O'Hagan and Kell *J Cheminform (2017) 9:18*
DOI 10.1186/s13321-017-0198-y

Journal of Cheminformatics

RESEARCH ARTICLE

Open Access

Analysis of drug–endogenous human metabolite similarities in terms of their maximum common substructures

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

Abstract

In previous work, we have assessed the structural similarities between marketed drugs ('drugs') and endogenous natural human metabolites ('metabolites' or 'endogenites'), using 'fingerprint' methods in common use, and the Tanimoto and Tversky similarity metrics, finding that the fingerprint encoding used had a dramatic effect on the apparent similarities observed. By contrast, the maximal common substructure (MCS), when the means of determining it is fixed, is a means of determining similarities that is largely independent of the fingerprints, and also has a clear chemical meaning. We here explored the utility of the MCS and metrics derived therefrom. In many cases, a shared scaffold helps cluster drugs and endogenites, and gives insight into enzymes (in particular transporters) that they both share. Tanimoto and Tversky similarities based on the MCS tend to be smaller than those based on the MACCS fingerprint-type encoding, though the converse is also true for a significant fraction of the comparisons. While no single molecular descriptor can account for these differences, a machine learning-based analysis of the nature of the differences (MACCS_Tanimoto vs MCS_Tversky) shows that they are indeed deterministic, although the features that are used in the model to account for this vary greatly with each individual drug. The extent of its utility and interpretability vary with the drug of interest, implying that while MCS is neither 'better' nor 'worse' for every drug–endogenite comparison, it is sufficiently different to be of value. The overall conclusion is thus that the use of the MCS provides an additional and valuable strategy for understanding the structural basis for similarities between synthetic, marketed drugs and natural intermediary metabolites.

Keywords: Drug transporters, Cheminformatics, Endogenites, Metabolomics, Encodings, Maximum common substructure

Background

It is becoming increasingly clear that the transmembrane transport of drugs and xenobiotics via any transphospholipid bilayer diffusion is probably negligible, and thus that they have to "hitchhike" on the transporters of intermediary metabolism in order to get into cells [1–19]. Consequently, we [2, 20–22] and others (e.g. [23–27]) have recognised, on the basis of the principle of 'molecular similarity' [28–30], that successful, marketed drugs ought to bear structural similarities to endogenous (intermediary) metabolites (that we shall sometimes call 'endogenites' [2]).

Following an earlier sortie [2], we have used the availability of a carefully curated reconstruction of the human metabolic network, Recon2 [31–33], to answer this question in a straightforward manner. Now 'similarity', as an essentially 'unsupervised' concept, depends on the metrics of similarity used, and arguably is best judged post hoc simply in terms of its utility [29, 34]. Most strategies for assessing the similarities of small molecules use a means of encoding their 2D structures as bitstrings and comparing the similarities of those bitstrings (e.g. [29, 30, 35–41]). Thus, for the drug–endogenite comparison, it was clear that even using the common Jaccard/Tanimoto

*Correspondence: dbk@manchester.ac.uk; http://dbkgroup.org/
[2] Manchester Institute of Biotechnology, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK
Full list of author information is available at the end of the article

similarity metric the rank and magnitude of the similarities could vary widely between different encodings [20].

However, there are many different similarity coefficients even for (binary) bitstrings (Todeschini and colleagues compared 51 [42]), and just using the MACSS166 encoding [43] and the Tversky similarity [44, 45] with different α and β coefficients we again found an enormous variation (both qualitative and quantitative) [22] in the similarities determined between two molecules as α and β were varied. A particular recognition here, however, was the utility of interrogating with just sub-fractions of the molecule that were effectively exploited when α and β (at a constant $\alpha + \beta$) were least similar to each other.

One scoring that is resistant to the detailed encoding used is based on the simple presence or absence of a given substructure, and assessing the frequencies and presence of some 600 common substructures provided a novel and useful metric, even with Tanimoto [21]. Again, however, the magnitude of the similarities determined depended on what fraction of the substructures (ranked in terms of their frequency) were used [21], and this encoding did not directly favour larger substructures over smaller ones.

All of these have been of value in recognising that approved, marketed drugs did share structural similarities with endogenous metabolites. A related question surrounds the "natural" substrates of particular transporters that transport pharmaceutical drugs, but this could not directly be answered from similarity considerations alone.

One structural feature that is largely (but not entirely, e.g. [46]) independent of both the encoding and the similarity used, at least if represented as a 2D graph of linked atom types, is the 'maximum common substructure' between two molecules (variously referred to as the MCS or MCSS). It has achieved especial prominence because of the frequent use of 'scaffolds' in medicinal chemistry, where the scaffold is effectively equivalent to a large, common substructure (e.g. [47–52]). Although its calculation is computationally much more demanding than are many of the other calculations in similarity cheminformatics [46, 53–65] (and see below), this essential independence from both the encoding and the similarity metric means that it is a principled strategy that we considered worth exploring for the drug–metabolite similarity problem. It was not necessarily clear that MCS would be better, but it was recognised that it would provide different information; in particular an MCS is a graph of connected atoms, with a clear chemical meaning, while a fingerprint is essentially uninterpretable without knowledge of how it was generated (and in many cases, e.g. for isomers, it is not unique, whereas an MCS is an MCS). The results of this analysis are given here.

Methods

The list of endogenous metabolites and marketed drugs was precisely as used previously [20–22], and we used the KNIME workflow environment (e.g. [66–72]) to write the appropriate workflows for these analyses. In particular, we used the RDKit [73] (http://rdkit.org/) MCS nodes for the MCS calculations. To provide a metric for the MCS, we followed the recent analyses of Bajorath and colleagues [65, 74, 75]. Thus they recognised that a similarity equivalent to the Tanimoto similarity for a molecule A with a total of $|A|_b$ heavy atoms and another molecule B with $|B|_b$ heavy atoms, could be written in the form [74]

$$\mathrm{Tc}_{\mathrm{MCS}}(A,B) = \frac{|\mathrm{MCS}(A,B)|_b}{|A|_b + |B|_b - |\mathrm{MCS}(A,B)|_b} \tag{1}$$

where $|\mathrm{MCS(A, B)}|_b$ is the number of heavy atoms in the MCS. Elementary inspection of Eq. (1) shows that the value of the $\mathrm{Tc}_{\mathrm{MCS}}$ does, as expected, range between 0 and 1.

The Tversky similarity coefficient Tv(A, B) coefficient [44, 76–78] is defined as:

$$\mathrm{Tv(A,B)} = \mathrm{c}/(\alpha\mathrm{a} + \beta\mathrm{b} + \mathrm{c}), \tag{2}$$

where a and b are the number of bits that are set to be 'on' (1 bits) only in molecular fingerprints A or B, respectively, and c is the number of on bits shared by both A and B. A is an interrogatory molecule while B is the molecule being interrogated as to its similarity. The smaller the value of α, the larger the contribution of B as a substructure of A (and hence to its similarity with A). The larger the value of α, the larger the contribution of B as a superstructure of A (equivalently A as a substructure of B). For $\alpha = \beta = 1$ the coefficient is numerically equivalent to the Tanimoto similarity.

A similar strategy could be followed [65, 75] (Eq. 3) to report a Tversky similarity as per Eq. 2, with α and β having their usual meanings as in the previous paragraph [22, 44, 76–78]. As before, we studied the effect of varying α while the sum of α and β was either 1 or 2.

$$\mathrm{Tv}_{\mathrm{MCS}}(A,B,\alpha,\beta) = \frac{|\mathrm{MCS}(A,B)|_b}{\alpha(|A|_b - |\mathrm{MCS}(A,B)|_b) + \beta(|B|_b - |\mathrm{MCS}(A,B)|_b) + |\mathrm{MCS}(A,B)|_b}, \quad \alpha,\beta \geq 0 \tag{3}$$

Specifically, the MCS algorithm used in this study was the fast connected MCS algorithm fMCS, as implemented in RDKit (see http://www.dalkescientific.com/writings/diary/archive/2012/05/13/fmcs.html and http://rdkit.org/Python_Docs/rdkit.Chem.fmcs.fmcs%27-pysrc.html). We used Python 2.7 + the Python RDKit package to generate [for all A and B's] the MCS SMARTS string, the a,b, and MCS Atom counts; as well as the Tanimoto-like MCSS.

Results

One drug versus all drugs plus endogenites

In our previous work [20], where we clustered marketed drugs on the basis of their chemical structures, this was simply a prelude to comparing them with metabolites but we did not dig down into the clusters so formed at any level of detail. Here, it was of initial interest to establish whether the MCS strategy did indeed return as most similar drugs containing a particular scaffold. To this end, we chose diazepam, as an example of a 'first generation' antipsychotic. As expected, it showed a shared pedigree with other related benzodiazepine molecules (Fig. 1). Such molecules were less similar to 'second generation' molecules such as clozapine and olanzapine [79–81] that are themselves part of a (large) family of such molecules with a complex pharmacological profile [82]. Figure 1a shows the various molecules as a function of the number of heavy atoms in the MCS when whole (aromatic) rings must be present in the MCS. Only 23 molecules have 9 or more heavy atoms in the MCS (Fig. 1a). All are well known antipsychotic drugs. The metabolites with the largest MCS (6 heavy atoms) are salsoline and salsolinol

