Metabolomics (2015) 11:323–339
DOI 10.1007/s11306-014-0733-z

ORIGINAL ARTICLE

A ‘rule of 0.5’ for the metabolite-likeness of approved pharmaceutical drugs

Steve O′Hagan · Neil Swainston · Julia Handl · Douglas B. Kell

Received: 4 July 2014 / Accepted: 8 September 2014 / Published online: 19 September 2014

Abstract We exploit the recent availability of a community reconstruction of the human metabolic network (‘Recon2’) to study how close in structural terms are marketed drugs to the nearest known metabolite(s) that Recon2 contains. While other encodings using different kinds of chemical fingerprints give greater differences, we find using the 166 Public MDL Molecular Access (MACCS) keys that 90 % of marketed drugs have a Tanimoto similarity of more than 0.5 to the (structurally) ‘nearest’ human metabolite. This suggests a ‘rule of 0.5’ mnemonic for assessing the metabolite-like properties that characterise successful, marketed drugs. Multiobjective clustering leads to a similar conclusion, while artificial (synthetic) structures are seen to be less human-metabolite-like. This ‘rule of 0.5’ may have considerable predictive value in chemical biology and drug discovery, and may represent a powerful filter for decision making processes.

Keywords Genome-wide metabolic reconstruction · Recon 2 · Cheminformatics · KNIME · Metabolite-likeness · Drug-likeness

1 Introduction

The declining productivity of the drug discovery process is well known (e.g. Empfield and Leeson 2010; Hay et al. 2014; Kell 2013; Kola 2008; Kola and Landis 2004; Rafols et al. 2014; van der Greef and McBurney 2005). Thus, many groups have sought to assess in silico those structural or biophysical properties of successful drugs that might be used as filters to enrich the contents of drug discovery libraries with molecules that share those properties. This has therefore led to concepts such as “drug-likeness” (e.g. Empfield and Leeson 2010; Hay et al. 2014; Kell 2013; Kola 2008; Kola and Landis 2004; van der Greef and McBurney 2005), “lead-likeness” (Gozalbes and Pineda-Lucena 2011; Holdgate 2007; Oprea et al. 2007, 2001; Wunberg et al. 2006), and “ligand efficiency” (Hopkins et al. 2014) by which the potentially desirable properties of such molecules have been assessed.

We recognise that any molecule bioactive in human cells (whether as a drug or for purposes of chemical genomics) must cross at least one membrane, that nutrients necessarily do so, that natural products remain a major source of successful (marketed) pharmaceutical drugs (Gozalbes and Pineda-Lucena 2011; Holdgate 2007; Oprea et al. 2007, 2001; van Deursen et al. 2011; Wunberg et al. 2006), and that successful drugs require or at least use membrane transporters (Dobson et al. 2009; Dobson and

Electronic supplementary material The online version of this article (doi:10.1007/s11306-014-0733-z) contains supplementary material, which is available to authorized users.

S. O′Hagan · D. B. Kell (✉)
School of Chemistry, The University of Manchester,
131 Princess St, Manchester M1 7DN, UK
e-mail: dbk@manchester.ac.uk
URL: http://dbkgroup.org/

S. O′Hagan · N. Swainston · D. B. Kell
The Manchester Institute of Biotechnology, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK

N. Swainston
School of Computer Science, The University of Manchester,
131 Princess St, Manchester M1 7DN, UK

J. Handl
Manchester Business School, The University of Manchester,
131 Princess St, Manchester M1 7DN, UK

Kell 2008; Giacomini and Huang 2013; Giacomini et al. 2010; Kell 2013; Kell and Dobson 2009; Kell et al. 2013, 2011; Kell and Goodacre 2014; Lanthaler et al. 2011) that normally are used for the transport of intermediary metabolites (Herrgård et al. 2008; Swainston et al. 2013; Thiele et al. 2013). Given the natural role for these transporters as transporters of intermediary metabolites, we and others have thus suggested (hypothesised) that successful drugs are in fact much more like metabolites (we use this term to mean the natural intermediary metabolites of human metabolism, and do not consider metabolites of the drugs) than are the typical structures found in drug discovery libraries (e.g. Chen et al. 2012; Dobson et al. 2009; Feher and Schmidt 2003; Gupta and Aires-de-Sousa 2007; Hamdalla et al. 2013; Karakoc et al. 2006; Khanna and Ranganathan 2009, 2011; Peironcely et al. 2011; Walters 2012; Zhang et al. 2011), and following the principle of molecular similarity (e.g. Bender and Glen 2004; Eckert and Bajorath 2007; Gasteiger 2003; Maldonado et al. 2006; Oprea 2004; Sheridan et al. 2004) that "metabolite-likeness" is therefore a useful criterion for the design of successful drugs (Dobson et al. 2009). At one level, this may not be seen as surprising given the fact that pharmaceutical drugs typically bind to proteins at sites to which endogenous metabolites normally bind, but the recognition of the importance of metabolite-likeness in drug discovery and chemical genomics remains less than complete.

While a variety of metabolite (pathway) databases exist (Ooi et al. 2010) [e.g. ChEBI (de Matos et al. 2012; Degtyarenko et al. 2009; Hastings et al. 2013), HMDB (Wishart et al. 2013), KEGG (Kanehisa et al. 2012, 2014), MetaCyc (Altman et al. 2013; Caspi et al. 2014; Karp and Caspi 2011) and MetaboLights (Haug et al. 2013)], the recent availability of a highly curated consensus map (Recon2) of the human metabolic network (and thus of intermediary metabolites) (Swainston et al. 2013; Thiele et al. 2013) now provides the most suitable starting point for the comparison of drugs that have been approved/marketed [available from DrugBank (Knox et al. 2011; Law et al. 2014)] and metabolites that are known to be part of the human metabolic network. We choose this latter over say HMDB since the measurable presence of a molecule in a human sample (e.g. Dunn et al. 2014) does not exclude that it has a nutritional, xenobiotic or gut microbial origin, and HMDB does contain many 'metabolites' that are not in fact produced via pathways containing proteins encoded by the human genome. Indeed Peironcely et al. (2011) noted, for instance, that the 'metabolite' debrisoquine was indeed classified in their scheme as a non-metabolite (and it is indeed a marketed drug).

Thus the primary purpose of this work (in contrast to our earlier work (Dobson et al. 2009) that included multiple metabolite databases that were not constrained as here), is to use the availability of Recon2 to assess precisely how 'metabolite-like' known drugs are, partly as an aid to developing metrics for determining whether drugs are likely to be substrates for relevant transporters and thus whether they are likely to be bioactive. The availability of Recon2 also allows us to reason sensibly about the nature and extent of metabolite space and how it differs from the kinds of molecules typically found in drug discovery libraries.

