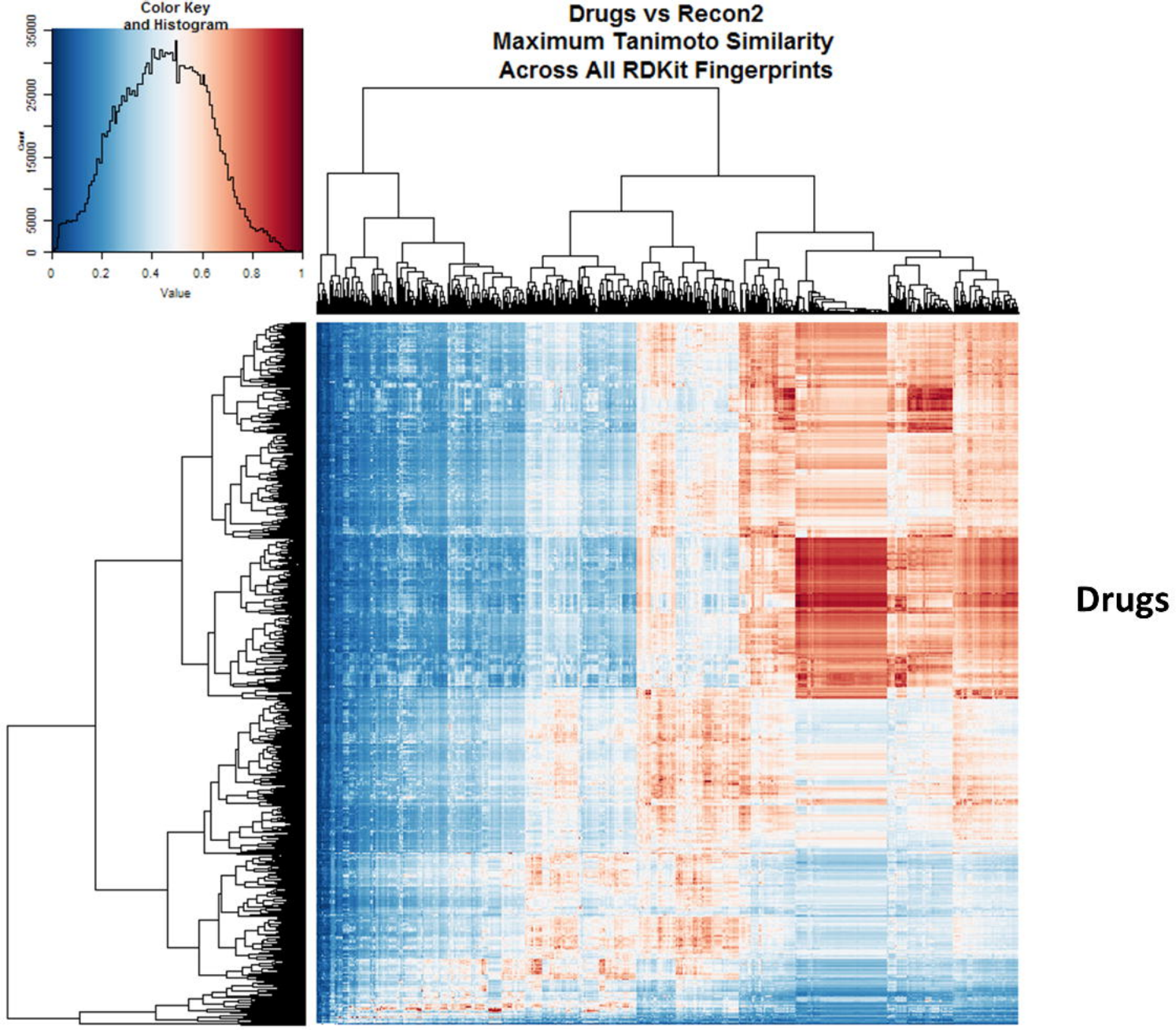


F

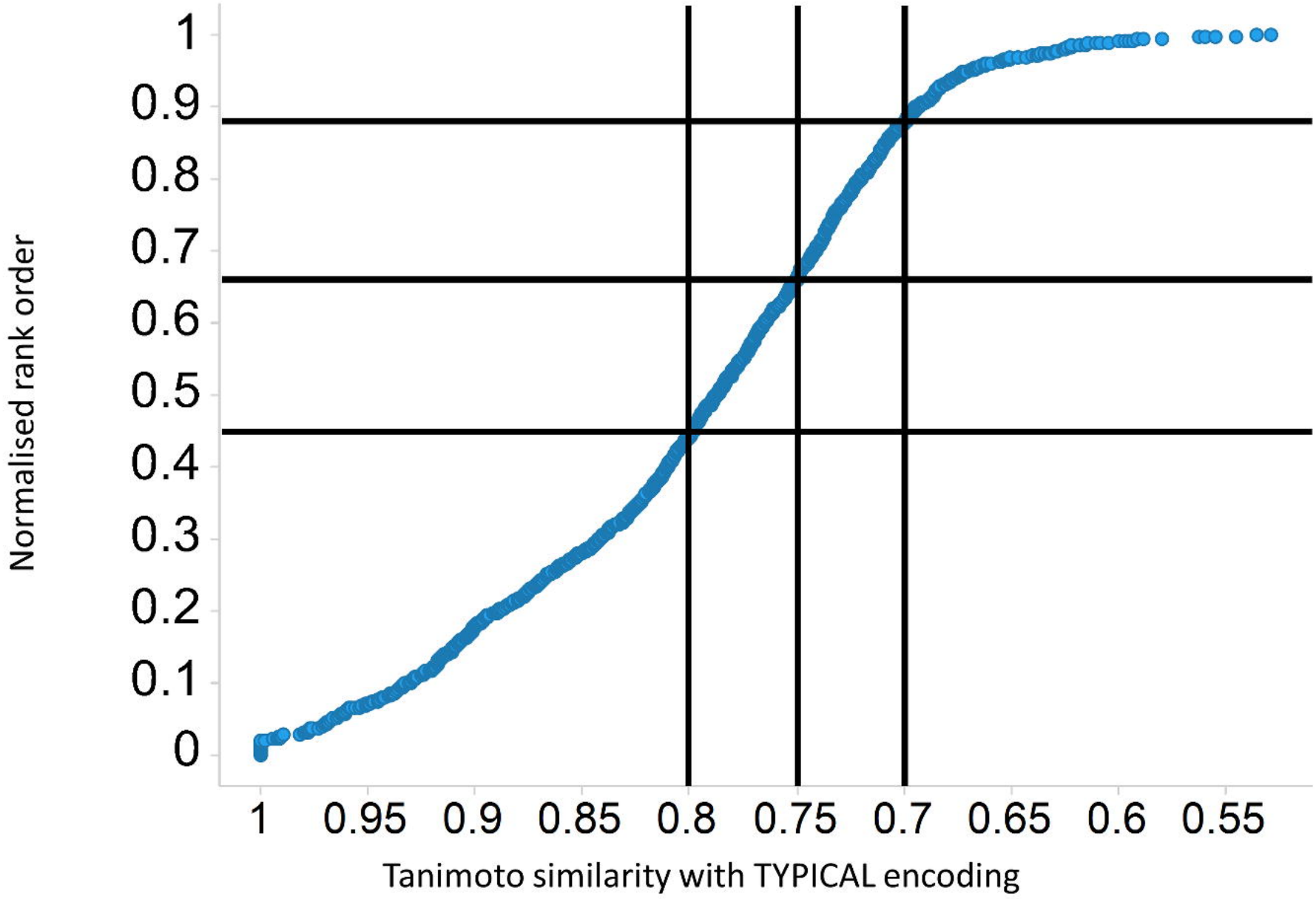
Mean of maximum_TS vs Maximum MACCS_TS

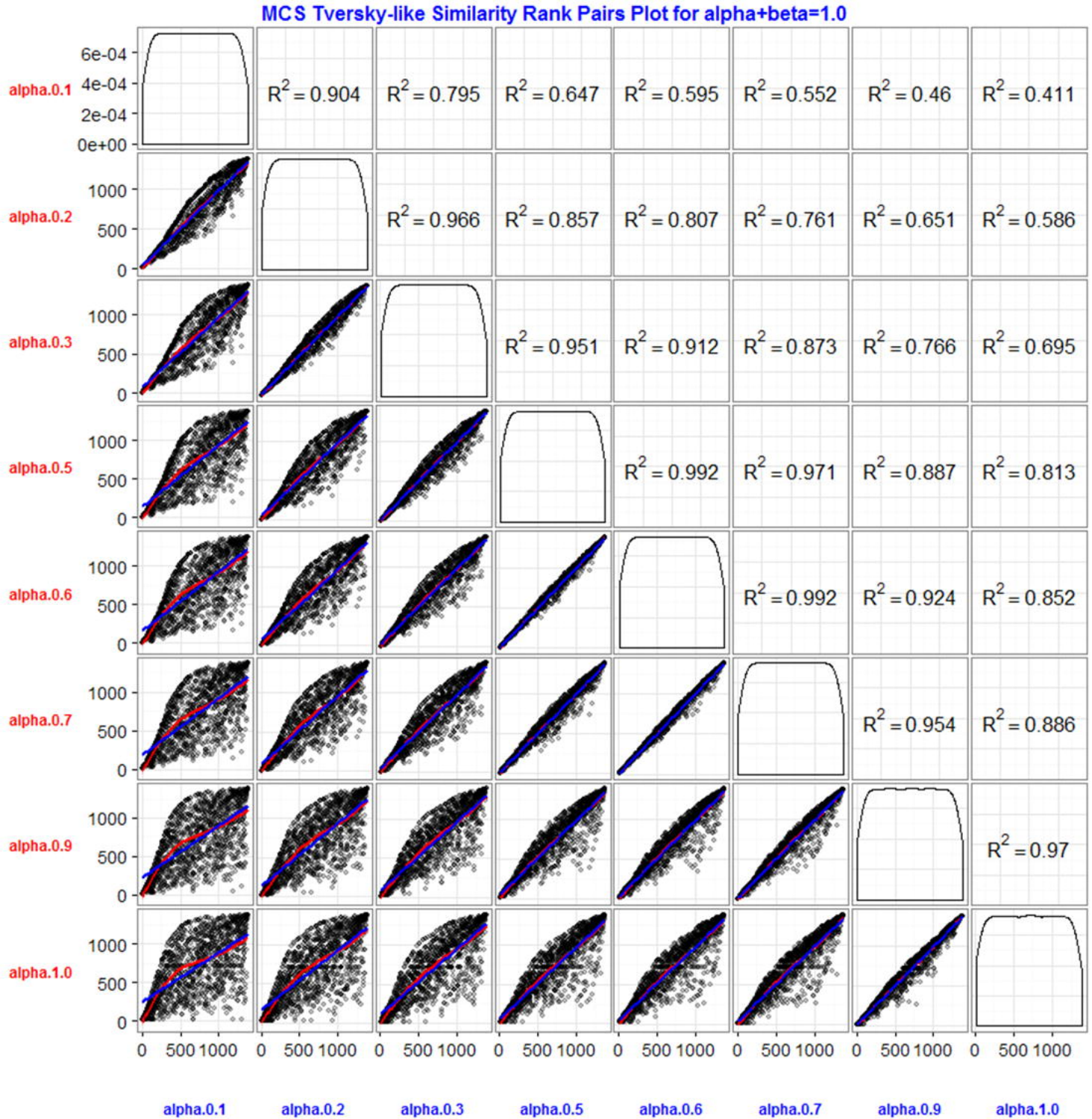


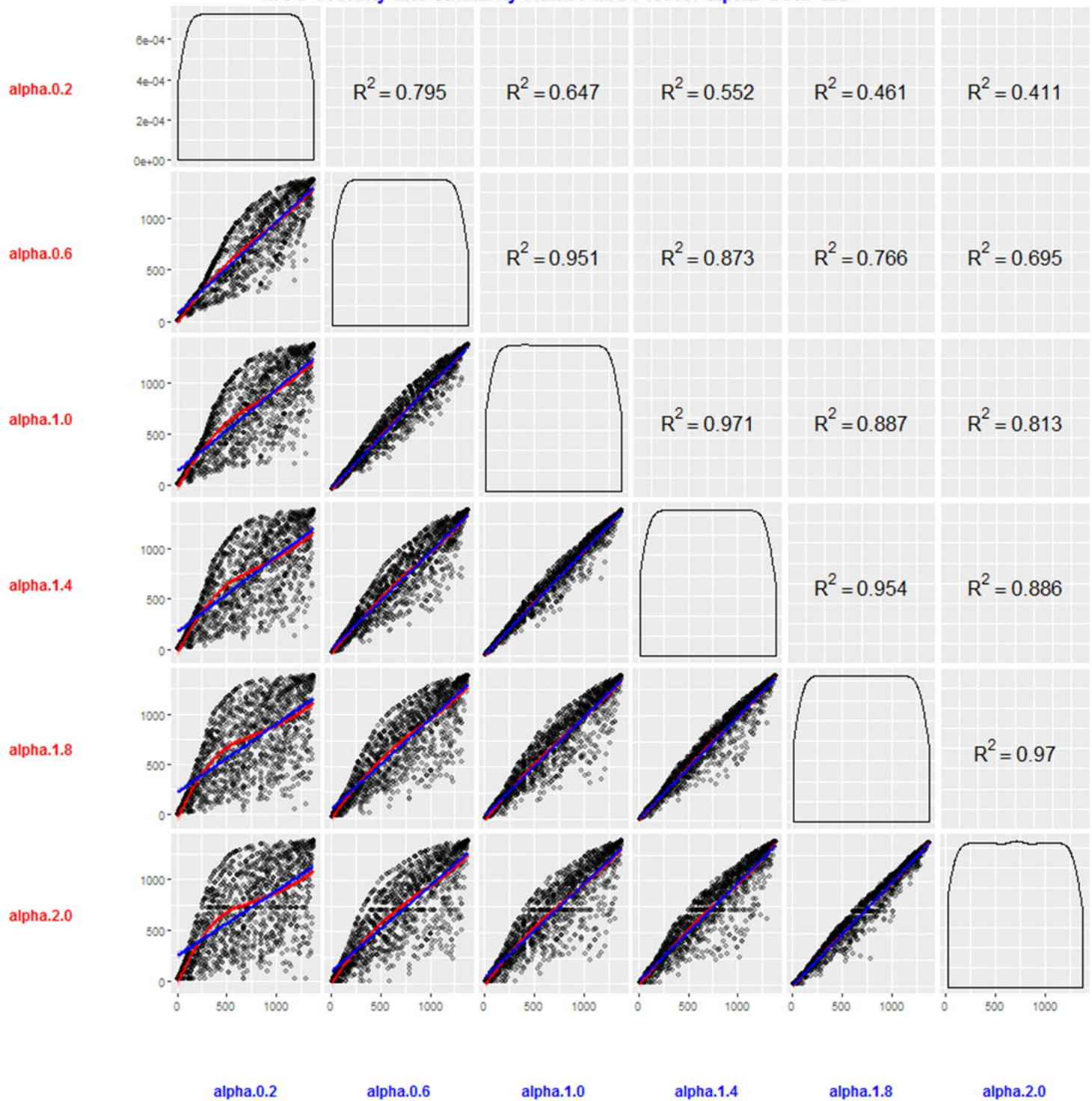
G

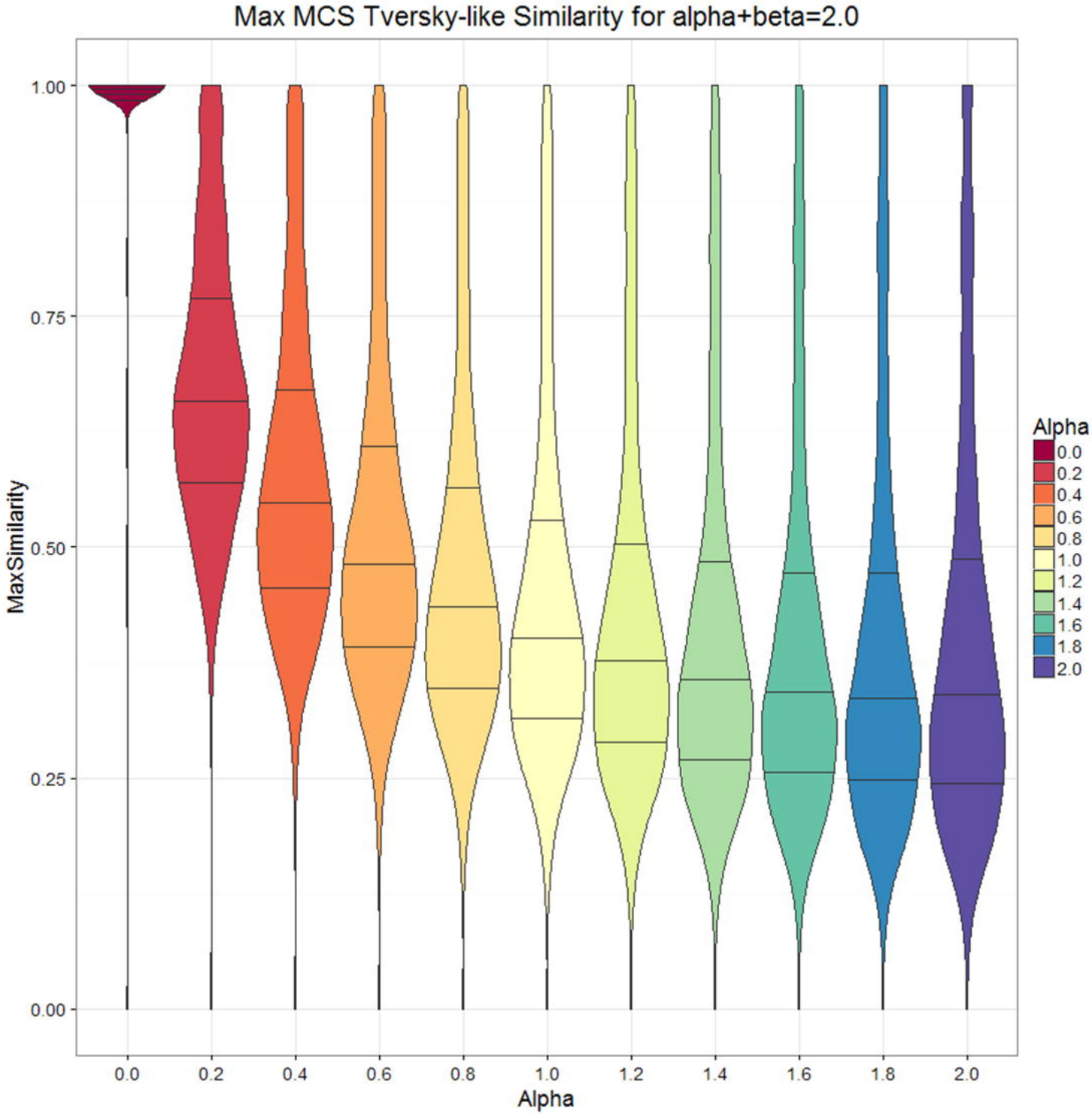

Color Key
and Histogram
Count
35000
25000
15000
5000
0
0
0.2
0.4
0.6
0.8
1
Value
Drugs vs Recon2
Maximum Tanimoto Similarity
Across All RDKit Fingerprints
Drugs
Recon2 metabolites

Drugs vs endogenites: normalised rank vs TYPICAL encoding (maximal Tanimoto similarity)