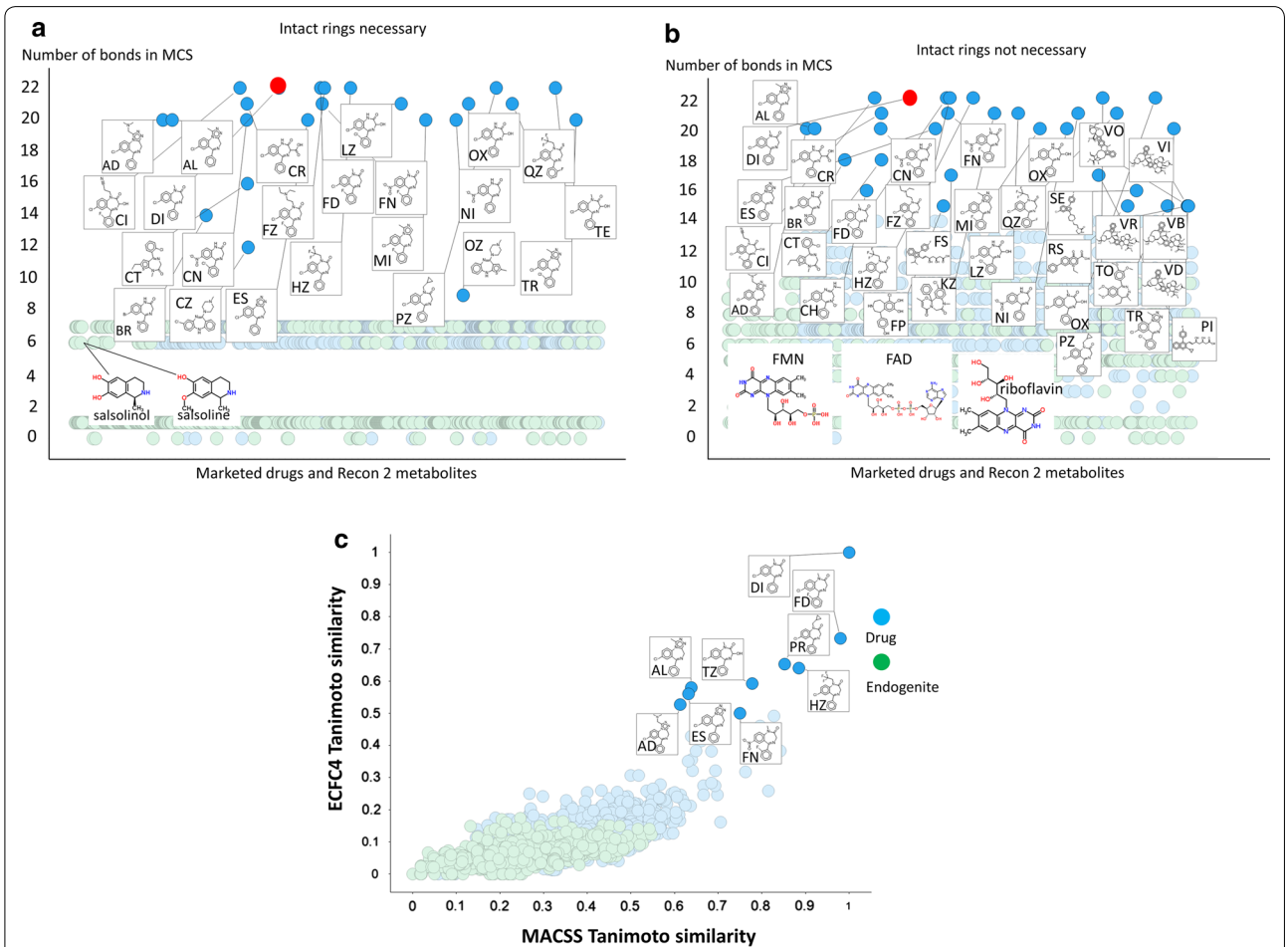

Fig. 1 Maximal common substructure (MCS) between diazepam (in *red*) and other marketed drugs and metabolites. The size of the MCS is plotted for various drugs (*blue*) and endogenous metabolites (*green*). A KNIME workflow was constructed, including using the RDKit MCS module and interrogated with the structure of diazepam. **a** Distribution of MCS values when the RDKit MCS was set to use only intact rings. **b** The same without that restriction. In both cases, the structures of the closest molecules are shown. **c** A comparison of the Tanimoto similarity of diazepam and other drugs and endogenites using two common fingerprint encodings (ECFP4 and MACCS). The structures of those exceeding 0.5 in each encoding are shown.

(which is not unreasonable, as they are condensation products of dopamine and acetaldehyde [83–87]). When this 'whole-ring' assumption is relaxed (Fig. 1b), a somewhat different pattern emerges, though we mark only those molecules with at least 16 heavy atoms in the MCS. Now the closest three metabolites (FAD, FMN and riboflavin) have 11 heavy atoms in the MCS, and while this strategy retains the main molecules of the 'rings-only' strategy, it now lets in molecules such as 'statins' (fluvastatin, pitastatin), anticancer *Vinca* alkaloids (vinblastine, vincristine, vindesine), and quinolone antibiotics (rosoxacin) whose basic scaffold is really nothing like that of a benzodiazepine. Note that Fig. 1 consists in total of 1112 metabolites and 1381 marketed drugs, making 2493 marketed drugs plus endogenous metabolites in toto. All 23 diazepams cluster together, and their lowest TS to diazepam when the encoding is the MCS is 0.667. By contrast, many more substances appear similar when some of the classical fingerprints are used. Figure 1c shows the Tanimoto similarities for diazepam versus all drugs (blue) and endogenites (green) for two RDKit encodings (MACCS and ECFP4), where 175 molecules have a MACCS-TS > 0.5, though only 9 molecules show similarities above 0.5 for both encodings. (The closest metabolites, which also do, are methylene tetrahydrofolate and vitamin D_2.) The simplest interpretation is really that the MCS is much more discriminating for what it says, i.e. the maximum common substructure or scaffold, but that this leads to a more natural and useful clustering. Finally, here, Fig. 2 and Additional file 1 shows the workflow used for Fig. 1a, b, and illustrates how we indicated the MCS in the Excel sheet to which the analyses were output. Thus we preferred the MCS that required that if rings were present they had to be present in their entirety in both molecules to contribute to the MCS.

MCS of all drugs and/or metabolites against each other

While this was considerably more demanding in computer time than our previous similarity analyses based on various fingerprints coupled to Tanimoto or Tversky similarity [20–22, 88], it proved possible and useful to do. A run of all drugs against all metabolites took approximately 3 days on a reasonably modern PC (Intel i7-4930K, 6 cores hyperthreaded cpu (12 virtual cores) @ 3.4 GHz, 64 GB Ram). We here used MACCS166 as the 'main' fingerprint. Others such as ECFP (and FCFP etc.) were not done since (1) comparison of MCS versus all possible fingerprints would have been completely unwieldy, and (2) we had compared the fingerprints with each other in our previous papers. Since MACCS gave among the largest similarities [20], we also considered that it would provide the sternest 'test' of the utility of MCS. Figure 3 shows heat maps for the three comparisons (endogenites–endogenites, drugs–drugs, drugs–endogenites), analogous to those performed [20] using molecular encodings. Relevant Excel sheets are given in the Additional files 3, 4, 5 to allow readers to explore further, but these are very rich in information. Thus, although (Fig. 1a) they tend to give more 'sensible' hits where scaffolds exist, numerically they only attain large Tanimoto similarities for rather similar drug or endogenite classes. These classes may be seen as blue clusters in Fig. 3, some of which are marked therein. As before, there are larger endogenite clusters, where CoA derivatives (bottom left of Fig. 3a) and sterols (bluest cluster nearer the middle) again clearly dominate, in contrast to the much 'bittier' population of drug space (Fig. 3b). The largest clusters of similarity of drugs versus endogenites (Fig. 3c) are again sterols (largest blue cluster, towards the top left), with others (marked in Fig. 3c) including amphetamines (similar to various neurotransmitters such as (nor)adrenaline), and nucleosides.

While the calculation of the MCS values was quite demanding, the calculation of other similarities (see "Methods" section) was much simpler, as those used depended only on the number of heavy atoms in the molecules being compared and those in their MCS. Since the Tversky similarity metric had proven (at some values of α and β) to be much more appropriate than Tanimoto for highlighting drug–endogenite similarities, we again used it. Comparing drugs (interrogating molecule) versus endogenites (interrogated library) it is clear (Fig. 4a) that for values of α such as 0.2 (when $\alpha + \beta = 1$) the Tversky similarity of at least one endogenite for virtually every drug exceeds 0.5 when using the MCS as the encoding, whereas this is much less true from when the Tanimoto similarity ($\alpha = \beta = 1$) is used (Fig. 4a). The same is true for the converse [where the interrogating molecule is an endogenite (Fig. 4b)].

Some examples

It seems that the MCS method of molecular comparison, when all rings are included intact, gives much more reliable measurements of useful similarity as judged by scaffolds. As ever, the different metrics give different indications of how similar two molecules seem to be. To this end, we interrogated the endogenites with a few drugs carefully chosen to illustrate the kinds of variation observable, first illustrating their differences with (1) an MCS-based similarity with Tversky α 0.2 and β 0.8 and (2) a MACCS encoding and a Tanimoto similarity as in [20].