2 Methods

2.1 Construction of datasets

The list of FDA-approved small molecule drugs was downloaded from DrugBank 3.0 (http://www.drugbank.ca/downloads) in November 2013 as an SDF file and consists of 1491 molecules. This is significantly smaller than the fuller list (7330 'drugs' via Drugbank and KEGGDrug) used previously (Dobson et al. 2009). The list of intermediary metabolites was extracted from the latest version of the Recon2 human metabolic network (Thiele et al. 2013). A further manual curation removed from the 'drugs' list (i) 'drugs' (mainly nutritional supplements) that are also intermediary metabolites produced by enzymes encoded by the genome and thus part of Recon 2 (though adrenaline was treated as a drug), and (ii) those 'metabolites' listed in Recon2 that are xenobiotic in nature or simply metals or salts. However, vitamins and essential amino acids and fatty acids, while not encoded by the human genome, were retained as 'metabolites' as they are both necessary for human metabolism and form part of the formal human metabolic network. The resultant data are in Supplementary information S3, and consist of 1113 'metabolites' [cf. 5333 'metabolites' previously (Dobson et al. 2009)] and 1381 'drugs'. In addition, data on antimalarial compounds were downloaded from the databases at the EBI (https://www.ebi.ac.uk/chemblntd).

2.2 Software

For the cheminformatics analyses we used the KoNstanz Information MinEr (KNIME, www.knime.org) (Beisken et al. 2013; Berthold et al. 2007; Mazanetz et al. 2012; Meinl et al. 2012; Stöter et al. 2013; Warr 2012). KNIME is a workflow environment somewhat similar to Taverna [with which we have previous experience in systems biology analyses (Li et al. 2008a, b)], but which is slightly more focussed on cheminformatics. The workflows we used here included nodes that made use of libAnnotationSBML (Swainston and Mendes 2009), the Chemistry Development Kit (Beisken et al. 2013;

Steinbeck et al. 2003) and the RDKit (Riniker and Landrum 2013a; b; Saubern et al. 2011) (www.rdkit.org/). We also used the software MOCK (Handl and Knowles 2007) for multiobjective clustering.

3 Results

3.1 Comparison of Tanimoto distances between drugs and natural metabolites

Our first task was to assess the average chemical (structure) distances between molecules according to a suitable metric. Many molecular descriptors exist for encoding molecules in a manner that allows this (e.g. Bender 2010; Duan et al. 2010; Koutsoukas et al. 2013; Sastry et al. 2010; Sheridan and Kearsley 2002; Todeschini and Consonni 2000; Wang and Bajorath 2010), most commonly referred to as fingerprints (e.g. Faulon and Bender 2010; Flower 1998) and sometimes with rather different properties and outcomes when matched against structures or biological activities (e.g. Dhanda et al. 2013; Medina-Franco and Maggiora 2014). Thus, and while some experience shows that they are not greatly different from each other when simply comparing chemical or structural similarity (Dobson et al. 2009; Riniker and Landrum 2013a), which is the focus of the present paper, we looked at a number of methods for producing molecular fingerprints. Probably most common are fingerprints derived from structural keys such as the 166 Public MDL (Molecular ACCess System) MACCS keys (Durant et al. 2002) based on a predefined dictionary of 166 substructures [that contain most of the important features of a larger 960-key set (McGregor and Pallai 1997)] and hashed to give 1,024 bits.

Given the molecular fingerprint method chosen, there is a more general acceptance of the metrics for the similarity of molecules whose (sub)structures are so encoded; although it has a size-dependence (that does not matter for this analysis), the Tanimoto distance, that effectively encodes the numbers of matching and non-matching substructures, is both easy to calculate and pre-eminent (Maggiora et al. 2014; Willett 2006).

We recognise that some 20 % of recent new chemical entities are prodrugs (15 % in the top 100 drugs) (Huttunen et al. 2011), and that some of these are converted non-enzymically to the active substances; however, these normally do not differ greatly in structural terms from the active substance in the marketed entities, so for convenience we shall use the latter. In contrast to Peironcely et al. (2011), who used supervised learning methods such as random forests [which are very powerful (Knight et al. 2009)] to predict whether a substance was or was not a metabolite, we are here interested only in the structural similarities between candidate molecules and Recon2 metabolites, and we confine ourselves strictly to unsupervised methods of analysis.

We checked a variety of implementations of the MACCS fingerprints (specifically those used in Open Babel, CDK and RDKit) and found very little difference between them, and for what is presented here we used those in the RDKit implementation. We therefore compared all metabolites against all metabolites (Fig. 1a), all drugs against all drugs (Fig. 1b), and all drugs against all metabolites (Fig. 1c). The metabolite-metabolite similarities (Fig. 1a) reveal multiple clusters, including one that is made up of CoA derivatives (full details in Figure S1), while the clusters of drug-drug similarities Fig. 1b are rather more heterogeneous (the trees are much 'bushier'). From Fig. 1c, the drug-metabolite similarities, there are some interesting clusters, e.g. the block of red and yellow towards the upper left represented sterols and steroids, while the larger swathe of red and yellow towards the bottom represents mainly CoA derivatives. All the data are given in an addressable form as Excel spreadsheets in Supplementary Information S1–S3.

A number of different fingerprints were used to determine if the extent of closeness of a drug to its nearest metabolite depended greatly on the fingerprint used. The various fingerprints used (http://www.rdkit.org/RDKit_Docs.current.pdf) were provided in the RDKit module (Riniker and Landrum 2013a) (https://code.google.com/p/rdkit/wiki/FingerprintsInTheRDKit) of KNIME (http://tech.knime.org/community/rdkit), and as stated in (Riniker and Landrum 2013b) were atom pairs (AP), feature-based circular fingerprint with radius 2 as bit vector (FeatMorgan2), and a circular fingerprint with radius 2 as bit vector (Morgan2). Morgan2 is the RDKit implementation of the familiar ECFP4, and FeatMorgan2 is equivalent to FCFP4 (Landrum et al. 2011). The features used by the RDKit for FeatMorgan2 consist of various donors, acceptors, aromatic atoms, halogens, basic and acidic atoms. We also used a representation (referred to in KNIME and here as 'RDKit') that is said to be a 'Daylightlike' topological fingerprint based on hashing molecular subgraphs. Most recently, RDKit has added some extra fingerprints, and for completeness we included these too. Thus, 'layered' is an experimental substructure fingerprint using hashed molecular subgraphs, while 'torsion' is said to be the bit vector topological-torsion fingerprint for a molecule. As indicated above, all of the data are tabulated in Fig S3.