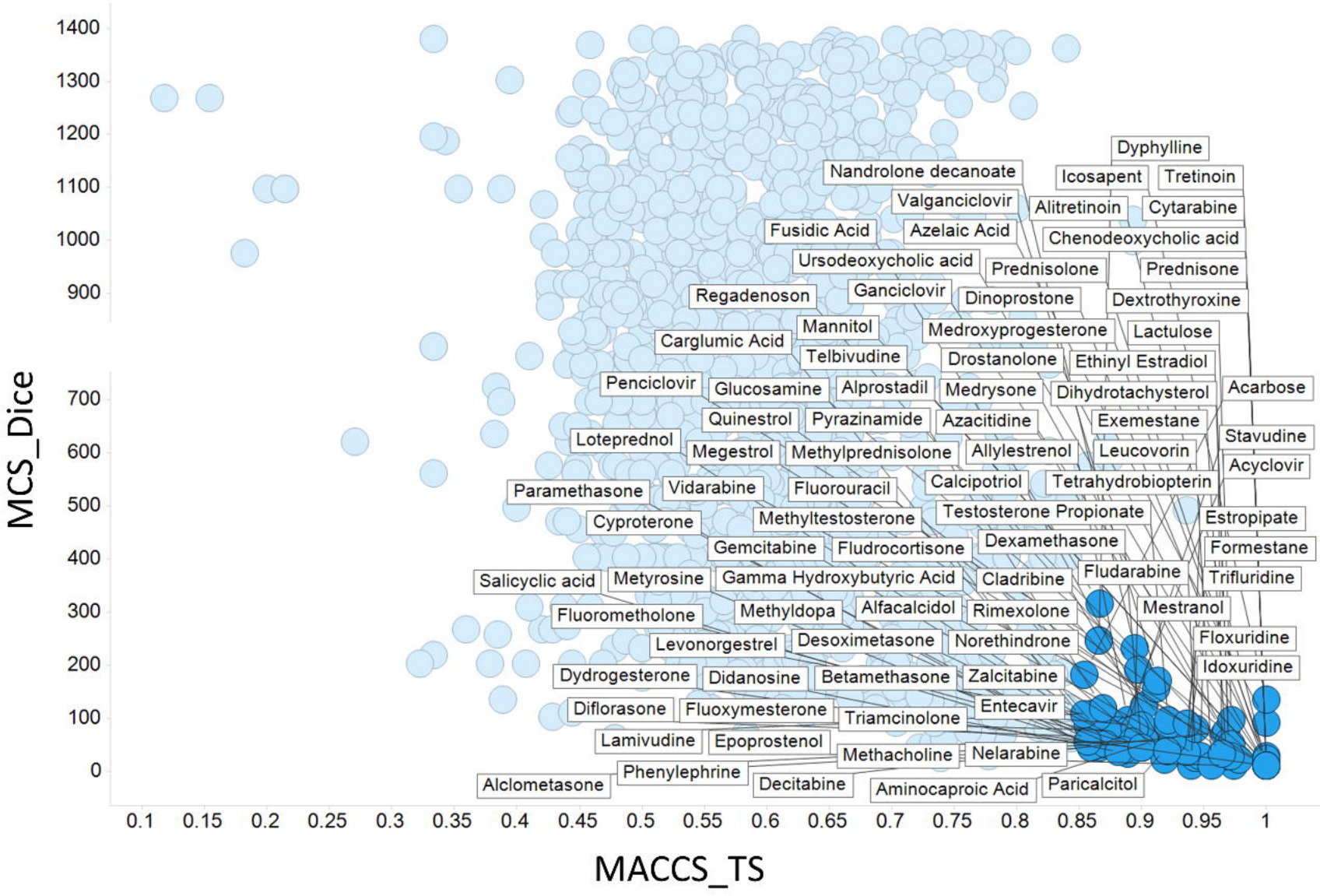

MCS Tversky-like Similarity Rank Pairs Plot for alpha+beta=1.0
A
alpha.0.1
alpha.0.2
alpha.0.3
alpha.0.5
alpha.0.6
alpha.0.7
alpha.0.9
alpha.1.0
6e-04
4e-04
2e-04
0e+00
1000
500
0
0 500 1000
R² = 0.904
R² = 0.795
R² = 0.647
R² = 0.595
R² = 0.552
R² = 0.46
R² = 0.411
R² = 0.966
R² = 0.857
R² = 0.807
R² = 0.761
R² = 0.651
R² = 0.586
R² = 0.951
R² = 0.912
R² = 0.873
R² = 0.766
R² = 0.695
R² = 0.992
R² = 0.971
R² = 0.887
R² = 0.813
R² = 0.992
R² = 0.924
R² = 0.852
R² = 0.954
R² = 0.886
R² = 0.97