Figure 5a shows the very small and hydrophilic metformin (MW 129.17), and how the MCS/Tversky encoding shows it to be much more metabolite-like than does the MACCS_Tanimoto analysis. Partly this is because its small size means that many bits are set low and so the

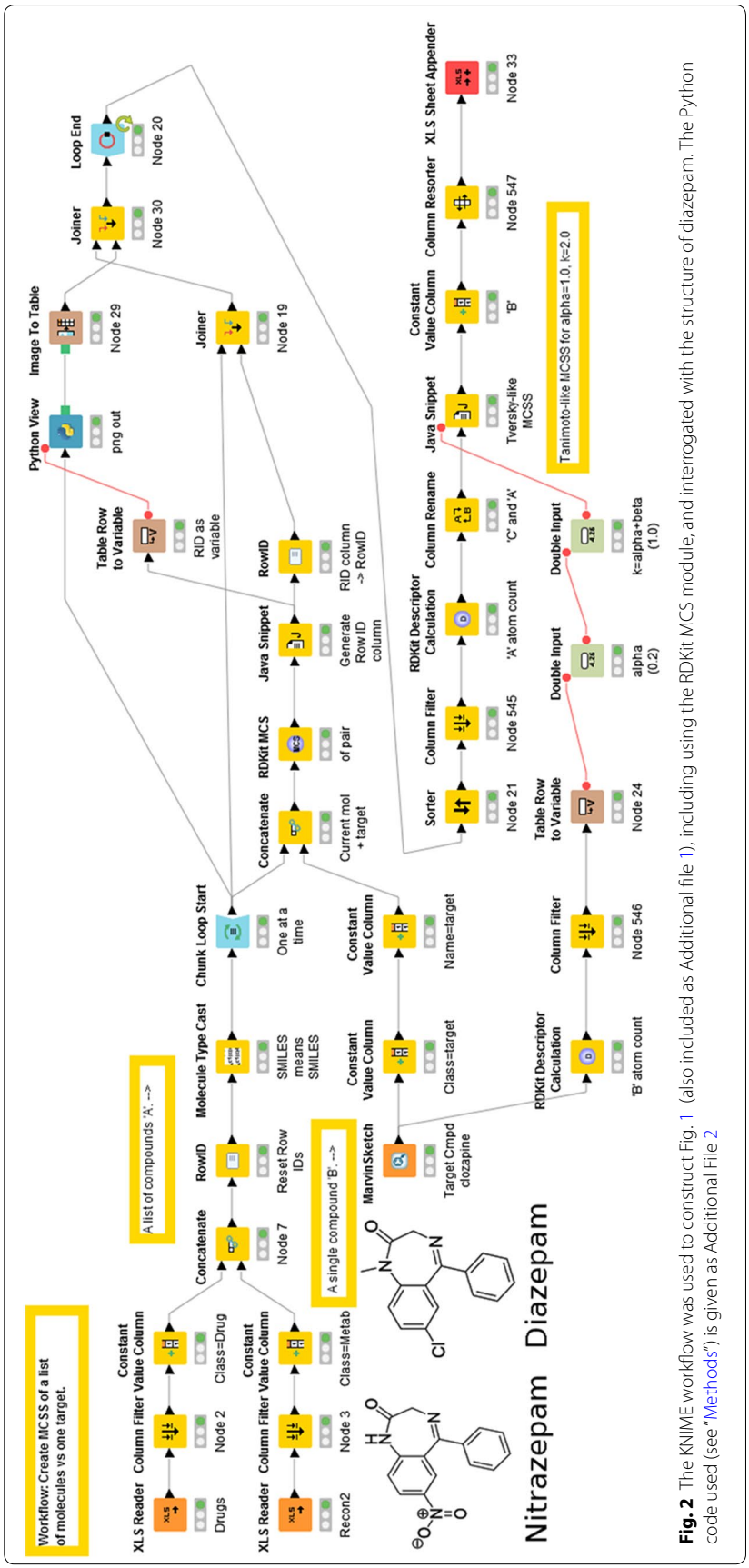

Fig. 2 The KNIME workflow was used to construct Fig. 1 (also included as Additional file 1), including using the RDKit MCS module, and interrogated with the structure of diazepam. The Python code used (see "Methods") is given as Additional File 2

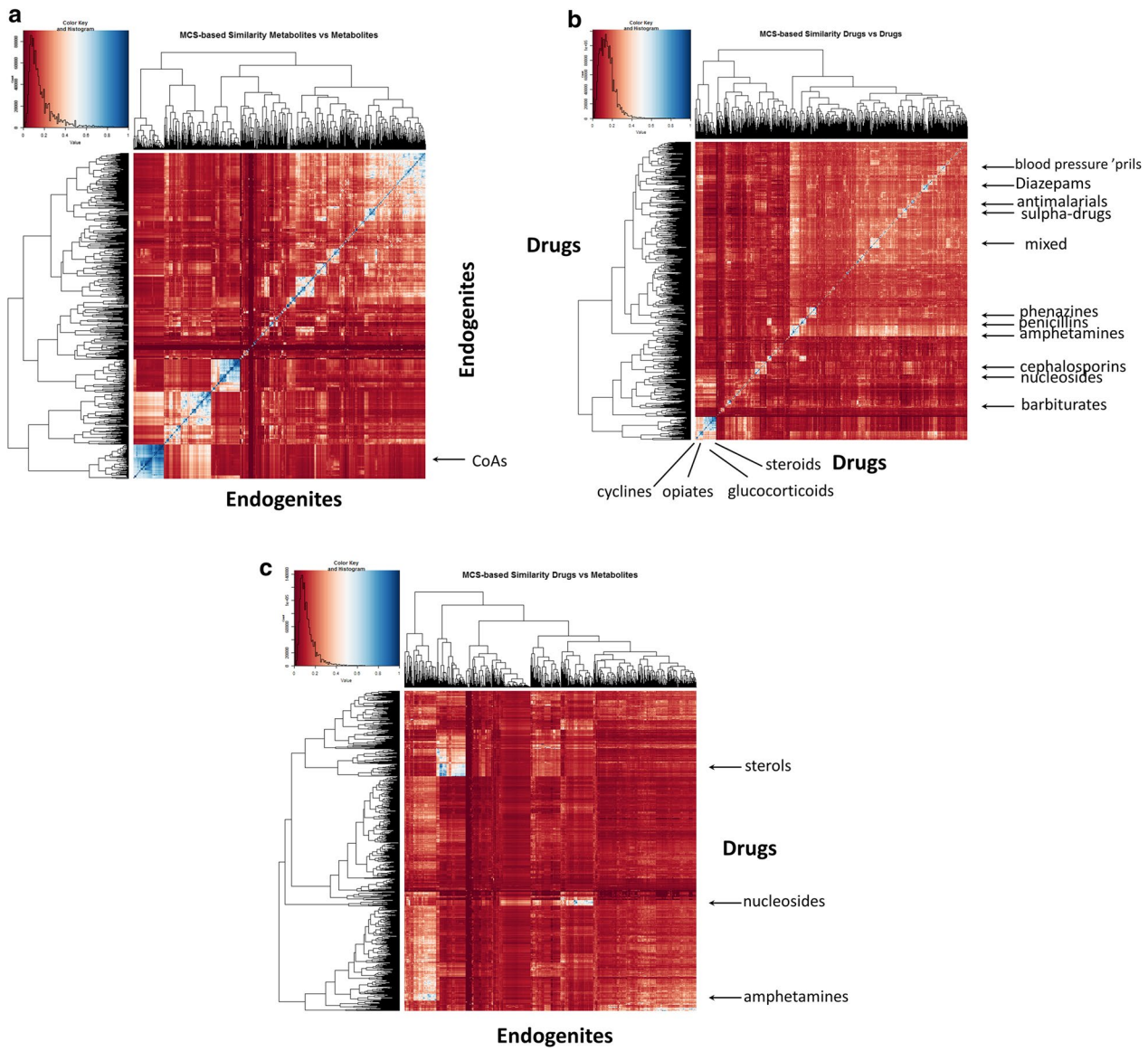

Fig. 3 Heat map of the comparison of the Tanimoto similarities of the MCS for **a** endogenites versus endogenites, **b** drugs versus drugs, **c** drugs versus endogenites

TS is low (see [22, 89–91]). Nevertheless, its structural similarity to creatine (most similar via the Tversky metric) and other organic cations is consistent with the fact that it is taken up by SLC22 family members (known as Organic Cation Transporters in the older literature [92–99]). Benzylpenicillin (334.39) illustrates a couple of interesting features (Fig. 5b). First is that among the drugs (in blue) it clusters most closely with the penicillins and then with the cepahlosporins, as expected. Secondly, the metabolites to which it is most similar include several N-substituted kynurenine derivatives, consistent with an anticipation that at least some of them might share a similar transporter. This is in fact the case (SLC15 family, e.g. [100–104]). Pravastatin (MW 424.53) is one of the so-called 'statin' class of drug that can inhibit HMGCoA reductase. As is clear from Fig. 6a, apart from the related natural products simvastatin and lovastatin, it does not show any obvious similarity or major MCS to any other so-called statin (e.g. atorvastatin (Lipitor) or rosuvatstain (Crestor)), even though they all share a glutarate or related lactone group. Arguably this reflects the fact that much of their activity is in fact due to

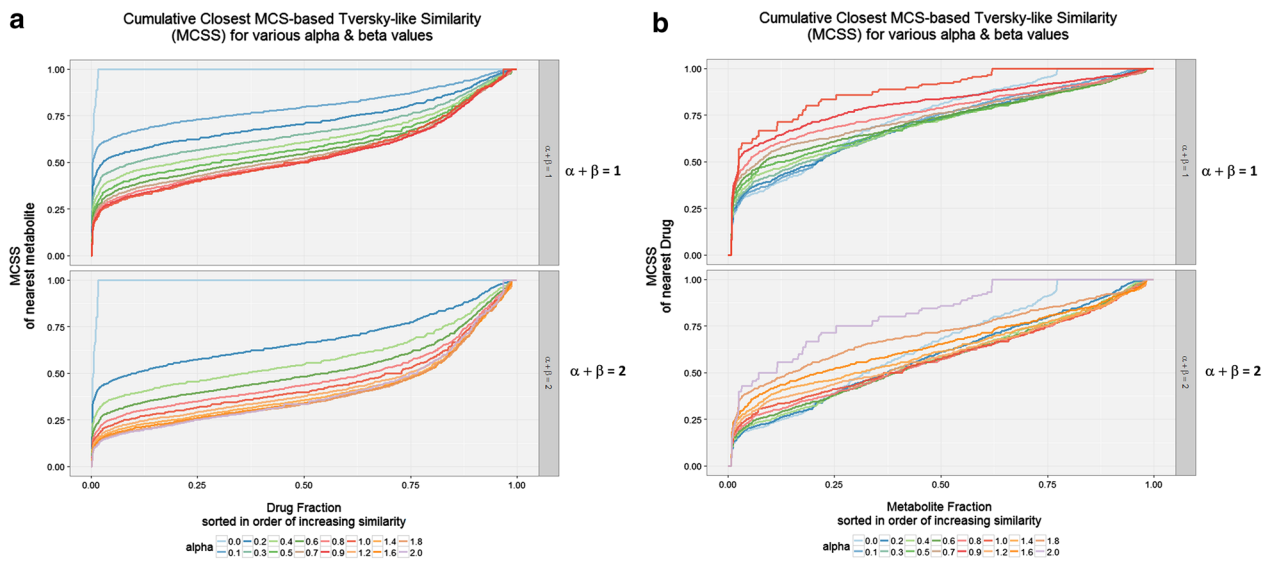

Fig. 4 Cumulative Tversky similarities for various values of Tversky α and β of **a** a drug to its closest endogenite, **b** an endogenite to its closest drug