Considering first just the Tanimoto similarity (TS) values using MACCS fingerprints and the 1,024 bitstring encoding, 90 % of marketed drugs have a 'nearest metabolite Tanimoto similarity' (NMTS, i.e. the TS to the nearest metabolite) of more than 0.5, 98.5 % over 0.4 and 99 % over 0.34, all highly significant values (Baldi and

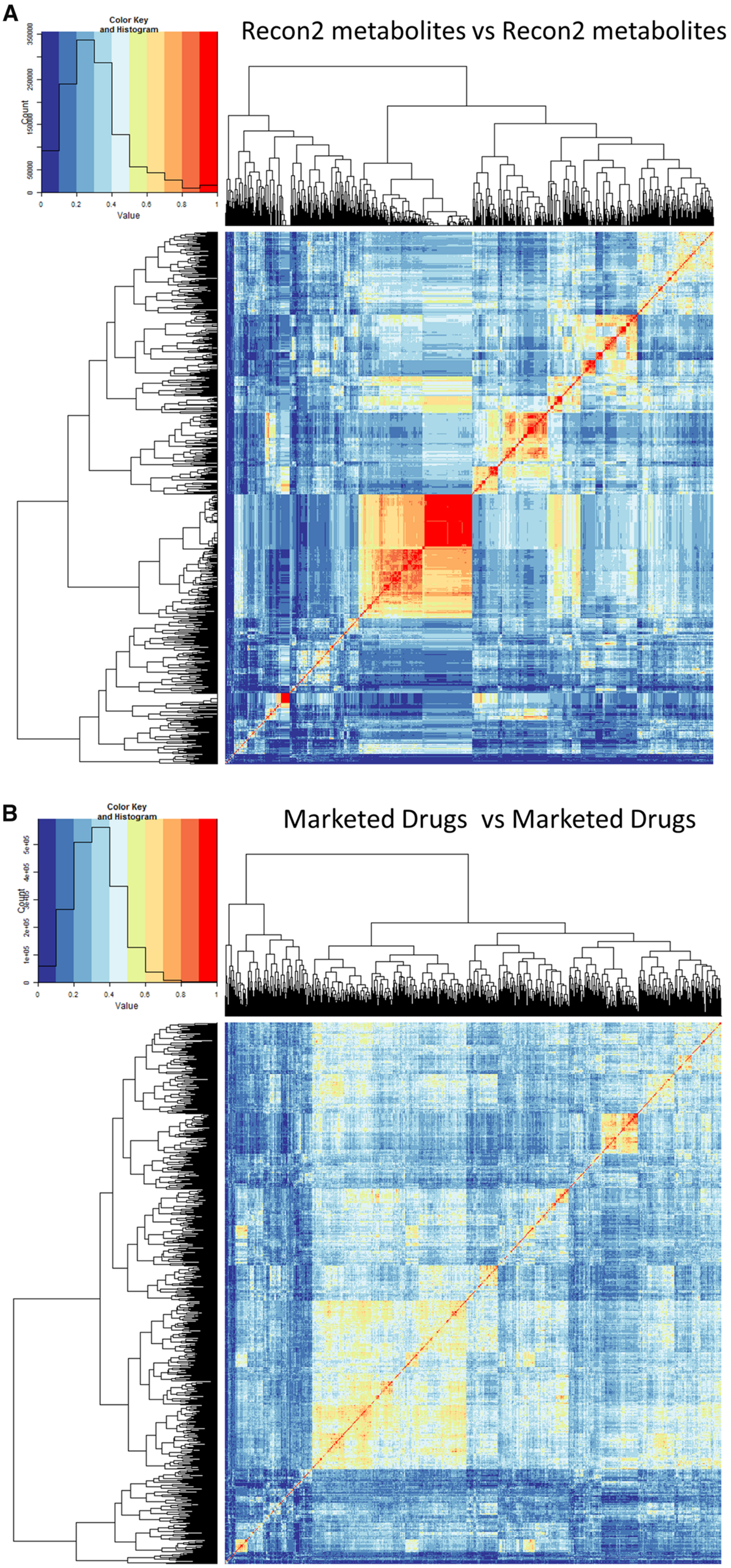
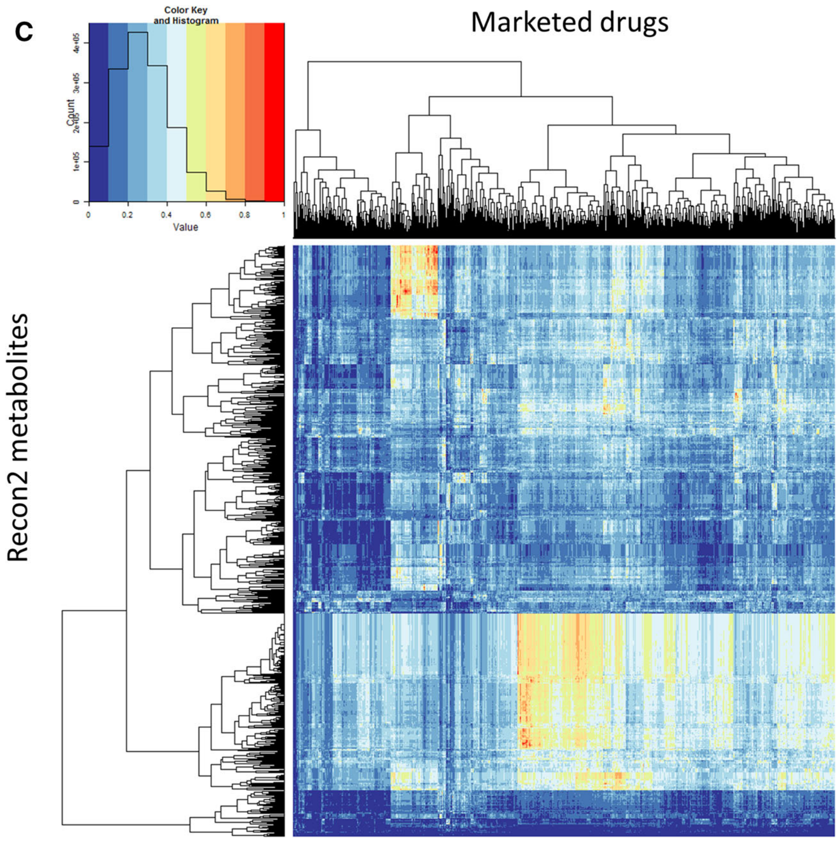


Fig. 1 Heat maps of the overall similarities between **a** Recon2 metabolites, **b** drugs and **c** each other. In the latter plot, the drugs lie on the *X*-axis and the metabolites on the *Y*-axis. Chemical structures were encoded using the MACCS encoding and Tanimoto distances calculated as described in Methods. The heat map representation (Eisen et al. 1998) encodes the numbers as a colour; in the present version, for ease of observation, we use ten discrete colours for the ten decades of Tanimoto similarity, with the colours chosen following the recommendations of Brewer et al. (1997) (see also http://www.colorbrewer2.org/). Also shown are hierarchical clusterings of the rows and columns (Eisen et al. 1998) using complete linkage and the default settings in the hclust function in R (Color figure online)