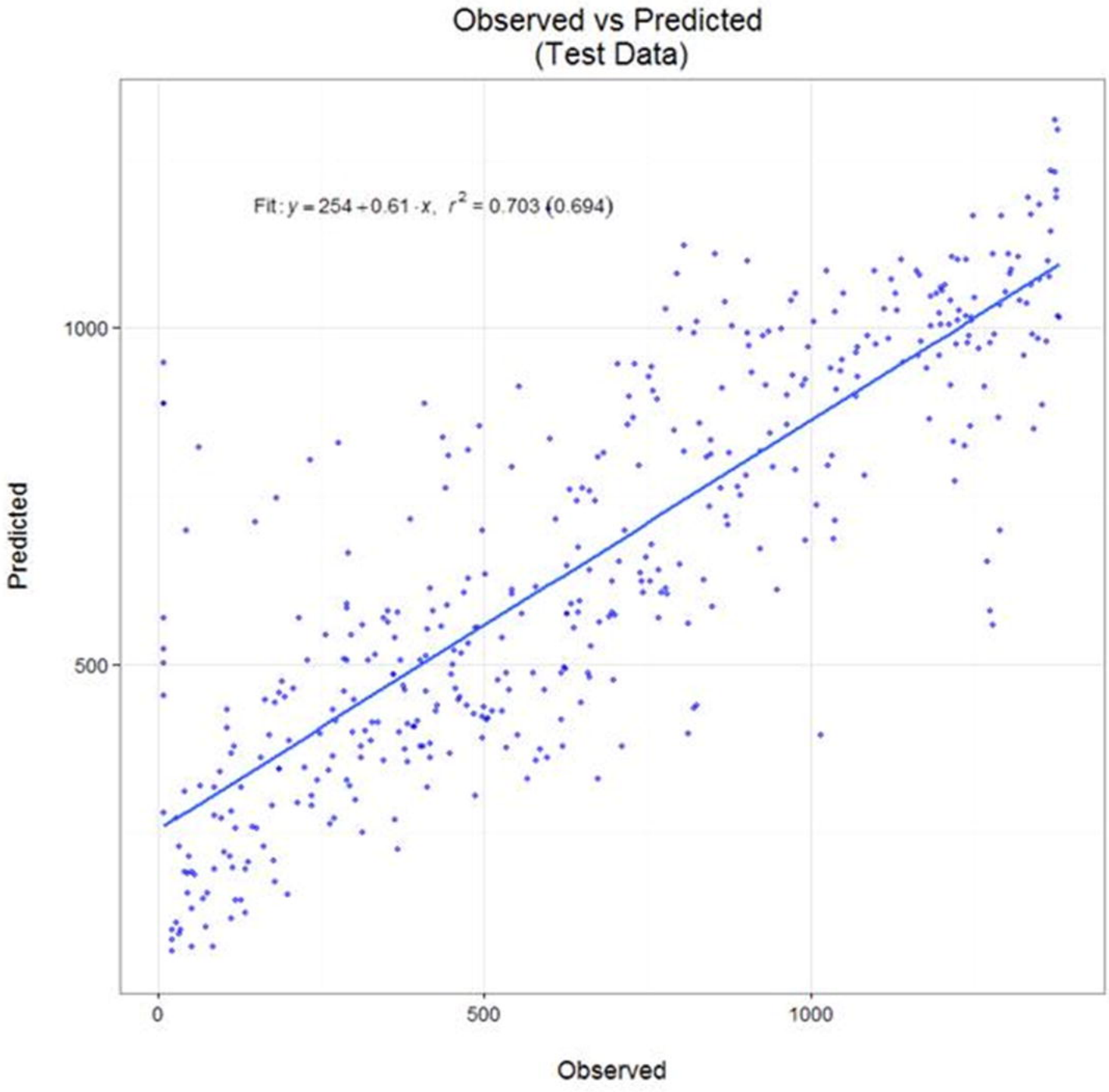

MCS Tversky-like Similarity Rank Pairs Plot for alpha+beta=2.0
B
alpha.0.2
alpha.0.6
alpha.1.0
alpha.1.4
alpha.1.8
alpha.2.0
R2 = 0.795
R2 = 0.647
R2 = 0.552
R2 = 0.461
R2 = 0.411
R2 = 0.951
R2 = 0.873
R2 = 0.766
R2 = 0.695
R2 = 0.971
R2 = 0.887
R2 = 0.813
R2 = 0.954
R2 = 0.886
R2 = 0.97
6e-04
4e-04
2e-04
0e+00
1000
500
0
0
500
1000


Max MCS Tversky-like Similarity for alpha+beta=2.0
C
MaxSimilarity
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Alpha
Alpha
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0

D

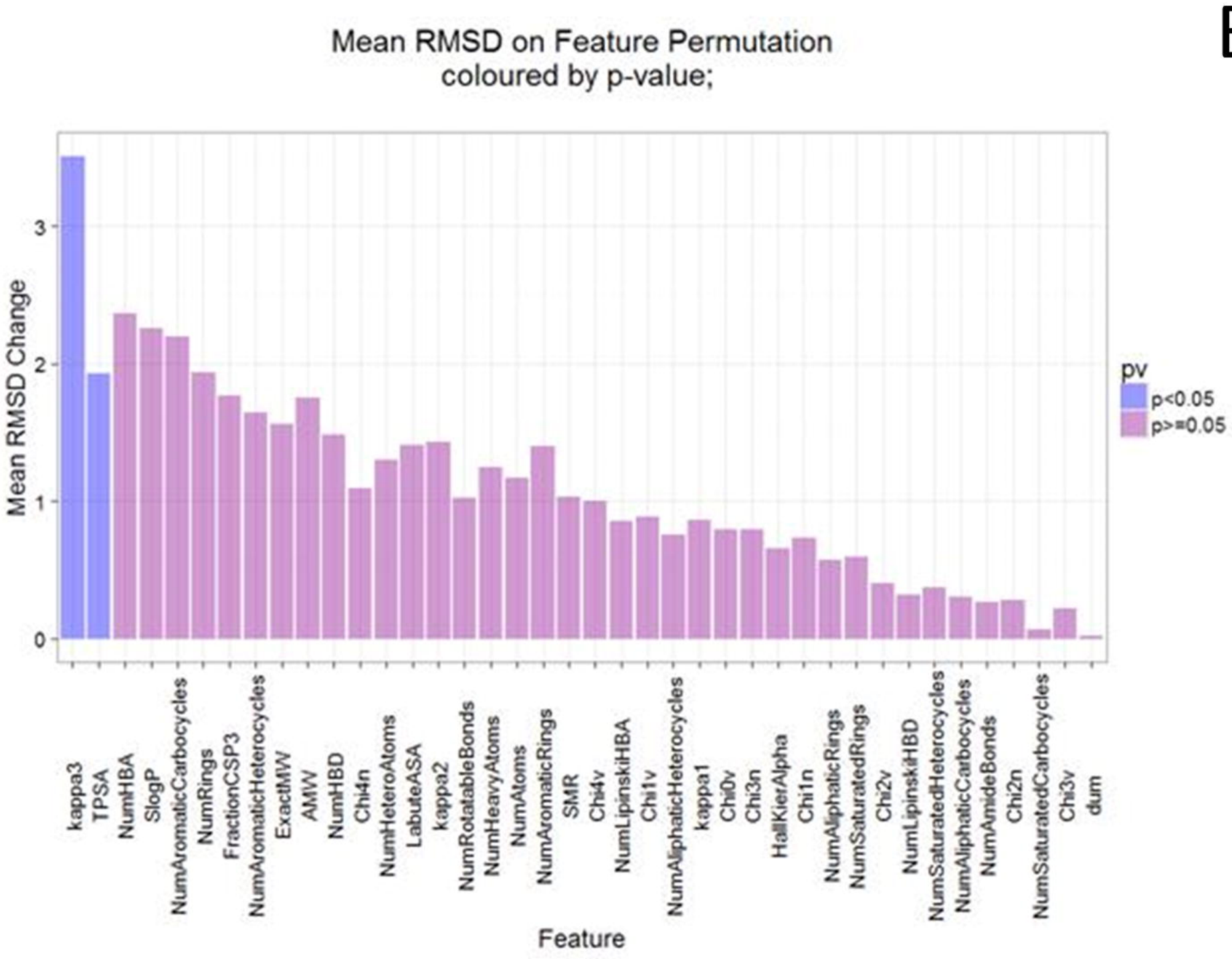
MCS Dice coefficient vs MACCS_TS



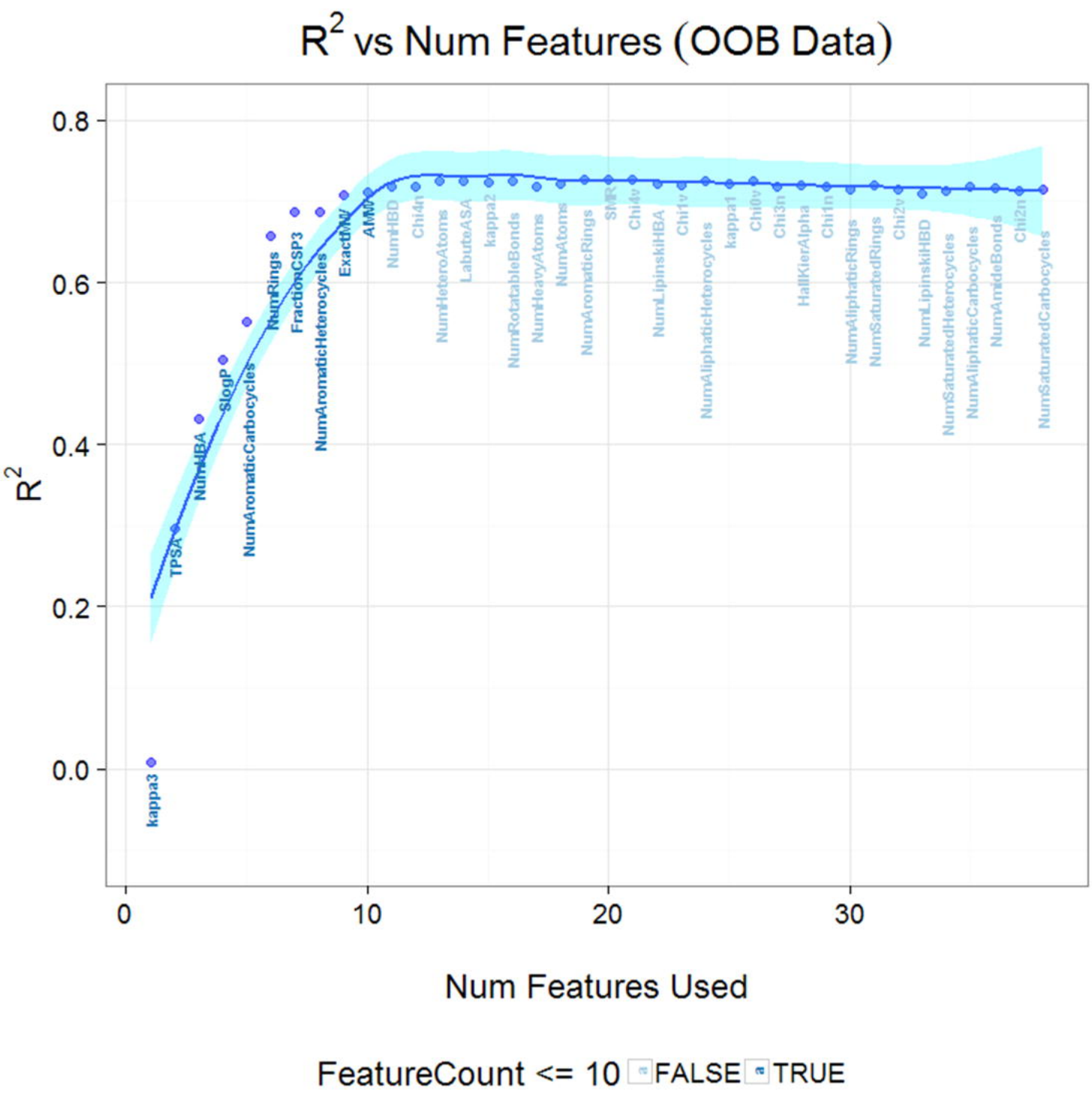

Observed vs Predicted
(Test Data)
Fit: y = 254 + 0.61 · x, r² = 0.703 (0.694)
1000
500
Predicted
0
500
1000
Observed