interactions (of the other parts of the molecule) with other targets (e.g. [105–119]), and expression profiling demonstrates clearly [120] that they lack a unitary mode of action. Consequently it is less surprising that MCS performs poorly in this regard, since they really do not have much of a common substructure. Verapamil (MW 454.6) is a Ca^{++}-channel blocker with multiple disease indications (implying considerable promiscuity, consistent with a log P value of 3.79 http://www.drugbank.ca/drugs/DB00661). It is also considered one of the more rapidly transported drugs in Caco-2 cells (e.g. [14, 15]). According to ChEMBL https://www.ebi.ac.uk/chembldb/index.php/compound/inspect/CHEMBL6966, it interacts with some 172 targets, including 11 uptake transporters, which presumably accounts for this. The central core, consisting of a long, branched and predominantly carbon-based linker, and the heterogeneous nature of the molecules to which it is 'similar' (Fig. 6b), would also be consistent with this.

Propranolol (Fig. 7a) (MW 259.15), another drug enjoying a high rate of transport through Caco-2 cells [14, 15], is a classical β-adrenergic receptor blocker. Unsurprisingly, the analysis pulls out many analogues both as drugs and (for metabolites) among analogues of (nor)adrenaline (synonym (nor)epinephrine) such as metanephine. As judged by the data deposited in ChEMBL https://www.ebi.ac.uk/chembldb/index.php/compound/inspect/CHEMBL27 it has 166 known targets, including 9 uptake transporters. Its structural similarity to noradrenaline means that unsurprisingly these include the very active serotonin, dopamine and noradrenaline transporters. Finally, we show a drug that is among the least obviously metabolite-like, viz. clozapine (Fig. 7b), and also rather hydrophobic; only two endogenites have a Tanimoto similarity exceeding 0.5, though its similarity to related drugs is indeed reasonably high. (The same phenomena attach to sepantronium bromide, a potent drug candidate for which significantly more than 99% of uptake flux into cells occurs via a single transporter (SLC35F2) [11], and for which any phospholipid bilayer transport is consequently negligible [10, 13, 17, 121]; data not shown.)

Although the data are implicit in Figs. 5, 6, 7, it is worthwhile (Table 1) just tabulating the number of molecules for which the difference in the encodings (MACCS_TS−MCS_Tv) is positive and negative for the six molecules, as this makes it clear how much they can differ in either direction.

Accounting for differences in the similarity metrics

Even just with these six drug molecules, it is clear that the degree of similarity with endogenites varies both qualitatively and quantitatively depending on what is the drug and what is the encoding and similarity metric. To this end, we have determined the differences in the similarity between these drugs and endogenites for each endogenite, and sought to understand what in structural or descriptor terms might account for it (in the way that we know that low numbers of bits in the bitstring, as occurs more for smaller molecules, necessarily makes the MACCS

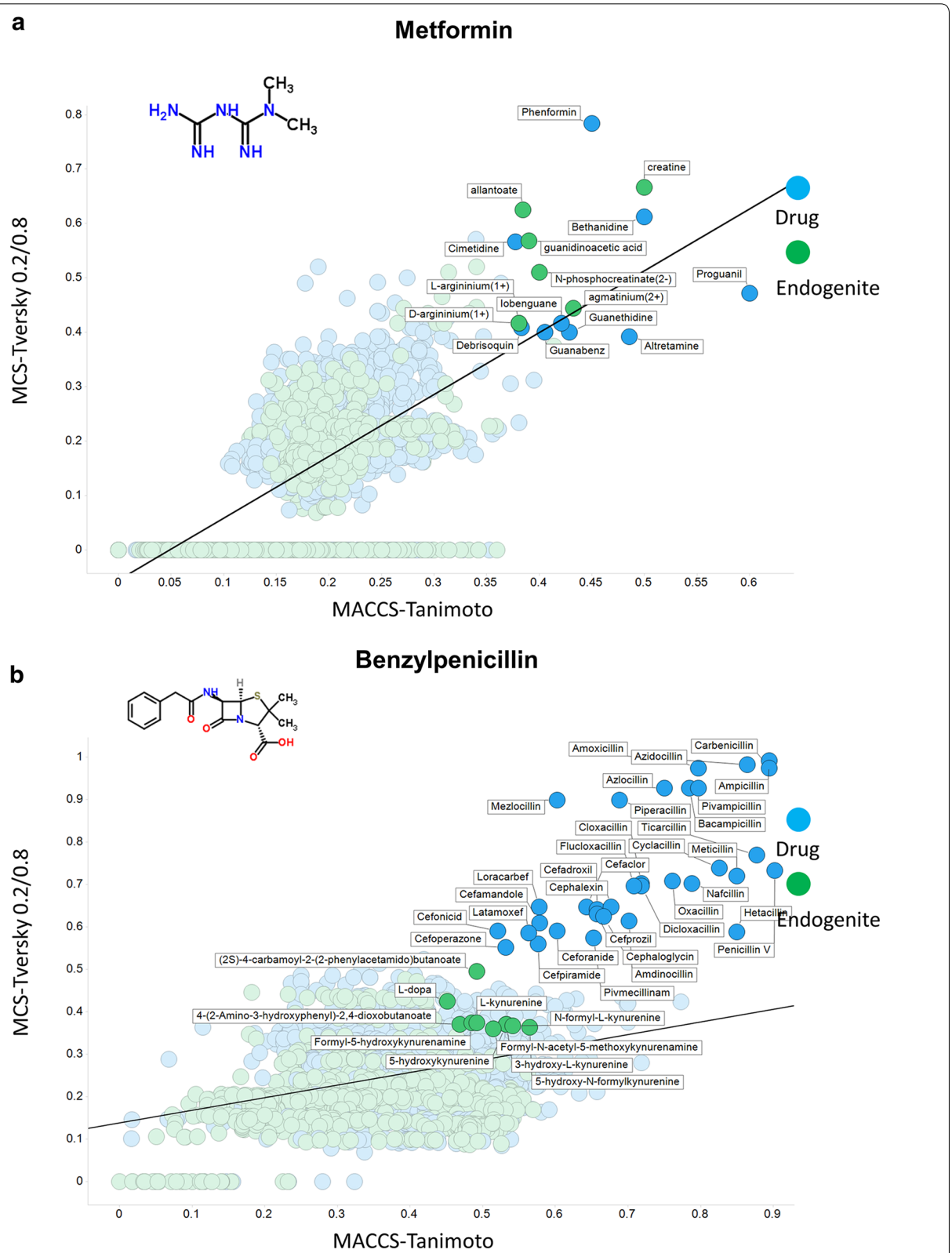

Fig. 5 Relationship between MCS encoded as a Tversky similarity (α, β = 0.2, 0.8) and MACCS-encoded Tanimoto similarity from selected drugs with other marketed drugs (*blue*) and endogenous metabolites (*green*), highlighted at an arbitrary 'break' for each class and where the numbers involved were small enough to permit legibility. The *straight lines* are those of best fit. **a** Metformin. **b** Benzylpenicillin