Fig. 1 continued



Nasr 2010). The first of those percentages compares with just 12 % when we did not use the ‘genuine’ human metabolites of Recon2 (Dobson et al. 2009) (note that there we used the nearest Tanimoto distance (=1 − TS)). Provided the molecule is not excessively halogenated, its NMTS is over 0.5 (e.g. 0.54 for Chlorzoxazone, 0.55 chlormerodrin, 0.6 diclofenac, 0.65 chlorphenesin and so on). This ‘rule’, by which the very great majority (90 % of) drugs are within a Tanimoto distance of 0.5 in MACCS fingerprint space, may be viewed in the context of the well-known ‘rule of 5’ (Lipinski et al. 1997) (Ro5) mnemonic for predicting drug lead quality. However, the cumulative plots of the NMTS for each drug using different fingerprints (Fig. 2a) do differ quite significantly depending on which fingerprint is used, and clearly the well-established MACCS fingerprints lead to a substantially greater degree of ‘metabolite-likeness’ than do almost all the other encodings (we do not pursue this here). Figure 2 also permits one to read off other metrics such as to note that more than 50 % of drugs have a TS greater than 0.6 to a metabolite for both MACCS and RDKit encodings.

Another indication of the rather different nature of the fingerprints comes from an analysis (Table 1) of the nature, and frequency of occurrence, of the nearest metabolite, where each fingerprint encoding has its own predilections for particular classes of metabolite, reflected also in the overall number of metabolites that are closest to at least one drug. These represent about one quarter of all drugs (or metabolites), an indication of the significant heterogeneity (Hopkins et al. 2014; Paolini et al. 2006) of drug space. RDKit has a slightly unusual predilection for cob(1)alamin and for protoheme, returned as the closest hits on 650 and 73 occasions, respectively (although removing these has negligible effects on the shape of the plot in Fig. 2a, indicating that this lower degree of metabolite-likeness, which is a continuous function, is inherent to the encoding). Scatter plots indicating correlations of ‘nearest metabolites’ with the different encodings are given in Fig. 2b, again illustrating the substantial differences found using the different encodings. Thus we would stress not only that similarity measures differ significantly for the different encodings, but that in functional terms the well-known existence of activity cliffs (e.g. Maggiora et al. 2014) means that quite small differences in molecular similarity may be highly significant with regard to pharmacological effects. In contrast to studies of related molecules that look at this (e.g. Muchmore et al. 2008; Papadatos et al. 2010), we discuss only the similarities themselves.

In a similar vein, the different encodings produce quite different assessments of the number of metabolites to