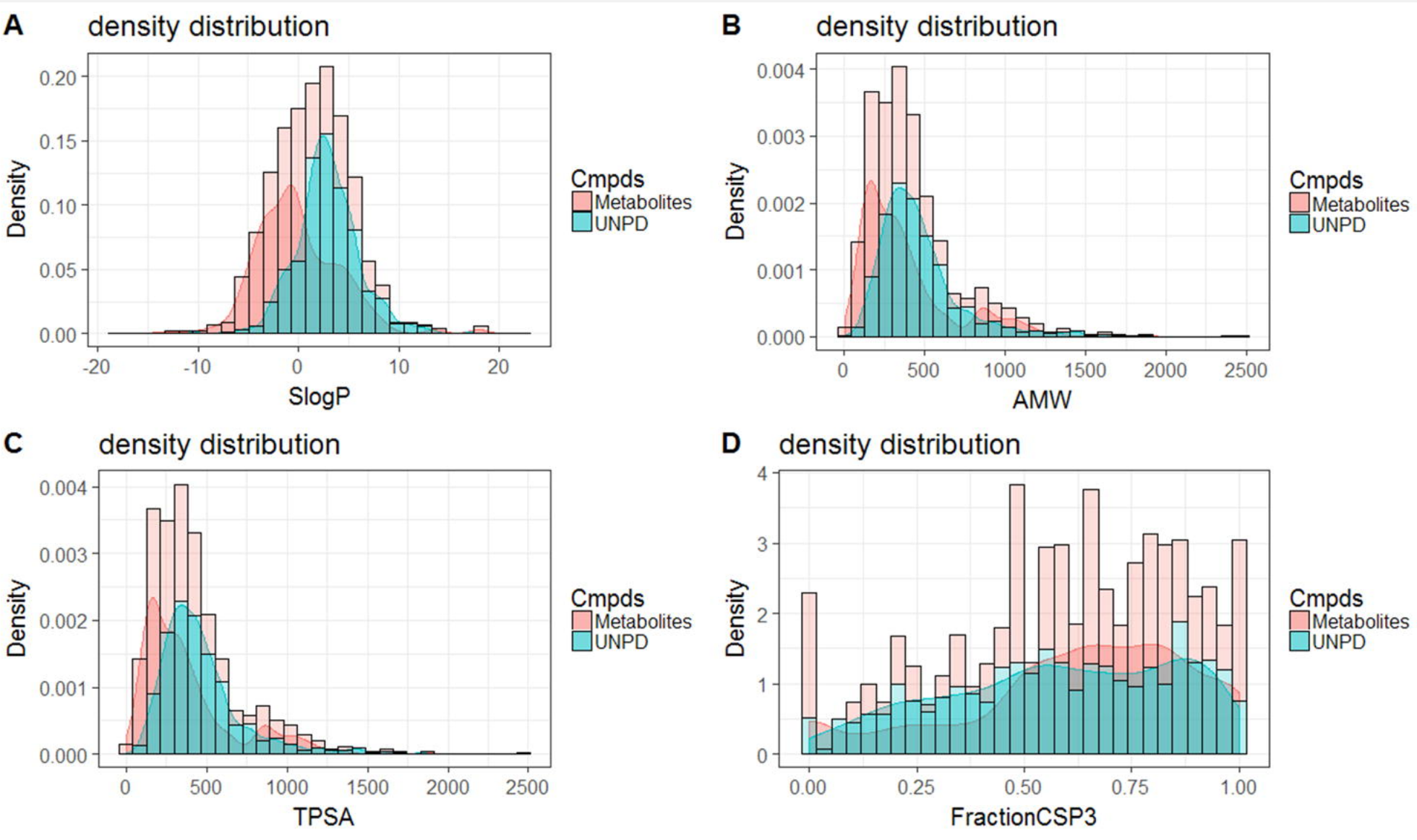
A

B

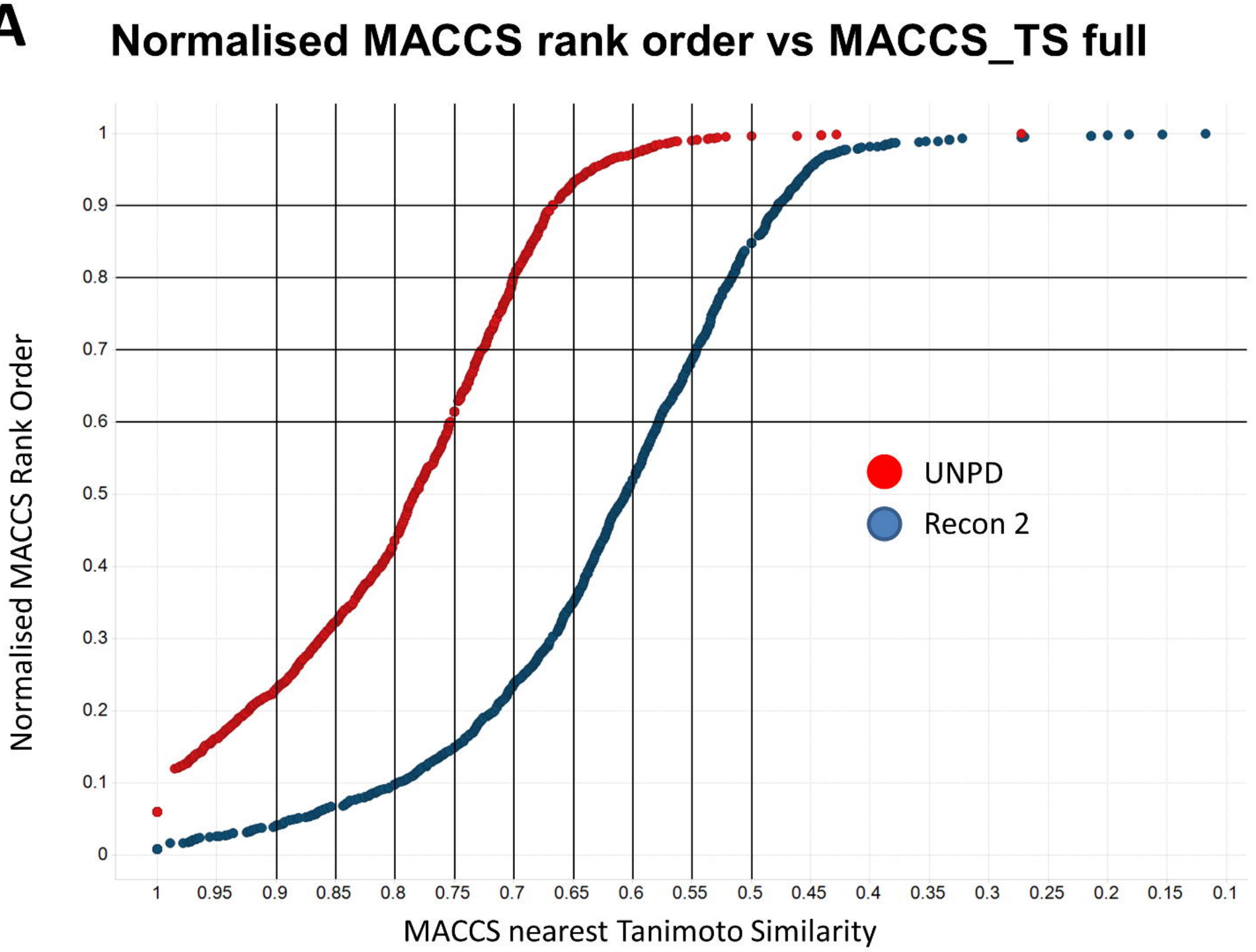

Mean RMSD on Feature Permutation
coloured by p-value;
Mean RMSD Change
0
1
2
3
pv
p<0.05
p>=0.05
kappa3
TPSA
NumHBA
SlogP
NumAromaticCarbocycles
NumRings
FractionCSP3
NumAromaticHeterocycles
ExactMW
AMW
NumHBD
Chi4n
NumHeteroAtoms
LabuteASA
kappa2
NumRotatableBonds
NumHeavyAtoms
NumAtoms
NumAromaticRings
SMR
Chi4v
NumLipinskiHBA
Chi1v
NumAliphaticHeterocycles
kappa1
Chi0v
Chi3n
HallKierAlpha
Chi1n
NumAliphaticRings
NumSaturatedRings
Chi2v
NumLipinskiHBD
NumSaturatedHeterocycles
NumAliphaticCarbocycles
NumAmideBonds
Chi2n
NumSaturatedCarbocycles
Chi3v
dum
Feature

C

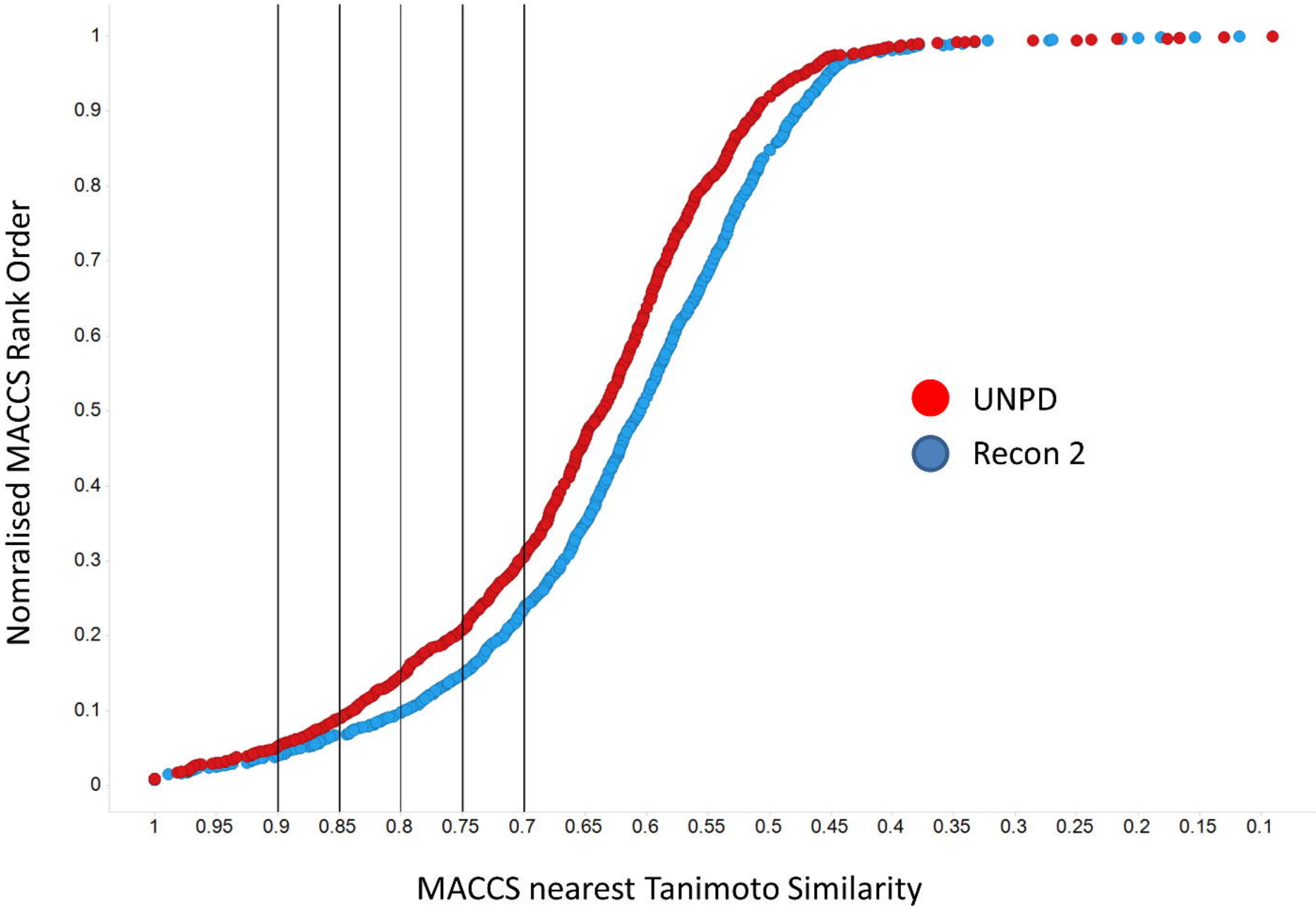

R2 vs Num Features (OOB Data)
R2
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Num Features Used
kappa3
TPSA
NumHBA
SlogP
NumAromaticCarbocycles
NumRings
FractionCSP3
NumAromaticHeterocycles
ExactMW
AMW
NumHBD
Chi4n
NumHeteroAtoms
LabuteASA
kappa2
NumRotatableBonds
NumHeavyAtoms
NumAromaticRings
Chi4v
NumLipinskiHBA
Chi1v
NumAliphaticHeterocycles
kappa1
Chi0v
Chi3n
HallKierAlpha
Chi1n
NumAliphaticRings
NumSaturatedRings
Chi2v
NumLipinskiHBD
NumSaturatedHeterocycles
NumAliphaticCarbocycles
NumAmideBonds
Chi2n
NumSaturatedCarbocycles
FeatureCount <= 10
FALSE
TRUE