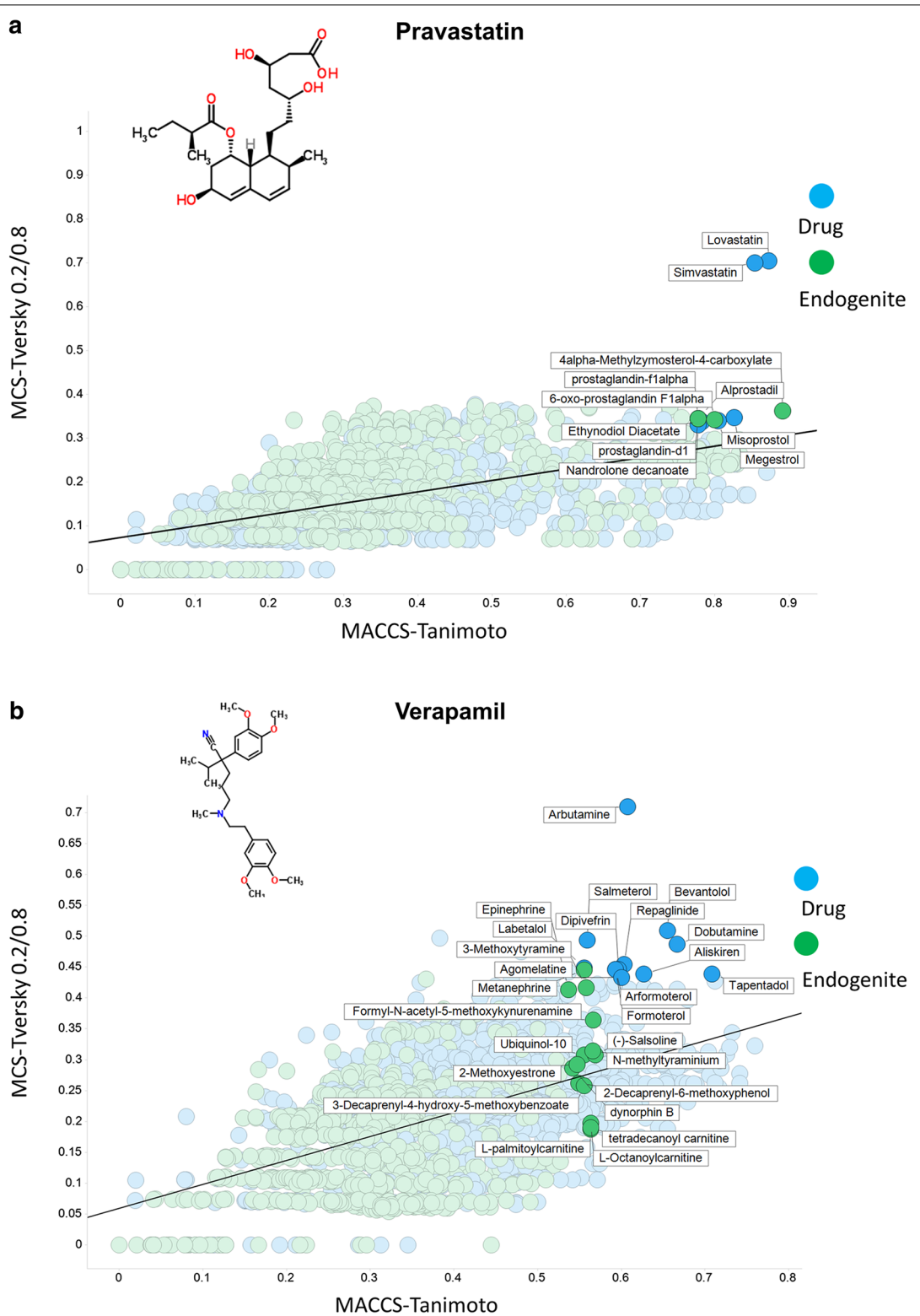

Fig. 6 Relationship between MCS encoded as a Tversky similarity (α, β = 0.2, 0.8) and MACCS-encoded Tanimoto similarity from selected drugs with other marketed drugs (*blue*) and endogenous metabolites (*green*), highlighted at an arbitrary 'break' for each class and where the numbers involved were small enough to permit legibility. The *straight lines* are those of best fit. **a** Pravastatin. **b** Verapamil

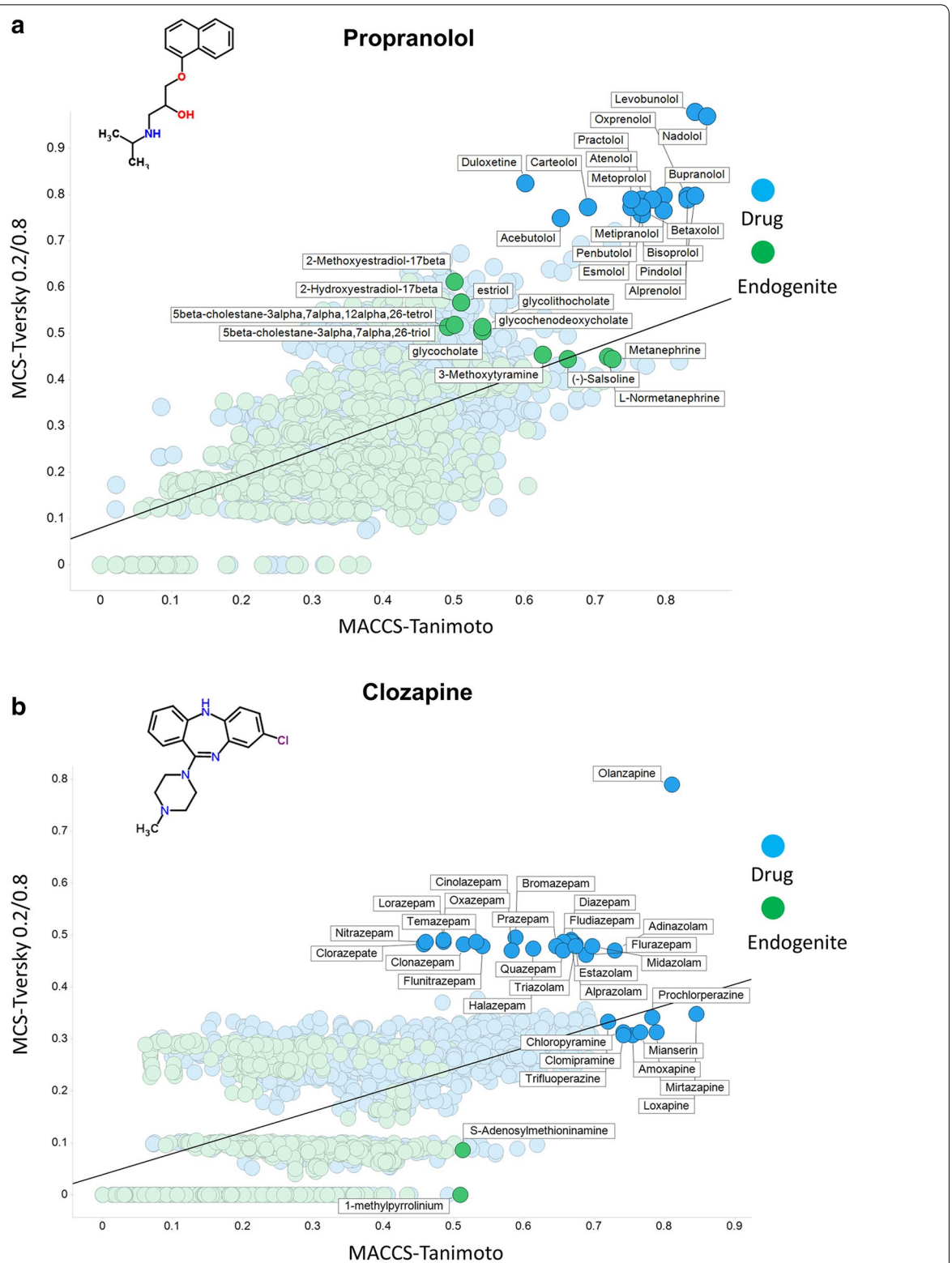

Fig. 7 Relationship between MCS encoded as a Tversky similarity (α, β = 0.2, 0.8) and MACCS-encoded Tanimoto similarity from selected drugs with other marketed drugs (*blue*) and endogenous metabolites (*green*), highlighted at an arbitrary 'break' for each class and where the numbers involved were small enough to permit legibility. The *straight lines* are those of best fit. **a** Propranolol. **b** Clozapine

Tanimoto similarity appear smaller [21, 36, 77, 89, 122–126]). To this end, we set up the following strategy:

Read Drugs + Recon2—the 'A' molecules. Then select the six named 'B' molecules, as in Figs. 5, 6, 7 and Table 1. Loop over each 'B'. For each 'A' paired with a 'B' calculate the MACCS-TS & Tversky-like MCS (alpha = 0.2, beta = 0.8), and their difference Delta. Calculate all available scalar (non-vector) RDKit descriptors of each 'A'—these are the input features of the model. Remove any constant features (there were none). Remove one of each pair of correlated features ($r \geq 0.98$); 13 feature columns removed. Split into 70:30 train:test set. Use a Random Forest regression model (200 trees; see [127, 128]) to predict delta as the objective function. Collect the Out-of-box and Test predictions for each molecule 'B'. Plot a Scatter plot of Actual versus Predicted for each 'B' on the test predictions [127].

Although trends varied for each of the 6 drugs in Figs. 5, 6, 7, no individual descriptor such as S log P could, on its own, account for the differences between MACCS_Tanimoto and MCS_Tversky. However, a random forest model could do so when out-of-bag tests were done, with the predictions and contributions of the descriptors given for the six drugs in Fig. 8. It is clear (1) that the differences are deterministic (Fig. 8a), but (2) that the basis for them, i.e. the features that contribute to those differences, is bespoke to each drug (Fig. 8b). The same was true of 10 other drugs selected at random (data not shown).

Discussion

It is clear that, even when using MCS and Tversky similarities where most drugs do manifest a reasonable similarity to at least one endogenite, the closeness of that similarity can be quite variable. If the effectiveness of drugs is indeed related to their ability to interact with binding sites of proteins, including transporters, that also interact with natural metabolites, this bears some explanation. One straightforward explanation, of course, is simply that we still have to discover many of the naturally occurring metabolites, and that the excellent Recon2—based on metabolic enzymes that are encoded by the genome sequence plus a few vitamins—is useful only insofar as it knows about them. Several general kinds of argument imply that this may indeed be the case. The first is that we can detect many more small molecules as mass spectral signals in biological samples than we can presently identify [129], possibly as a result of unknown enzyme promiscuity [130–132]. Similarly, from the point of view of metabolic network reconstructions, the latest version of Recon2, Recon2.2 [33], contains 2652 unique chemical species, some 60% more than in Recon1 [31, 133], implying that we are far from discovering them all, and some are known still to be absent [9]. Thirdly, many of the metabolites may not be entirely the result of the host's biosynthesis, being derived from dietary sources [134, 135] and including biotransformations in the gut. At an elementary level this is clearly true, since essential amino acids, fatty acids and vitamins are (by definition) not synthesised by the host. However, as known elements of human metabolism, these are generally taken into account and appear in the metabolic reconstructions, albeit many 'known' metabolites still do not [9]. The ability to transport such compounds may be of relatively recent evolutionary origin, much as is the ability of mammals to digest lactose in adulthood [136–138] (which is also highly variable between individuals and indeed races [139, 140]). We also note that the experimental serum metabolome listed at http://www.serummetabolome.ca/statistics [141] refers to 2243 endogenous metabolites but 3363 exogenous metabolites, with the corresponding numbers for the human urine metabolome [142] being 1665 endogenous metabolites and 3363 exogenous metabolites.