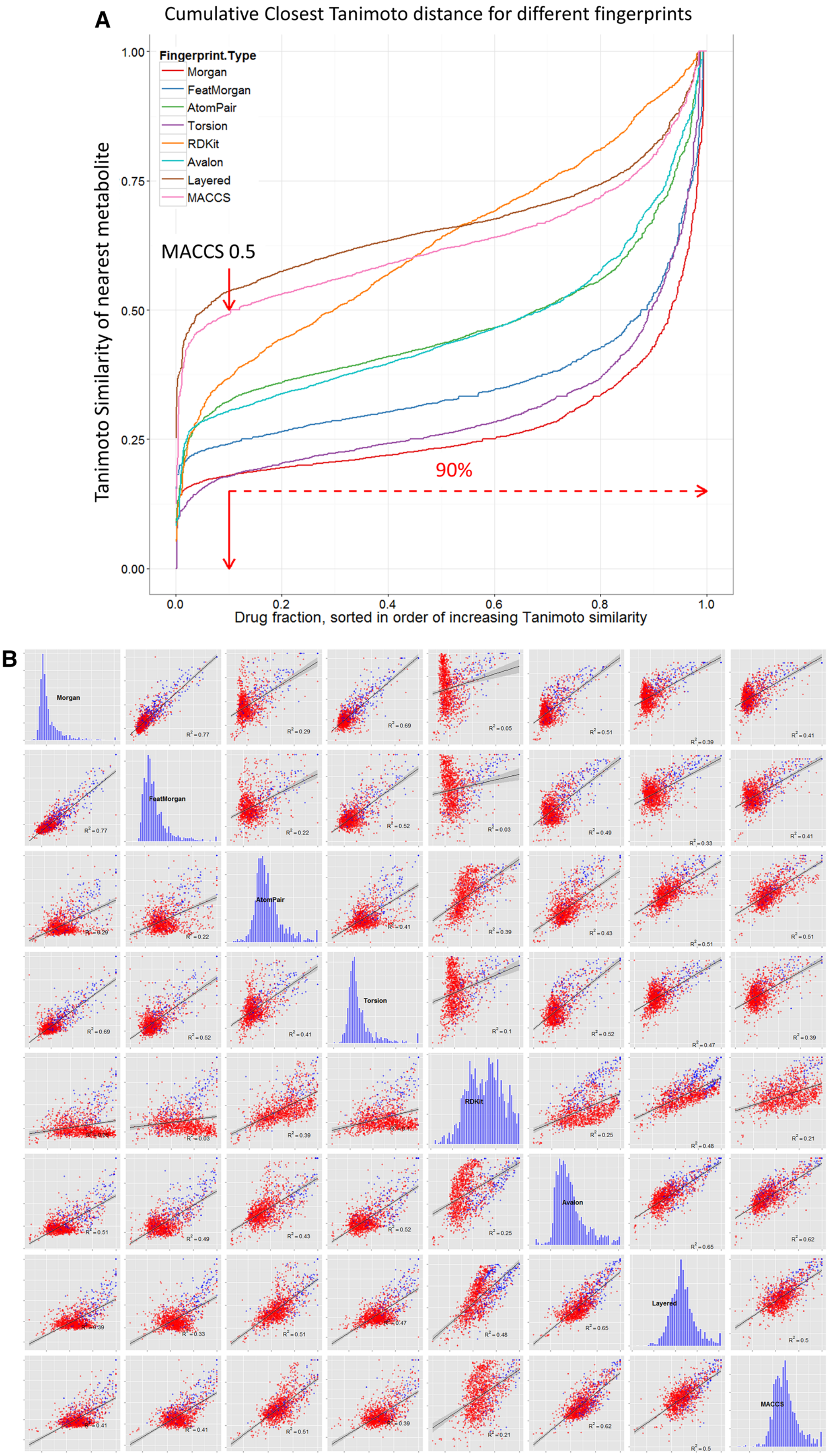


Fig. 2 Different structural encodings produce different drug-metabolite distances. **a** Cumulative plots of nearest drug-metabolite Tanimoto distances using various fingerprints. The number of drugs with a Tanimoto similarity of 0.5 or smaller is *arrowed* (i.e. all of those to the right, ca 90 %) have a Tanimoto similarity greater than 0.5. **b** Scatter plots relating the nearest Tanimoto distance to a metabolite for each drug; when the closest metabolites are the same for both encodings they are coloured *red*. Correlation coefficients are as given. The *blue* histograms represent the distributions of Tanimoto similarities for each of the encodings (scaled to fit the relevant windows). **c** Cumulative numbers of metabolites with a Tanimoto similarity ≥ 0.5 for various drugs and encodings. **d** The variation of the numbers of metabolites with a Tanimoto similarity ≥ 0.5 for all drugs using the MACCS encoding, with some of the highest labelled by name and with the chemical structure of arbekacin, the 'most promiscuously metabolite-like' of all, shown. **e** The 14 least metabolite-like drugs when using the MACCS encoding. **f** An assessment of part of drug-metabolite space where drugs are largely but not entirely distant from metabolites (Color figure online)