A
density distribution
Density
SlogP
B
density distribution
Density
AMW
C
density distribution
Density
TPSA
D
density distribution
Density
FractionCSP3
Cmpds
Metabolites
UNPD

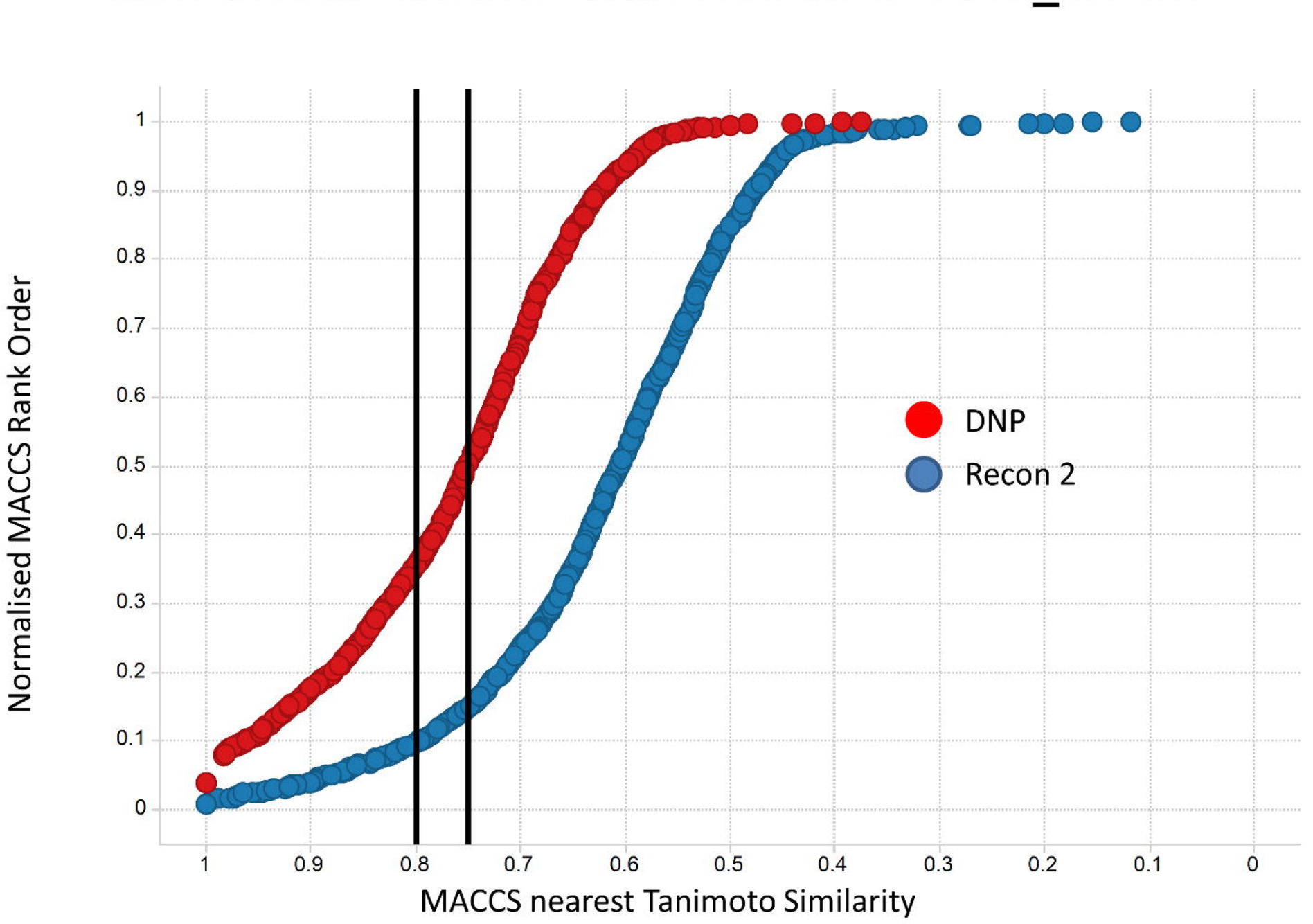
A


Normalised MACCS rank order vs MACCS_TS full
Normalised MACCS Rank Order
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
MACCS nearest Tanimoto Similarity
UNPD
Recon 2

B Normalised MACCS rank order vs MACCS_TS sampled

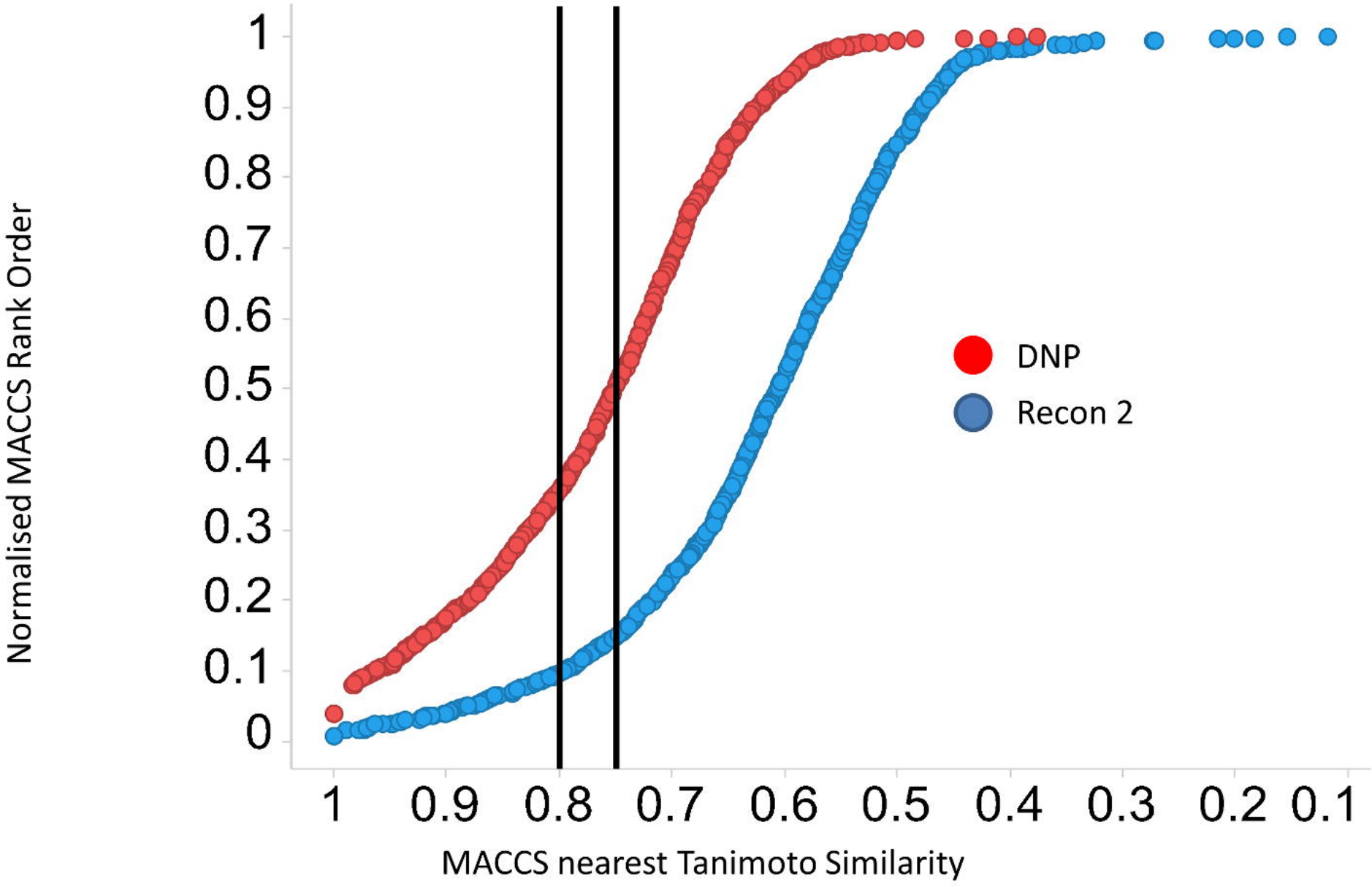


A

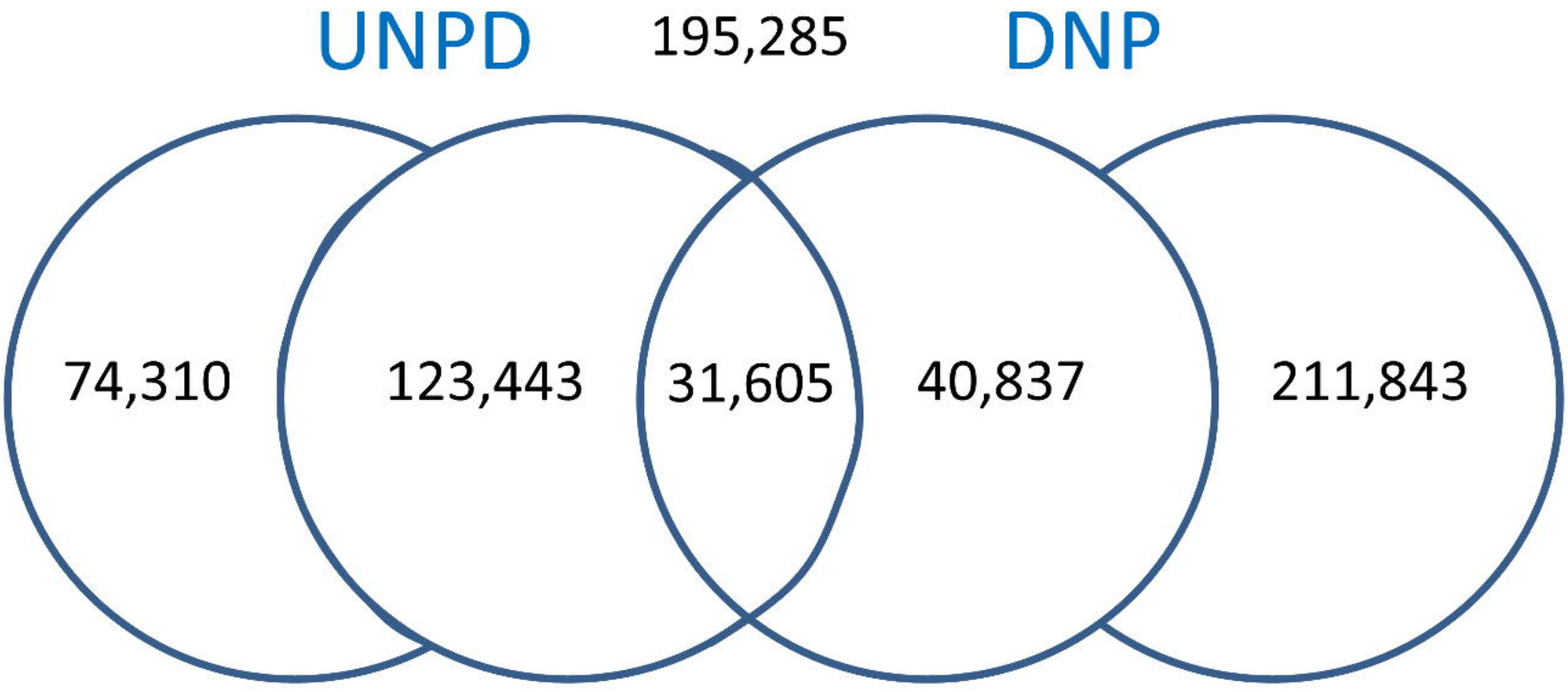

Normalised MACCS rank order vs MACCS_TS full
Normalised MACCS Rank Order
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
DNP
Recon 2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
MACCS nearest Tanimoto Similarity