At all events, when we compared the differences in the magnitude of the similarity between MACCS_Tanimoto and MCS_Tversky, it was clear that they could be positive or negative, although MACCS was more often the larger, but that no individual descriptor could account for these differences, even though they were clearly deterministic (as are the analyses). Overall, though, it is clear that the use of the MCS adds significantly to the armoury of similarity strategies for those seeking to compare the structural similarities between synthetic drugs and natural biomolecules.

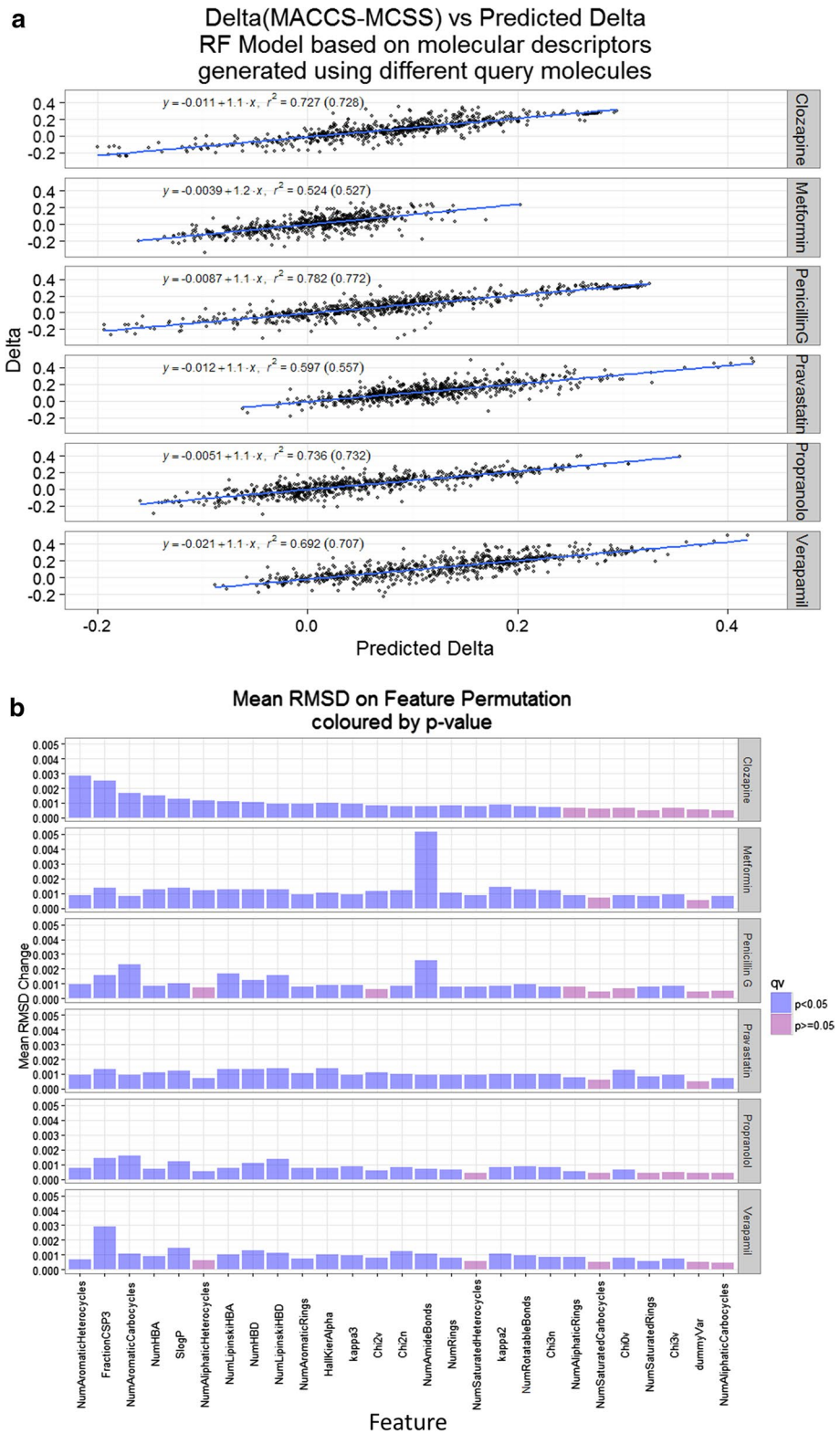

Fig. 8 Random Forest prediction of the differences (Delta) between MACCS_Tanimoto and MCS_Tversky similarities. **a** *Scatterplot* with regression coefficients for 6 drugs. **b** Contribution of each of the retained RDKit features for each drug

Table 1 Variation in sign

Molecule	Positive difference MACCS_TS–MCS_Tv	Negative difference MACCS_TS–MCS_Tv	% with a positive difference
Clozapine	1366	379	78.3
Metformin	1034	711	59.3
Benzylpenicillin	1282	463	73.5
Pravastatin	1575	170	90.3
Propranolol	1172	573	67.2
Verapamil	1496	249	85.7

Conclusion

The extent to which two molecules are to be seen as 'similar' in purely (2D) structural terms depends strongly on both their encoding and the similarity metric used, and this was the case for our drug–endogenite analyses as performed previously [20–22]. In the absence of 'activity' or 'functional' data, the only comparators for 'closeness' rely on purely unsupervised methods of analysis. It is clear that not all of a drug will typically bind to its 'target' (not least since some molecular features will have been designed in for other purposes, e.g. ADME). However, the extent of this is normally not known, and probably not knowable, and that necessarily underpins part of the functional variation in similarity.

One strategy to ensure that we pick up pertinent similarities is to use as many methods as possible for encoding them, and we here sought to assess the maximal common substructure (MCS) as an additional useful similarity measure. MCS also has the advantage of having a clear chemical meaning in terms of a linked set of atoms. Although, again, the extent to which the MCS showed up similarities observable via the MACCS fingerprint varied significantly between drugs, the corresponding conclusion was precisely that, as a consequence of this, the MCS was valuable as an additional method in such comparisons. To reiterate, we do not imply that MCS is 'better' or 'worse' than other methods, but we do think that the evidence shows that it is different and correspondingly valuable, and should thus be used in parallel with fingerprinting methods, whether separately or (as often done to advantage, e.g. [63, 143, 144]), via fusion methods. Finally, a referee wondered whether there might be a correlation between MCS-similarity to the nearest endogenite and bioavailability. The present analysis now opens up the possibility of answering precisely these and other such questions.

Additional files

Additional file 1. Workflow of Fig. 2 used to generate the data shown in Fig. 1.

Additional file 2. Python code used to generate substructures.

Additional file 3. Comparison of endogenites with endogenites in terms of their maximum common substructures.

Additional file 4. Comparison of marketed drugs with marketed drugs in terms of their maximum common substructures.

Additional file 5. Comparison of endogenites with marketed drugs in terms of their maximum common substructures.

Authors' contributions

SO'H wrote most of the workflows; some were modified by DBK. Both authors read and approved the final manuscript.

Author details

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

Acknowledgements

DBK thanks Dr David Hepworth for a useful discussion that finally motivated him to look more closely at MCS analyses, and Prof Terry Brown for reminding him of the recent evolutionary origin of lactase persistence. We thank the BBSRC for financial support (Grants BB/K019783/1 and BB/M017702/1). Two anonymous reviewers provided excellent, fair and detailed comments that helped us to improve this paper significantly during the refereeing process.

Competing interests

The authors declare that they have no competing interests.

Received: 14 November 2016 Accepted: 9 February 2017
Published online: 09 March 2017

References

1. Dobson PD, Kell DB (2008) Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? Nat Rev Drug Disc 7:205–220
2. Dobson PD, Patel Y, Kell DB (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 SG, Kell DB (2009) Implications of the dominant role of cellular transporters in drug uptake. Curr Top Med Chem 9:163–184
4. Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KL, Chu X, Dahlin A, Evers R, Fischer V, Hillgren KM et al (2010) Membrane transporters in drug development. Nat Rev Drug Discov 9(3):215–236
5. Kell DB, Dobson PD, Oliver SG (2011) Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only. Drug Disc Today 16(15/16):704–714
6. Kell DB, Dobson PD, Bilsland E, Oliver SG (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(5/6):218–239
7. Kell DB (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 (eds) (2013) Transporters in drug development: discovery, optimization, clinical study and regulation. AAPS/Springer, New York
9. Kell DB, Goodacre R (2014) Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. Drug Disc Today 19(2):171–182
10. Kell DB, Oliver SG (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 GE, Radic B, Mayor-Ruiz C, Blomen VA, Trefzer C, Kandasamy RK, Huber KVM, Gridling M, Chen D, Klampfl T et al (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 RA, Hepworth D, Hediger MA, Edwards AM et al (2015) A call for systematic research on solute carriers. Cell 162(3):478–487
13. Kell DB (2015) What would be the observable consequences if phospholipid bilayer diffusion of drugs into cells is negligible? Trends Pharmacol Sci 36(1):15–21
14. Mendes P, Oliver SG, Kell DB (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 DB (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 DB (2016) Implications of endogenous roles of transporters for drug discovery: hitchhiking and metabolite-likeness. Nat Rev Drug Disc 15(2):143–144
17. Kell DB (2016) How drugs pass through biological cell membranes—a paradigm shift in our understanding? Beilstein Mag 2(5). http://www.beilstein-institut.de/download/628/609_kell.pdf
18. Mooij MG, Nies AT, Knibbe CAJ, Schaeffeler E, Tibboel D, Schwab M, de Wildt SN (2016) Development of human membrane transporters: drug disposition and pharmacogenetics. Clin Pharmacokinet 55(5):507–524
19. Govindarajan R, Sparreboom A (2016) Drug transporters: advances and opportunities. Clin Pharmacol Ther 100(5):398–403
20. O'Hagan S, Swainston N, Handl J, Kell DB (2015) A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. Metabolomics 11(2):323–339
21. O'Hagan S, Kell DB (2015) Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites. Front Pharmacol 6:105
22. O'Hagan S, Kell DB (2016) MetMaxStruct: a Tversky-similarity-based strategy for analysing the (sub)structural similarities of drugs and endogenous metabolites. Front Pharmacol 7:266
23. Karakoc E, Sahinalp SC, Cherkasov A (2006) Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. J Chem Inf Model 46(5):2167–2182
24. Gupta S, Aires-de-Sousa J (2007) Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. Mol Divers 11(1):23–36
25. Khanna V, Ranganathan S (2009) Physicochemical property space distribution among human metabolites, drugs and toxins. BMC Bioinform 10(Suppl 15):S10
26. Peironcely JE, Reijmers T, Coulier L, Bender A, Hankemeier T (2011) Understanding and classifying metabolite space and metabolite-likeness. PLoS ONE 6(12):e28966
27. Hamdalla MA, Mandoiu II, Hill DW, Rajasekaran S, Grant DF (2013) BioSM: metabolomics tool for identifying endogenous mammalian biochemical structures in chemical structure space. J Chem Inf Model 53(3):601–612
28. Gasteiger J (ed) (2003) Handbook of chemoinformatics: from data to knowledge. Wiley/VCH, Weinheim
29. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2(22):3204–3218
30. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular similarity in medicinal chemistry. J Med Chem 57:3186–3204
31. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, Haraldsdottír H, Mo ML, Rolfsson O, Stobbe MD et al (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31(5):419–425