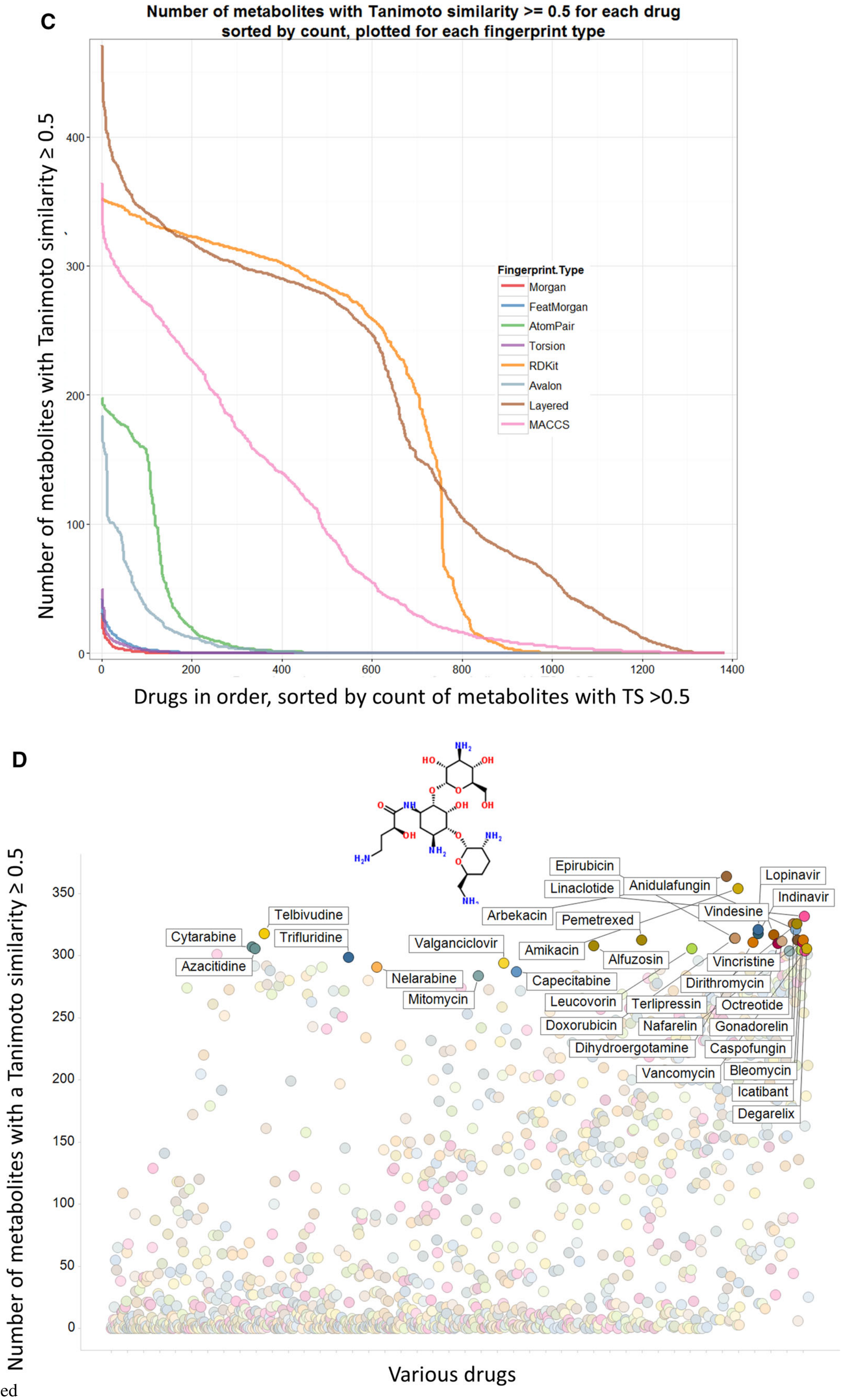


Fig. 2 continued

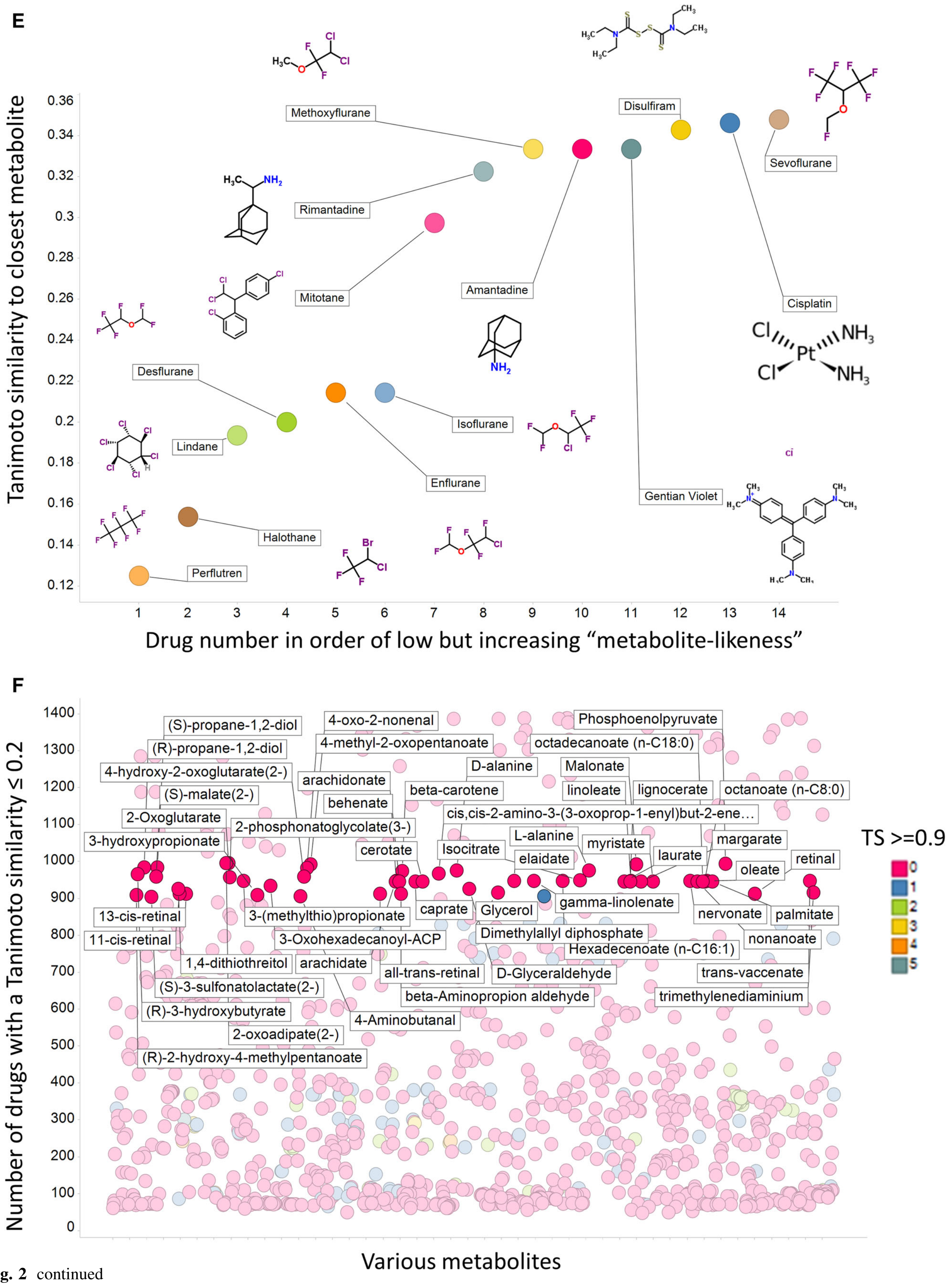


Fig. 2 continued

which each drug displays a Tanimoto similarity exceeding 0.5 (Fig. 2c), with (unsurprisingly, given the data in Fig. 2a) the MACCS, RDKit and Layered encodings showing the greatest tendency towards 'metabolite-likeness'. Based on MACCS, 50 % of marketed drugs have at least 31 metabolites with a TS of 0.5 or more. The 'winner' (i.e. the drug with the most metabolites to which it bears a TS greater than or equal to 0.5) is arbekacin, with 364, and the relevant data, plus a few named drugs, are given in Fig. 2d. It is probably worth commenting, albeit this is not necessarily a surprising finding, that these 'highly metabolite-like' drugs are natural products or molecules derived therefrom [see also (Kell 2013; Newman and Cragg 2012)]. The average greatest TS to a metabolite of the five most drug-like drugs (0.547), the five least drug-like drugs (0.683), the five most drug-like Ro5 failures (0.496) and the five least drug-like Ro5 passes (0.557, but minus tegaserod, not present in our list) as listed by Bickerton et al. (2012) are as noted.

By contrast, the substance with the lowest NMTS (perflutren, 0.125) is in fact an injectable contrast agent of lipid microspheres marketed precisely because it does not enter cells, while the next three lowest (NTS $\leq$ 0.2) are halothane (an inhalational narcotic), lindane (a topical chlorinated insecticide) and desflurane (a polyfluorinated inhalational anaesthetic), consistent with the fact that virtually no natural human metabolites are halogenated. Ten of the 14 least metabolite-like drugs contain at least two halogens (Fig. 2e).

In a similar vein, it is possible to enquire as to which metabolites have the most or fewest marketed drugs closely associated with them in terms of Tanimoto similarity, the latter in particular as a possible indication of areas of chemical space that might be deemed to be relatively underexplored. The metabolites with the very lowest TS to drugs are small and uninteresting (ammonia, water, etc.), so Fig. 2f illustrates those metabolites that are least similar to numbers of drugs between 900 and 1,000, at the same time illustrating the nonlinearity of drug and metabolite spaces by encoding with colours those metabolites that nonetheless have 1–5 drugs with a TS greater than or equal to 0.9 (glycerol is marked and has one, viz. mannitol). One might consider the sparsely populated areas of 'metabolite-likeness space' to be ones worth pursuing in drug discovery.

Another means of displaying the data, and a convenient means of interrogating them for a drug of interest, is given in Fig. 3, where we display the Tanimoto similarity to all metabolites for the beta-(adrenergic receptor) blocker propranolol. All metabolites with a TS greater than 0.5 are labelled, and structures are shown for (from left to right) propranolol itself, (−)-salsoline, adrenaline, L-normetanephrine, metanephrine and norepinephrine. While 'structural similarity' may be seen as a subjective matter, in this case the chemical similarities are obvious, and it is probably not surprising that a beta-adrenergic antagonist should have similarities of this type.

3.2 Multiobjective clustering of drugs and metabolites

In the above, we clustered (or bi-clustered) the drugs and the metabolites separately. Another approach to assessing the mapping of drug and metabolite spaces, and the extent to which they overlap or otherwise), is to use clustering methods of both together. These algorithms differ widely [there is no single 'correct' clustering (Everitt 1993)] but the state of the art is represented by methods such as MOCK (Handl and Knowles 2007) (MultiObjective Clustering with automatic K) that use multiple objectives [specifically both closeness and connectivity (Handl and Knowles 2007; Handl et al. 2005)] simultaneously to cluster objects on the basis of their 'similarity'. As with any multiobjective method, there are multiple 'best' solutions represented by a Pareto front (Kell 2012), and we illustrate this in Fig. 4. Figure 4a shows the overall variation of 'optimal' cluster number for the Pareto front, with 'knees' at e.g. 3, 7, 25, 30, 42 and 64 clusters, while Fig. 4b shows the distribution of drugs and metabolites in the MOCK solution for 25 clusters. Also marked are the 'top ten' blockbuster drugs by sales from 2010 [NB fluticasone propionate and salmeterol are part of a combined medicine; see also (Kell et al. 2013)], while the colour encodes the cluster membership of compounds when there are only seven clusters. Cluster 0 is mainly small metabolites like bicarbonate, but it is evident that the lower clusters all contain both metabolites and drugs. We also looked at the distribution of various molecular properties (such as polar surface area, molecular mass, log P etc.) between clusters, but no trends nor hotspots were apparent for particular clusters (not shown).