B

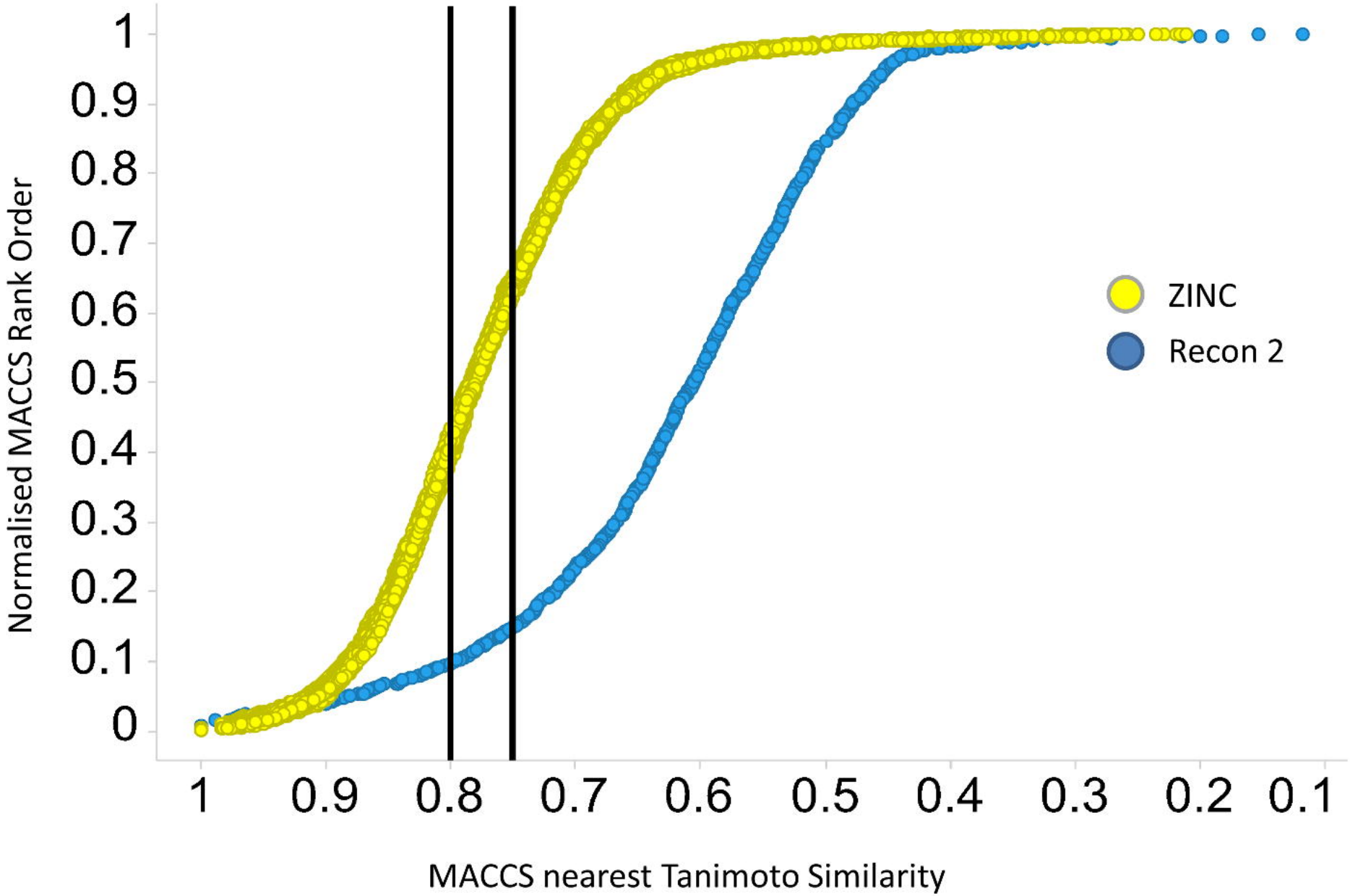
Normalised MACCS rank order vs MACCS_TS sampled

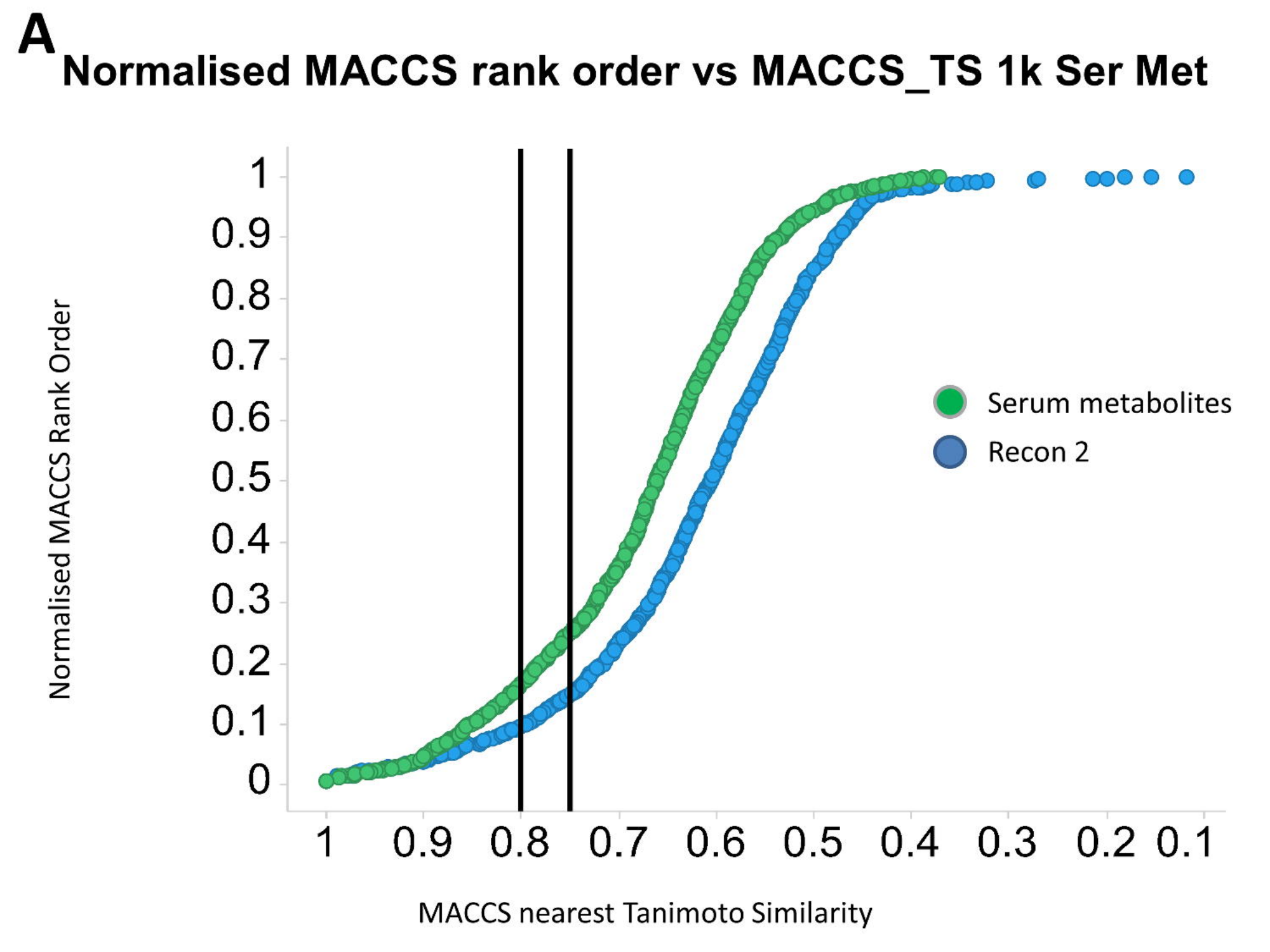


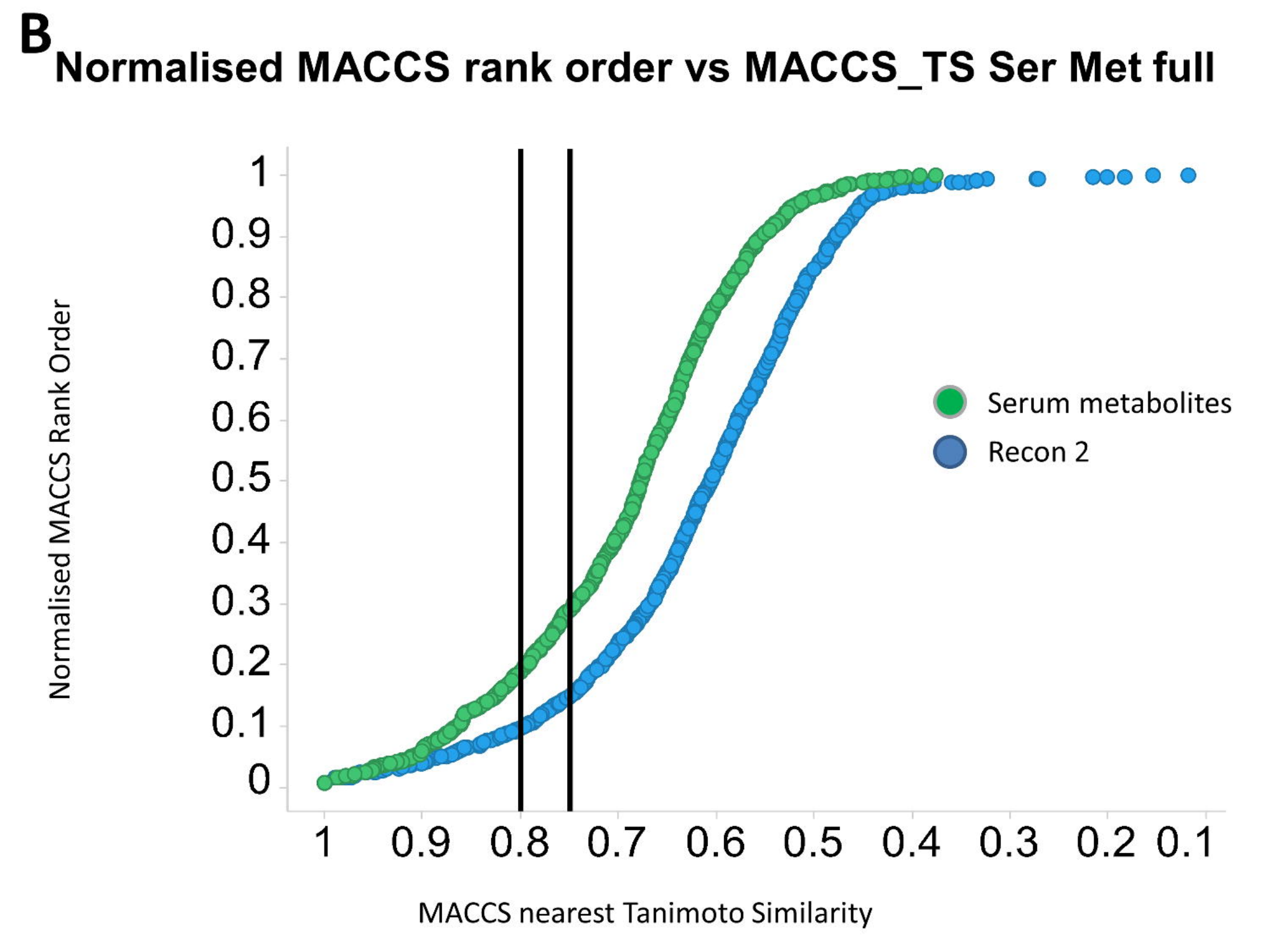
Distributions of molecules in full and cleaned UNPD and DNP databases