32. Swainston N, Mendes P, Kell DB (2013) An analysis of a 'community-driven' reconstruction of the human metabolic network. Metabolomics 9(4):757–764
33. Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, Zielinski DC, Ang KS, Gardiner NJ, Gutierrez JM et al (2016) Recon 2.2: from reconstruction to model of human metabolism. Metabolomics 12:109
34. Everitt BS (1993) Cluster analysis. Edward Arnold, London
35. Maldonado AG, Doucet JP, Petitjean M, Fan BT (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers 10(1):39–79
36. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11(23–24):1046–1053
37. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today 12(5–6):225–233
38. Maggiora GM, Shanmugasundaram V (2011) Molecular similarity measures. Methods Mol Biol 672:39–100
39. Willett P (2011) Similarity searching using 2D structural fingerprints. Meth Mol Biol 672:133–158
40. Willett P (2014) The calculation of molecular structural similarity: principles and practice. Mol Inform 33(6–7):403–413
41. O'Boyle NM, Sayle RA (2016) Comparing structural fingerprints using a literature-based similarity benchmark. J Cheminform 8:36
42. Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P (2012) Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. J Chem Inf Model 52(11):2884–2901
43. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42(6):1273–1280
44. Tversky A (1977) Features of similarity. Psychol Rev 84(4):327–352
45. Horvath D, Marcou G, Varnek A (2013) Do not hesitate to use Tversky- and other hints for successful active analogue searches with feature count descriptors. J Chem Inf Model 53(7):1543–1562
46. Kawabata T (2011) Build-up algorithm for atomic correspondence between chemical structures. J Chem Inf Model 51(8):1775–1787
47. Barker EJ, Buttar D, Cosgrove DA, Gardiner EJ, Kitts P, Willett P, Gillet VJ (2006) Scaffold hopping using clique detection applied to reduced graphs. J Chem Inf Model 46(2):503–511
48. Renner S, Schneider G (2006) Scaffold-hopping potential of ligand-based similarity concepts. ChemMedChem 1(2):181–185
49. Cao Y, Jiang T, Girke T (2008) A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Bioinformatics 24(13):i366–i374
50. Krueger BA, Dietrich A, Baringhaus KH, Schneider G (2009) Scaffold-hopping potential of fragment-based de novo design: the chances and limits of variation. Comb Chem High Throughput Screen 12(4):383–396
51. Vogt M, Stumpfe D, Geppert H, Bajorath J (2010) Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. J Med Chem 53(15):5707–5715
52. Hu Y, Bajorath J (2011) Combining horizontal and vertical substructure relationships in scaffold hierarchies for activity prediction. J Chem Inf Model 51(2):248–257
53. Bone RGA, Villar HO (1997) Exhaustive enumeration of molecular substructures. J Comput Chem 18(1):86–107
54. Raymond JW, Willett P (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. J Comput Aided Mol Des 16(7):521–533
55. Raymond JW, Willett P (2002) Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. J Comput Aided Mol Des 16(1):59–71
56. Cerruela García G, Luque Ruiz I, Gómez-Nieto MA (2004) Step-by-step calculation of all maximum common substructures through a constraint satisfaction based algorithm. J Chem Inf Comput Sci 44(1):30–41
57. Grosso A, Locatelli M, Pullan W (2008) Simple ingredients leading to very efficient heuristics for the maximum clique problem. J Heurist 14(6):587–612
58. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50(3):339–348
59. Hariharan R, Janakiraman A, Nilakantan R, Singh B, Varghese S, Landrum G, Schuffenhauer A (2011) MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. J Chem Inf Model 51(4):788–806
60. Wang Y, Backman TWH, Horan K, Girke T (2013) fmcsR: mismatch tolerant maximum common substructure searching in R. Bioinformatics 29(21):2792–2794
61. Chen J, Sheng J, Lv D, Zhong Y, Zhang G, Nan P (2014) The optimization of running time for a maximum common substructure-based algorithm and its application in drug design. Comput Biol Chem 48:14–20
62. Kumar A, Maranas CD (2014) CLCA: maximum common molecular substructure queries within the MetRxn database. J Chem Inf Model 54(12):3417–3438
63. Duesbury E, Holliday J, Willett P (2015) Maximum common substructure-based data fusion in similarity searching. J Chem Inf Model 55(2):222–230
64. Englert P, Kovács P (2015) Efficient heuristics for maximum common substructure search. J Chem Inf Model 55(5):941–955
65. Kunimoto R, Vogt M, Bajorath J (2016) Maximum common substructure-based Tversky index: an asymmetric hybrid similarity measure. J Comput Aided Mol Des 30(7):523–531
66. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: the Konstanz Information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) Data analysis, machine learning and applications. Springer, Berlin, pp 319–326
67. Mazanetz MP, Marmon RJ, Reisser CBT, Morao I (2012) Drug discovery applications for KNIME: an open source data mining platform. Curr Top Med Chem 12(18):1965–1979
68. Meinl T, Jagla B, Berthold MR (2012) Integrated data analysis with KNIME. Woodh Pub Ser Biomed 16:151–171
69. Nicola G, Berthold MR, Hedrick MP, Gilson MK (2015) Connecting proteins with drug-like compounds: open source drug discovery workflows with BindingDB and KNIME. Database (Oxf) 2015:1–22
70. O'Hagan S, Kell DB (2015) Software review: the KNIME workflow environment and its applications in Genetic Programming and machine learning. Genet Progr Evol Mach 16:387–391
71. Saubern S, Guha R, Baell JB (2011) KNIME workflow to assess PAINS filters in SMARTS format. Comparison of RDKit and Indigo cheminformatics libraries. Mol Inform 30(10):847–850
72. Steinmetz FP, Mellor CL, Meinl T, Cronin MTD (2015) Screening chemicals for receptor-mediated toxicological and pharmacological endpoints: using public data to build screening tools within a KNIME workflow. Mol Inform 34(2–3):171–178
73. Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform 5(1):26
74. Zhang BJ, Vogt M, Maggiora GM, Bajorath J (2015) Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. J Comput Aided Mol Des 29(10):937–950
75. Wu MJ, Vogt M, Maggiora GM, Bajorath J (2016) Design of chemical space networks on the basis of Tversky similarity. J Comput Aided Mol Des 30(1):1–12
76. Geitmann M, Elinder M, Seeger C, Brandt P, de Esch IJP, Danielson UH (2011) Identification of a novel scaffold for allosteric inhibition of wild type and drug resistant HIV-1 reverse transcriptase by fragment library screening. J Med Chem 54(3):699–708
77. Senger S (2009) Using Tversky similarity searches for core hopping: finding the needles in the haystack. J Chem Inf Model 49(6):1514–1524
78. Gan S, Cosgrove DA, Gardiner EJ, Gillet VJ (2014) Investigation of the use of spectral clustering for the analysis of molecular data. J Chem Inf Model 54(12):3302–3319
79. Leucht S, Corves C, Arbter D, Engel RR, Li C, Davis JM (2009) Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis. Lancet 373(9657):31–41
80. Farooq S, Taylor M (2011) Clozapine: dangerous orphan or neglected friend? Br J Psychiatry 198(4):247–249
81. Leucht S, Cipriani A, Spineli L, Mavridis D, Orey D, Richter F, Samara M, Barbui C, Engel RR, Geddes JR et al (2013) Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. Lancet 382(9896):951–962
82. Selent J, Marti-Solano M, Rodríguez J, Atanes P, Brea J, Castro M, Sanz F, Loza MI, Pastor M (2014) Novel insights on the structural determinants