3.3 The drug-likeness of synthetic 'druglike' molecules and 'fragments' and of natural products

Having seen the closeness of successful, marketed drugs to metabolites when both are MACCS-encoded, it was important to establish that (while unlikely) this was not a strange artefact of the MACCS encoding itself. To this end, and while we and others (e.g. Dobson et al. 2009; Feher and Schmidt 2003; Khanna and Ranganathan 2011; Medina-Franco and Maggiora 2014; Ohno et al. 2010) have recognised that marketed drugs do differ structurally from most molecules in drug discovery libraries, despite their 'biogenic bias' (Hert et al. 2009), we sought to see how similar such non-marketed drug molecules or compounds are to marketed drugs when we compare them in the same way. The comparison is not entirely favourable to metabolites since we already know (Fig. 2) that many of the very smallest metabolite molecules are simply not druglike, and this is reflected in the data of Fig. 5. Figure 5a shows a heat map relating 2,000 structures taken randomly from the 30,000 in the Maybridge fragment library (similar kinds of map were obtained using subsets of varying sizes up to 15,000) relative to marketed drugs, while Fig. 5b shows that of a random subset of the Maybridge library vs Recon2 metabolites. Figure 5c shows the cumulative similarities (all using MACCS encodings) to metabolites for a collection of molecules from a subset of 1,000 molecules from the Maybridge fragment library, from the 13,533 compounds in the Tres Cantos Antimalarial Drug Set (Gamo

Table 1 Summary of the most frequently represented 'closest metabolite' to FDA-approved drugs, the number of times they appear, and the number of metabolites that are closest to a drug at least once

Fingerprint encoding	Most common 'closest metabolite'	Times represented	Total number of different metabolites that are 'closest' to a marketed drug at least once
MACCS	Docosa-4,7,10,13,16-pentaenoic acid	52	359
Atom pair	Linoleic coenzyme A	68	346
Feats Morgan	Docosa-4,7,10,13,16-pentaenoic acid	124	319
Morgan	Docosa-4,7,10,13,16-pentaenoic acid	77	338
RDKit	Methylcobalamin	650	268
Layered	Adenosylcobalamin	87	300
Avalon	Cortisol	65	213
Torsion	Vaccenyl coenzyme A	44	327

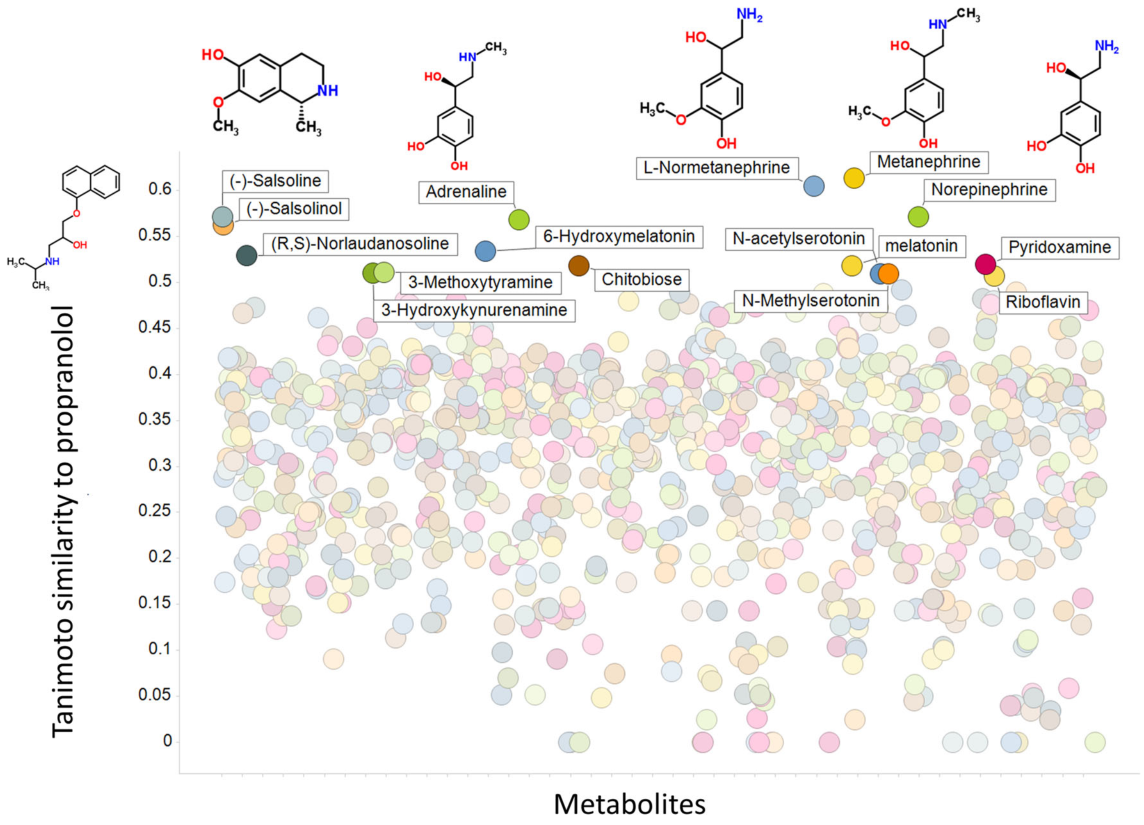


Fig. 3 Variation of the Tanimoto similarity for a marketed drug, propranolol, with various metabolites, those with a TS of over 0.5 being labelled, and structures given for a representative set to illustrate the close chemical similarity (Color figure online)