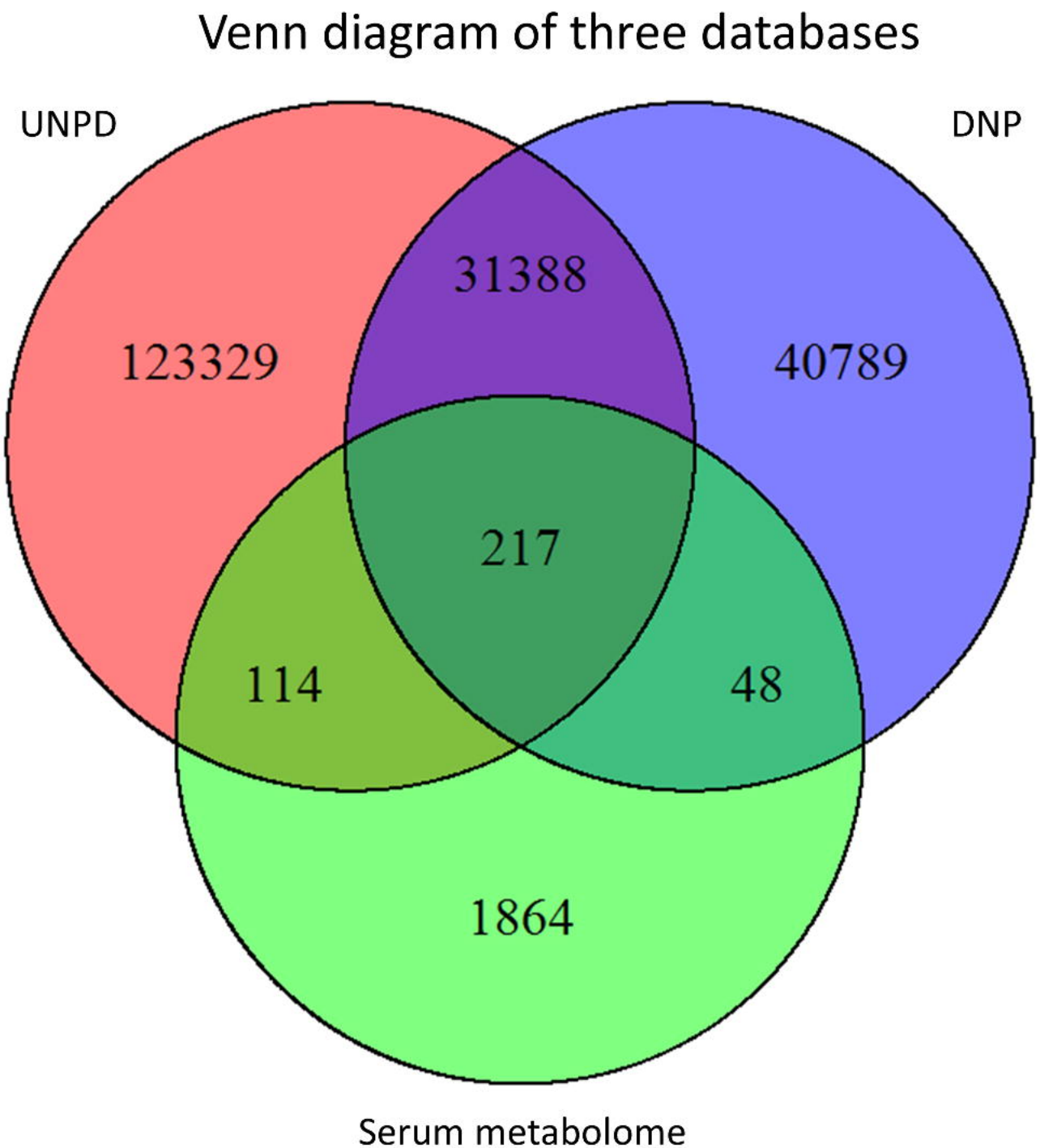
Normalised MACCS rank order vs MACCS_TS sampled ZINC

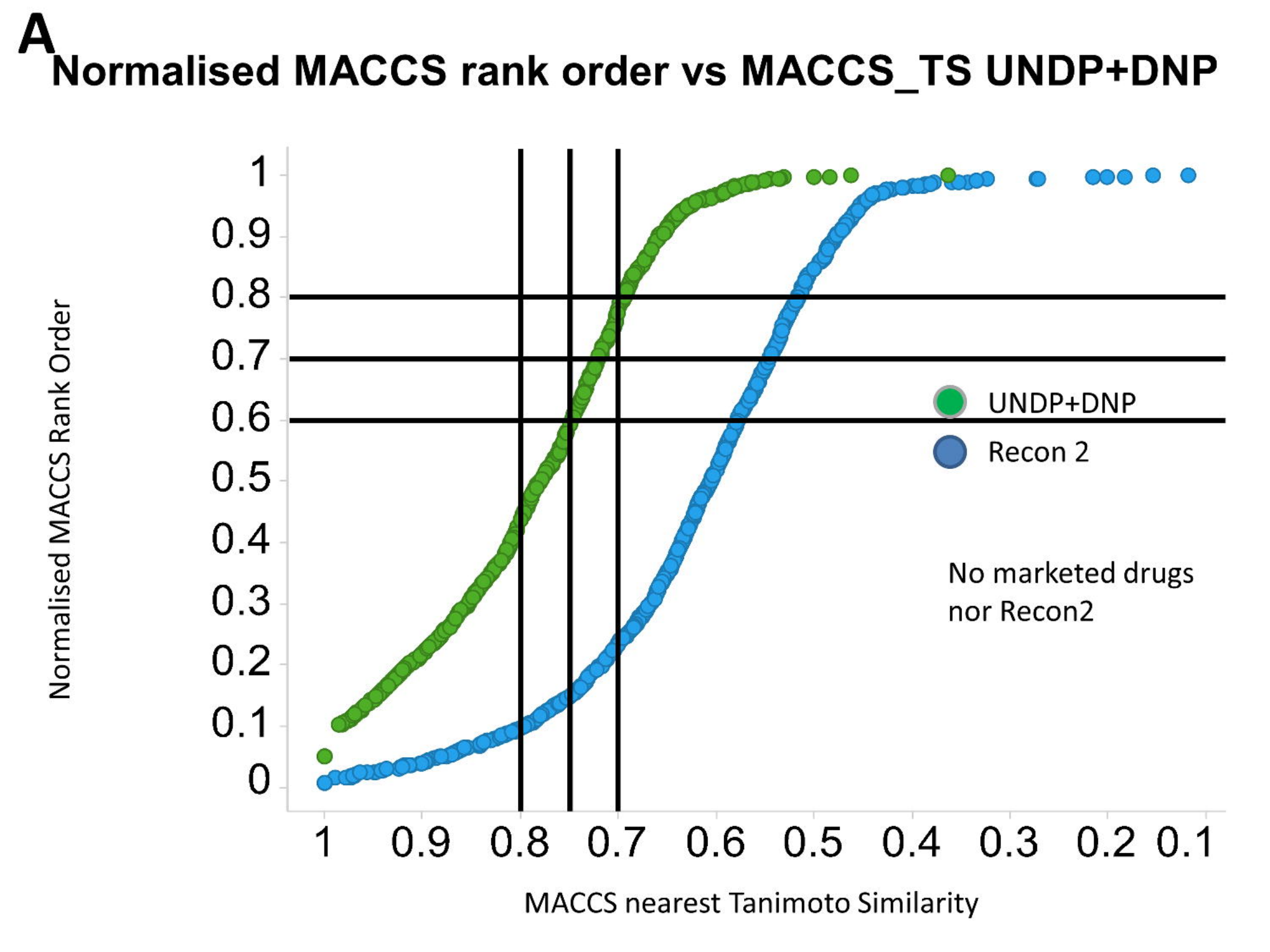


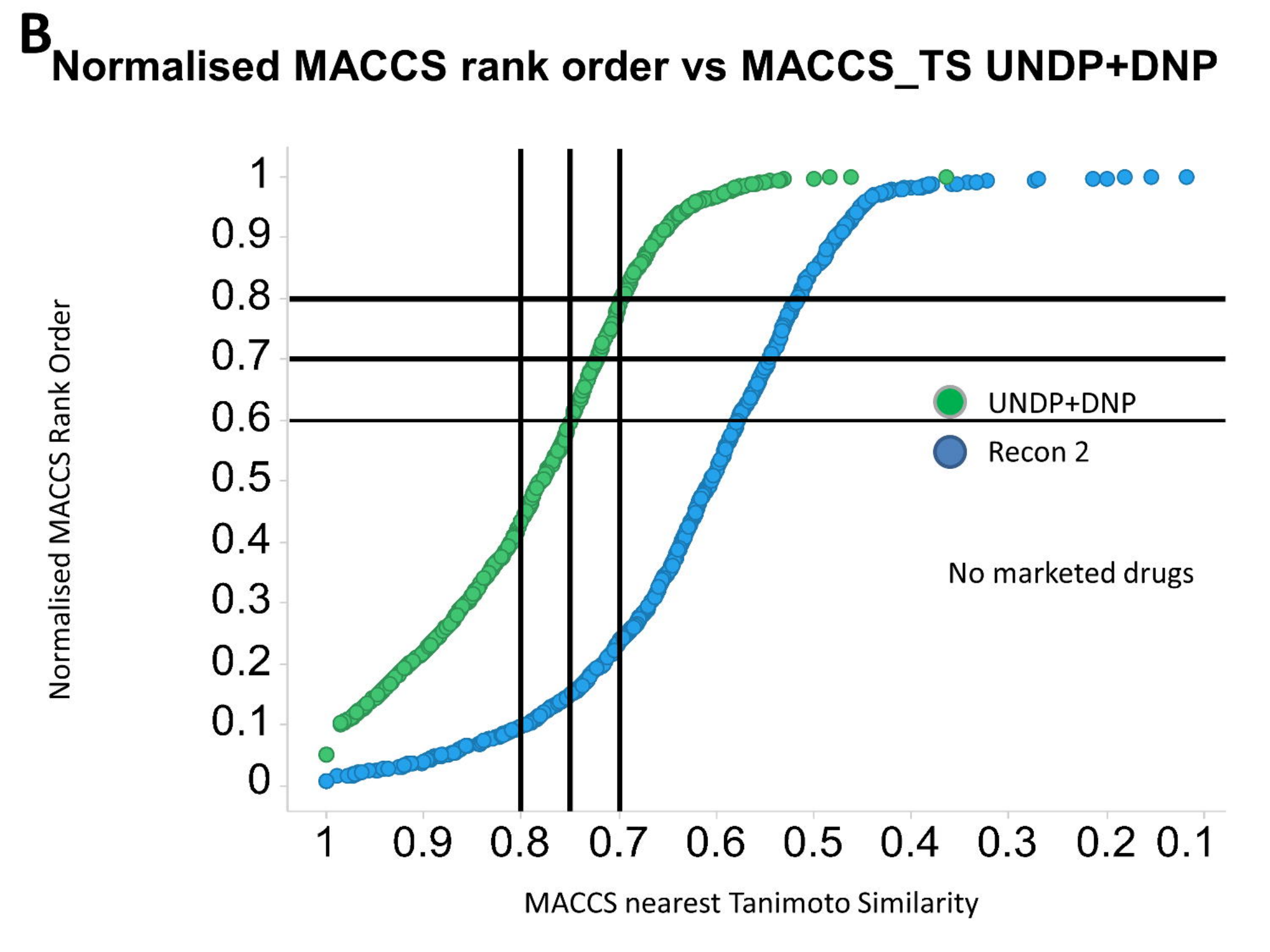

A
Normalised MACCS rank order vs MACCS_TS 1k Ser Met
Normalised MACCS Rank Order
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
MACCS nearest Tanimoto Similarity
Serum metabolites
Recon 2


B
Normalised MACCS rank order vs MACCS_TS Ser Met full
Normalised MACCS Rank Order
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
MACCS nearest Tanimoto Similarity
Serum metabolites
Recon 2