of clozapine and olanzapine multi-target binding profiles. Eur J Med Chem 77:91–95
83. Deehan GA Jr, Brodie MS, Rodd ZA (2013) What is in that drink: the biological actions of ethanol, acetaldehyde, and salsolinol. Curr Top Behav Neurosci 13:163–184
84. Hipólito L, Sánchez-Catalán MJ, Martí-Prats L, Granero L, Polache A (2012) Revisiting the controversial role of salsolinol in the neurobiological effects of ethanol: old and new vistas. Neurosci Biobehav Rev 36(1):362–378
85. Mravec B (2006) Salsolinol, a derivate of dopamine, is a possible modulator of catecholaminergic transmission: a review of recent developments. Physiol Res 55(4):353–364
86. Naoi M, Maruyama W, Akao Y, Yi H (2002) Dopamine-derived endogenous *N*-methyl-(*R*)-salsolinol: its role in Parkinson's disease. Neurotoxicol Teratol 24(5):579–591
87. Naoi M, Maruyama W, Nagy GM (2004) Dopamine-derived salsolinol derivatives as endogenous monoamine oxidase inhibitors: occurrence, metabolism and function in human brains. Neurotoxicology 25(1–2):193–204
88. O'Hagan S, Dunn WB, Brown M, Knowles JD, Kell DB (2005) Closed-loop, multiobjective optimisation of analytical instrumentation: gas-chromatography-time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. Anal Chem 77:290–303
89. Flower DR (1998) On the properties of bit string-based measures of chemical similarity. J Chem Inf Comput Sci 38(3):379–386
90. Godden JW, Stahura FL, Bajorath J (2000) Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. J Chem Inf Comput Sci 40(3):796–800
91. Al Khalifa A, Haranczyk M, Holliday J (2009) Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. J Chem Inf Model 49(5):1193–1201
92. Kimura N, Masuda S, Tanihara Y, Ueo H, Okuda M, Katsura T, Inui K (2005) Metformin is a superior substrate for renal organic cation transporter OCT2 rather than hepatic OCT1. Drug Metab Pharmacokinet 20(5):379–386
93. Becker ML, Visser LE, van Schaik RHN, Hofman A, Uitterlinden AG, Stricker BHC (2009) Genetic variation in the organic cation transporter 1 is associated with metformin response in patients with diabetes mellitus. Pharmacogenom J 9(4):242–247
94. Tzvetkov MV, Vormfelde SV, Balen D, Meineke I, Schmidt T, Sehrt D, Sabolić I, Koepsell H, Brockmöller J (2009) The effects of genetic polymorphisms in the organic cation transporters OCT1, OCT2, and OCT3 on the renal clearance of metformin. Clin Pharmacol Ther 86(3):299–306
95. Zolk O (2009) Current understanding of the pharmacogenomics of metformin. Clin Pharmacol Ther 86(6):595–598
96. Minematsu T, Giacomini KM (2011) Interactions of tyrosine kinase inhibitors with organic cation transporters and multidrug and toxic compound extrusion proteins. Mol Cancer Ther 10(3):531–539
97. Nies AT, Hofmann U, Resch C, Schaeffeler E, Rius M, Schwab M (2011) Proton pump inhibitors inhibit metformin uptake by organic cation transporters (OCTs). PloS ONE 6(7):e22163
98. Han TK, Proctor WR, Costales CL, Cai H, Everett RS, Thakker DR (2015) Four cation-selective transporters contribute to apical uptake and accumulation of metformin in Caco-2 cell monolayers. J Pharmacol Exp Ther 352(3):519–528
99. Ciarimboli G, Gautron S, Schlatter E (eds) (2016) Organic cation transporters: integration of physiology, pathology and pharmacology. Springer, Heidelberg
100. Bretschneider B, Brandsch M, Neubert R (1999) Intestinal transport of beta-lactam antibiotics: analysis of the affinity at the H^+/peptide symporter (PEPT1), the uptake into Caco-2 cell monolayers and the transepithelial flux. Pharm Res 16(1):55–61
101. Luckner P, Brandsch M (2005) Interaction of 31 beta-lactam antibiotics with the H^+/peptide symporter PEPT2: analysis of affinity constants and comparison with PEPT1. Eur J Pharm Biopharm 59(1):17–24
102. Bailey PD, Boyd CA, Collier ID, George JP, Kellett GL, Meredith D, Morgan KM, Pettecrew R, Price RA (2006) Affinity prediction for substrates of the peptide transporter PepT1. Chem Commun (Camb) 3:323–325
103. Rubio-Aliaga I, Daniel H (2008) Peptide transporters and their roles in physiological processes and drug disposition. Xenobiotica 38(7–8):1022–1042
104. Smith DE, Clémençon B, Hediger MA (2013) Proton-coupled oligopeptide transporter family SLC15: physiological, pharmacological and pathological implications. Mol Aspects Med 34(2–3):323–336
105. Liao JK (2002) Beyond lipid lowering: the role of statins in vascular protection. Int J Cardiol 86(1):5–18
106. Undas A, Brozek J, Musial J (2002) Anti-inflammatory and antithrombotic effects of statins in the management of coronary artery disease. Clin Lab 48(5–6):287–296
107. Weitz-Schmidt G (2002) Statins as anti-inflammatory agents. Trends Pharmacol Sci 23(10):482–486
108. Blanco-Colio LM, Tuñon J, Martin-Ventura JL, Egido J (2003) Anti-inflammatory and immunomodulatory effects of statins. Kidney Int 63(1):12–23
109. Kwak BR, Mulhaupt F, Mach F (2003) Atherosclerosis: anti-inflammatory and immunomodulatory activities of statins. Autoimmun Rev 2(6):332–338
110. Steffens S, Mach F (2004) Anti-inflammatory properties of statins. Semin Vasc Med 4(4):417–422
111. Jain MK, Ridker PM (2005) Anti-inflammatory effects of statins: clinical evidence and basic mechanisms. Nat Rev Drug Discov 4(12):977–987
112. Abeles AM, Pillinger MH (2006) Statins as antiinflammatory and immunomodulatory agents: a future in rheumatologic therapy? Arthritis Rheum 54(2):393–407
113. Endres M (2006) Statins: potential new indications in inflammatory conditions. Atheroscler Suppl 7(1):31–35
114. Li JJ, Zheng X, Li J (2007) Statins may be beneficial for patients with slow coronary flow syndrome due to its anti-inflammatory property. Med Hypotheses 69(2):333–337
115. Mira E, Manes S (2009) Immunomodulatory and anti-inflammatory activities of statins. Endocr Metab Immune Disord Drug Targets 9(3):237–247
116. Dinarello CA (2010) Anti-inflammatory agents: present and future. Cell 140(6):935–950
117. Bu DX, Griffin G, Lichtman AH (2011) Mechanisms for the anti-inflammatory effects of statins. Curr Opin Lipidol 22(3):165–170
118. Antonopoulos AS, Margaritis M, Lee R, Channon K, Antoniades C (2012) Statins as anti-inflammatory agents in atherogenesis: molecular mechanisms and lessons from the recent clinical trials. Curr Pharm Des 18(11):1519–1530
119. Kell DB (2009) Iron behaving badly: inappropriate iron chelation as a major contributor to the aetiology of vascular and other progressive inflammatory and degenerative diseases. BMC Med Genom 2:2
120. Wagner BK, Kitami T, Gilbert TJ, Peck D, Ramanathan A, Schreiber SL, Golub TR, Mootha VK (2008) Large-scale chemical dissection of mitochondrial function. Nat Biotechnol 26:343–351
121. Kell DB (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
122. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38(6):983–996
123. Dixon SL, Koehler RT (1999) The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. J Med Chem 42(15):2887–2900
124. Salim N, Holliday J, Willett P (2003) Combination of fingerprint-based similarity coefficients using data fusion. J Chem Inf Comput Sci 43(2):435–442
125. Wang YA, Eckert H, Bajorath J (2007) Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. ChemMedChem 2(7):1037–1042
126. Wang Y, Bajorath J (2008) Balancing the influence of molecular complexity on fingerprint similarity searching. J Chem Inf Model 48(1):75–84
127. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
128. Knight CG, Platt M, Rowe W, Wedge DC, Khan F, Day P, McShea A, Knowles J, Kell DB (2009) Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. Nucleic Acids Res 37(1):e6

129. Carbonell P, Parutto P, Baudier C, Junot C, Faulon JL (2014) Retropath: automated pipeline for embedded metabolic circuits. ACS Synth Biol 3(8):565–577
130. Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 79:471–505
131. Carbonell P, Faulon JL (2010) Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics 26(16):2012–2019
132. Carbonell P, Lecointre G, Faulon JL (2011) Origins of specificity and promiscuity in metabolic networks. J Biol Chem 286(51):43994–44004
133. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srvivas R, Palsson BØ (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci 104(6):1777–1782
134. Scalbert A, Brennan L, Manach C, Andres-Lacueva C, Dragsted LO, Draper J, Rappaport SM, van der Hooft JJ, Wishart DS (2014) The food metabolome: a window over dietary exposure. Am J Clin Nutr 99(6):1286–1308
135. Gibbons H, Brennan L (2017) Metabolomics as a tool in the identification of dietary biomarkers. Proc Nutr Soc 76(1):42–53
136. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6):1111–1120
137. Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG (2011) Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond B Biol Sci 366(1566):863–877
138. Walter J, Ley R (2011) The human gut microbiome: ecology and recent evolutionary changes. Annu Rev Microbiol 65:411–429
139. Sibley E (2004) Genetic variation and lactose intolerance: detection methods and clinical implications. Am J Pharmacogenomics 4(4):239–245
140. Mattar R, de Campos Mazo DF, Carrilho FJ (2012) Lactose intolerance: diagnosis, genetic, and clinical factors. Clin Exp Gastroenterol 5:113–121
141. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B et al (2011) The human serum metabolome. PLoS ONE 6(2):e16957
142. Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, Bjorndahl TC, Krishnamurthy R, Saleem F, Liu P et al (2013) The human urine metabolome. PLoS ONE 8(9):e73076
143. Willett P (2006) Enhancing the effectiveness of ligand-based virtual screening using data fusion. QSAR Combin Sci 25(12):1143–1152
144. Willett P (2013) Combination of similarity rankings using data fusion. J Chem Inf Model 53(1):1–10