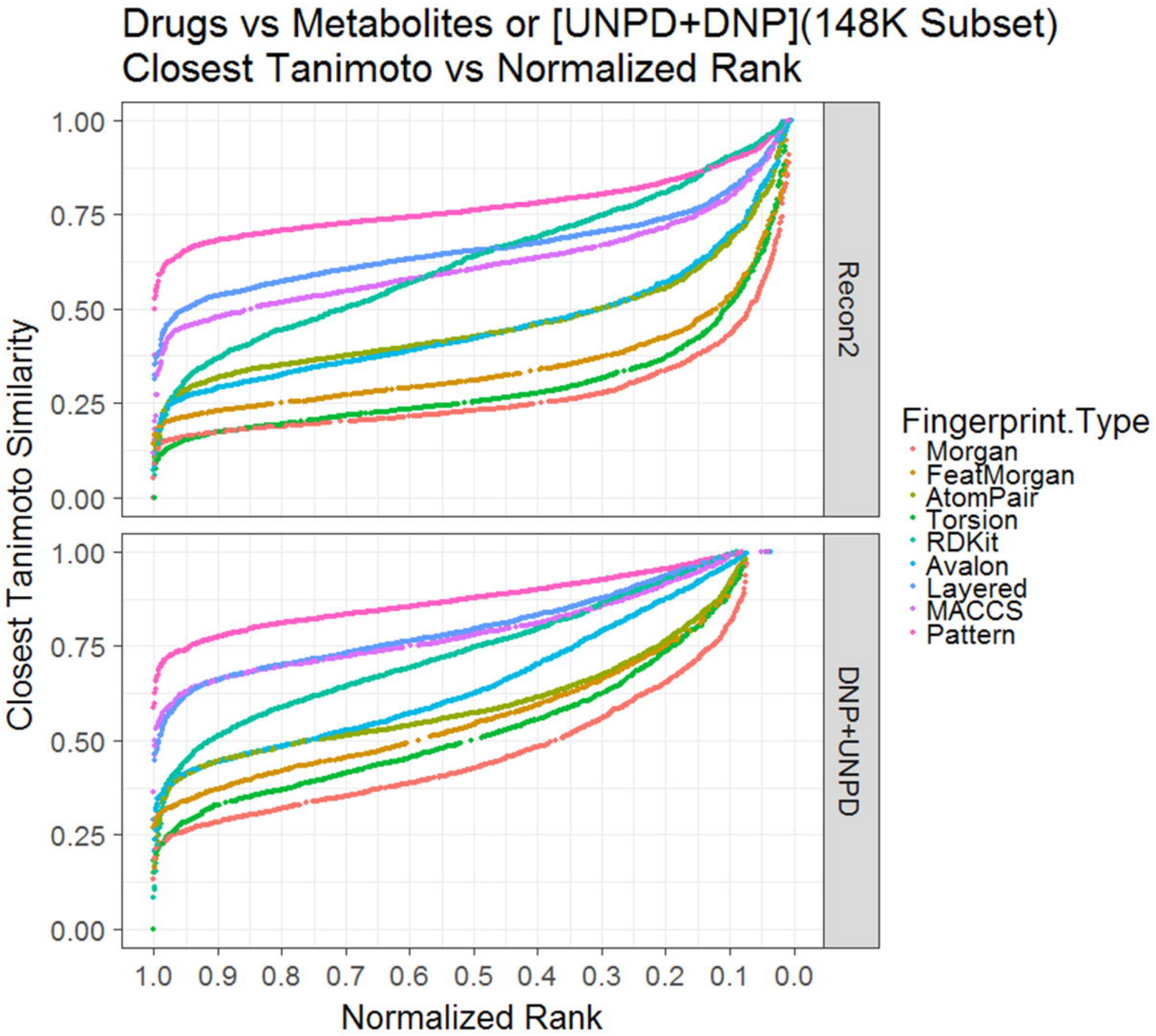
C


Venn diagram of three databases
UNPD
DNP
123329
31388
40789
217
114
48
1864
Serum metabolome


A
Normalised MACCS rank order vs MACCS_TS UNDP+DNP
Normalised MACCS Rank Order
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
UNDP+DNP
Recon 2
No marketed drugs
nor Recon2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
MACCS nearest Tanimoto Similarity

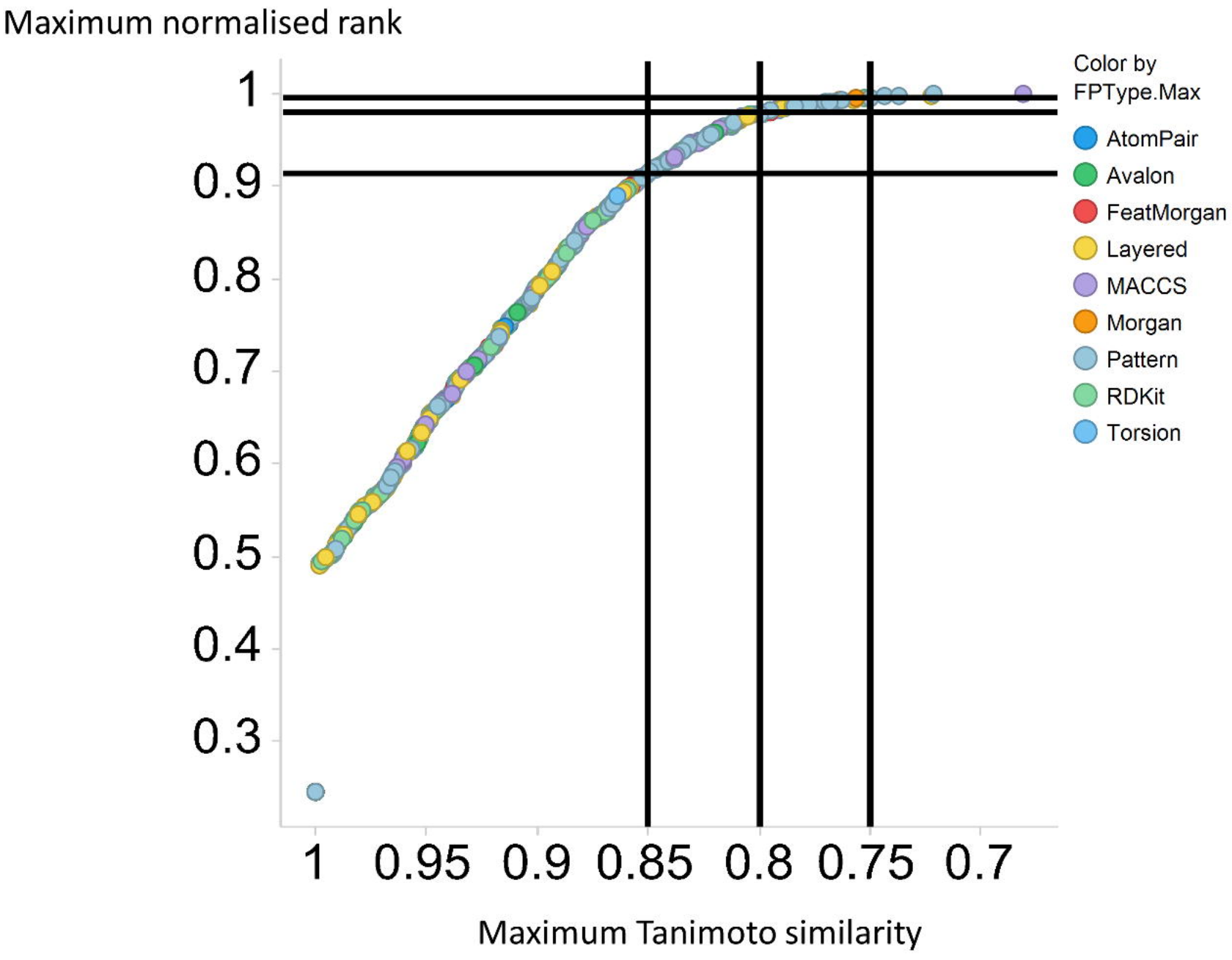

B
Normalised MACCS rank order vs MACCS_TS UNDP+DNP
Normalised MACCS Rank Order
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
UNDP+DNP
Recon 2
No marketed drugs
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
MACCS nearest Tanimoto Similarity

C

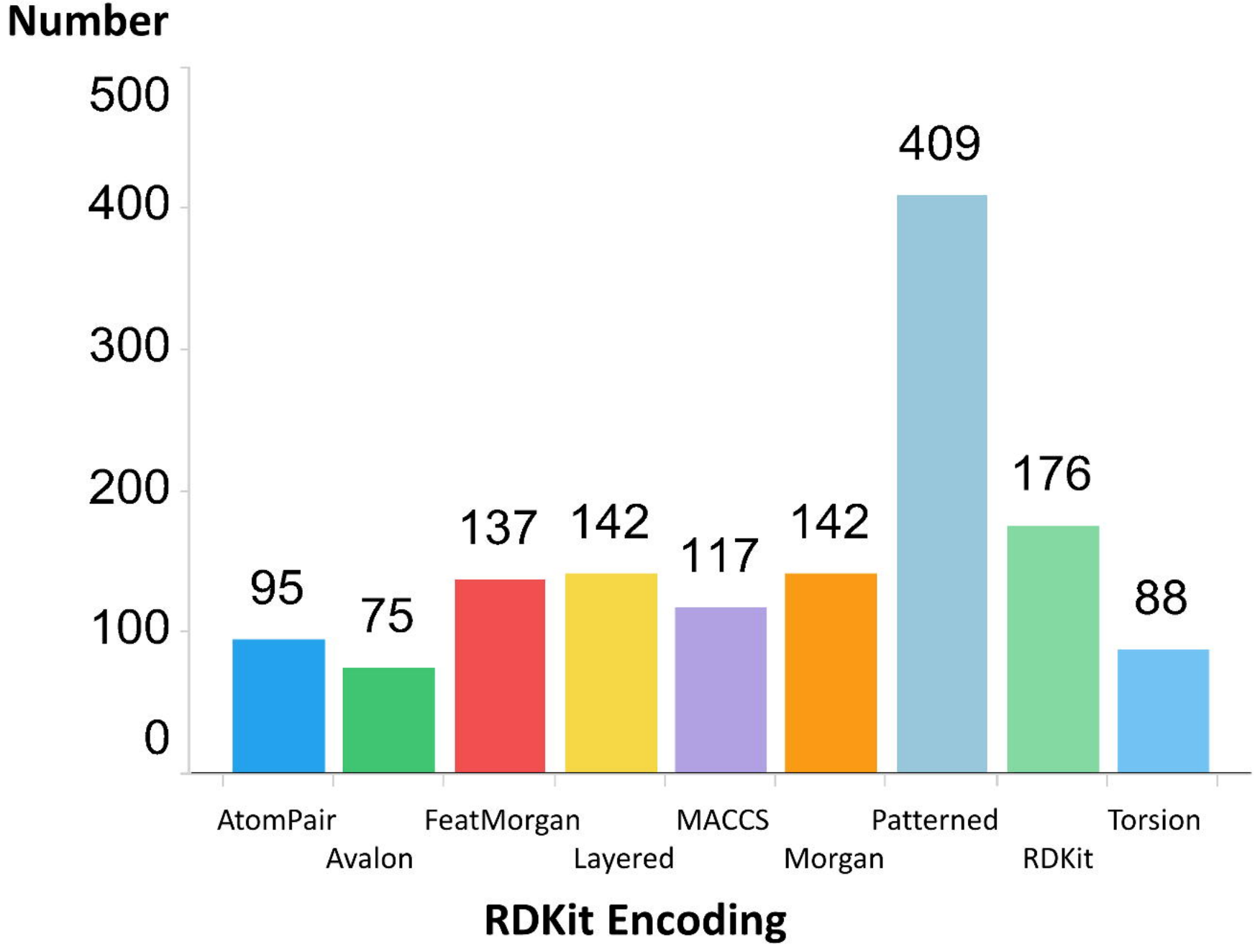

Drugs vs Metabolites or [UNPD+DNP](148K Subset)
Closest Tanimoto vs Normalized Rank
Closest Tanimoto Similarity
1.00
0.75
0.50
0.25
0.00
Recon2
DNP+UNPD
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Normalized Rank
Fingerprint.Type
Morgan
FeatMorgan
AtomPair
Torsion
RDKit
Avalon
Layered
MACCS
Pattern

D

Maximum rankwith TYPICAL encoding, 196k NPs



E Distribution of maximum values of different encodings



F

No relationship between Caco-2 permeability and natural product likeness