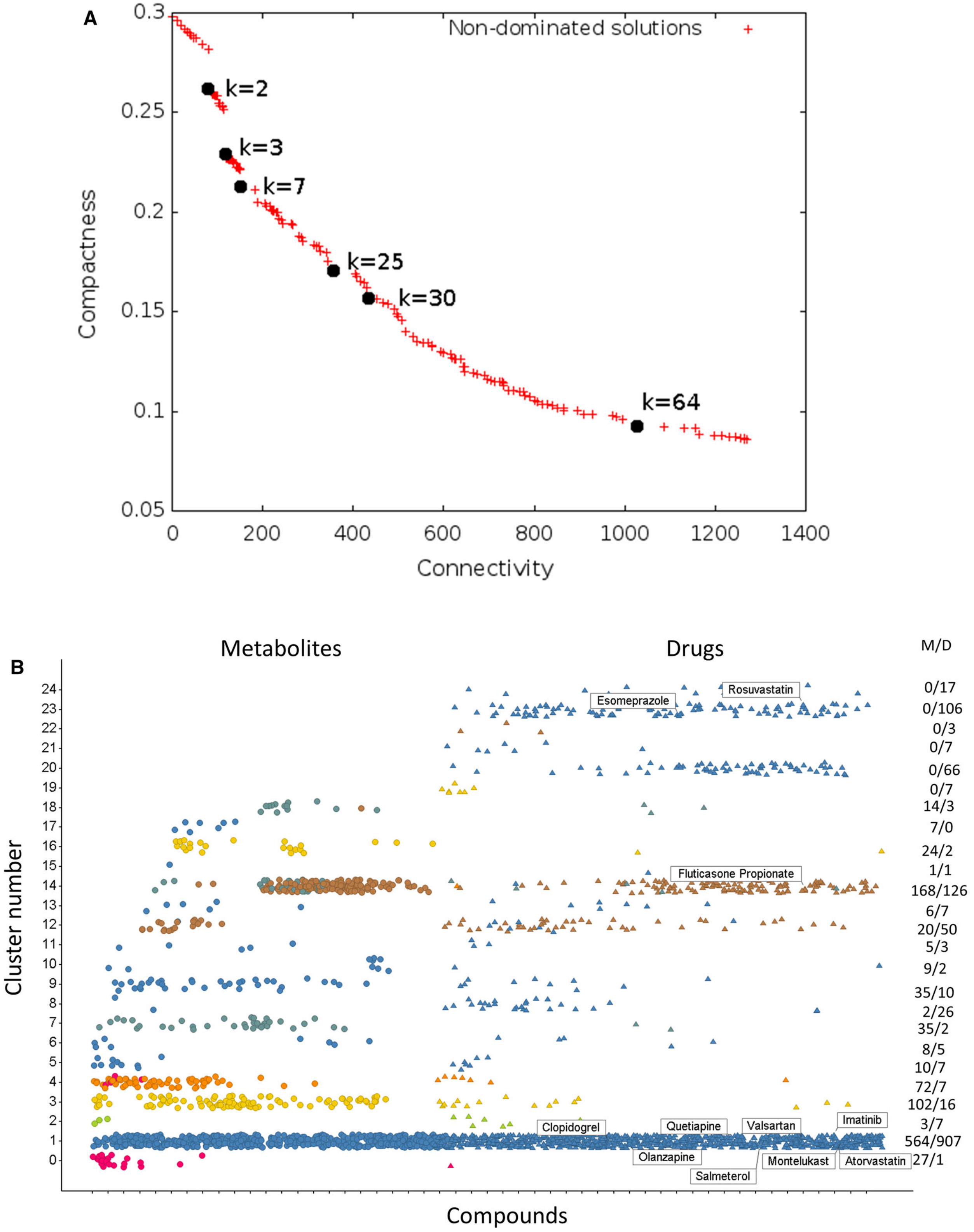


Fig. 4 Drug-metabolite clustering using the MACCS encoding and MOCK, a multiobjective clustering algorithm. **a** Dependence of cluster numbers as the weightings of the two main objectives are varied. The ‘knees’ at cluster numbers of 2, 3, 7, 25, 30 and 64 are marked. **b** Cluster membership and its distribution between drugs and metabolites for when 25 clusters are chosen. Data are ‘jittered’ in the *Y* direction to make them clearer (Color figure online)

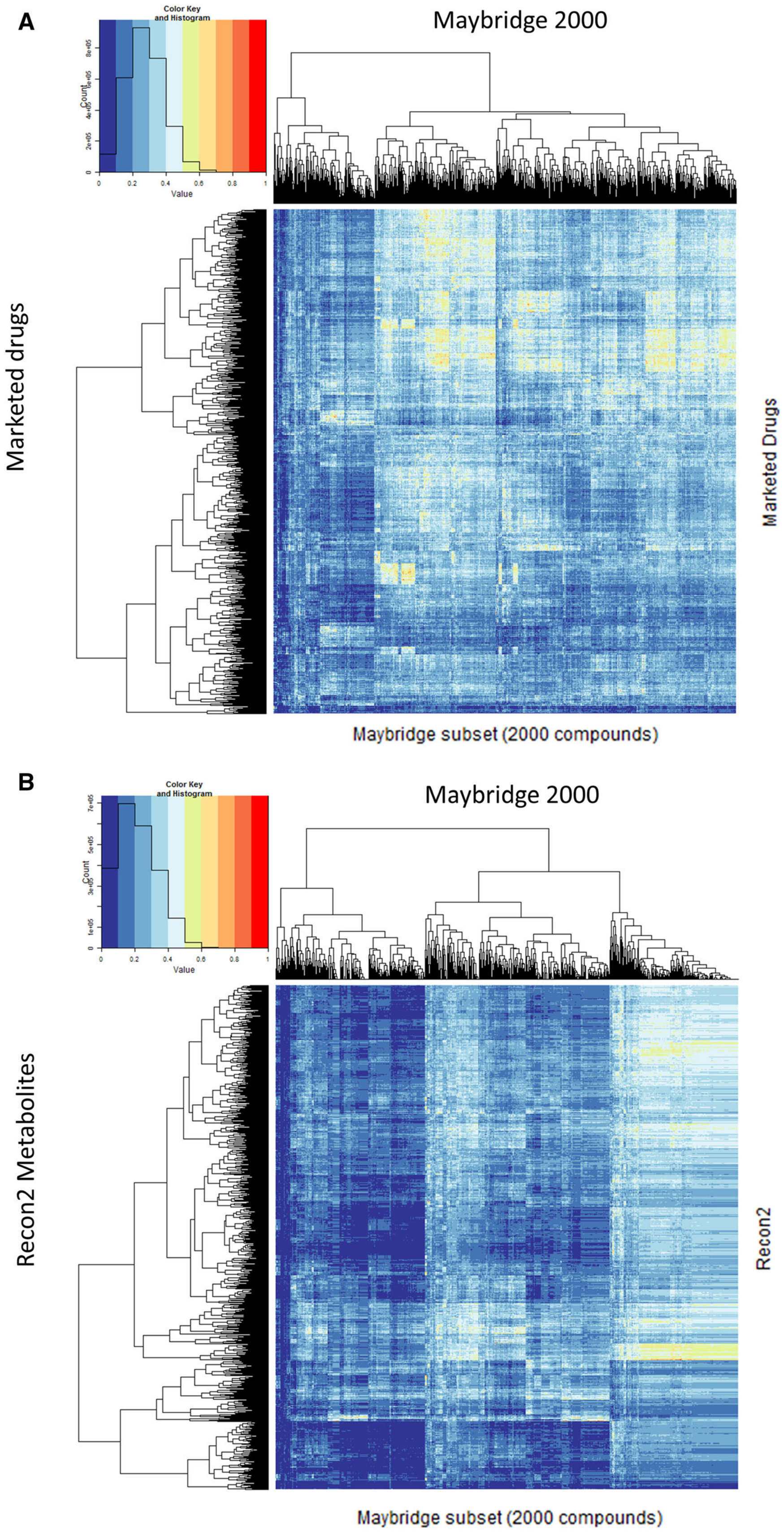
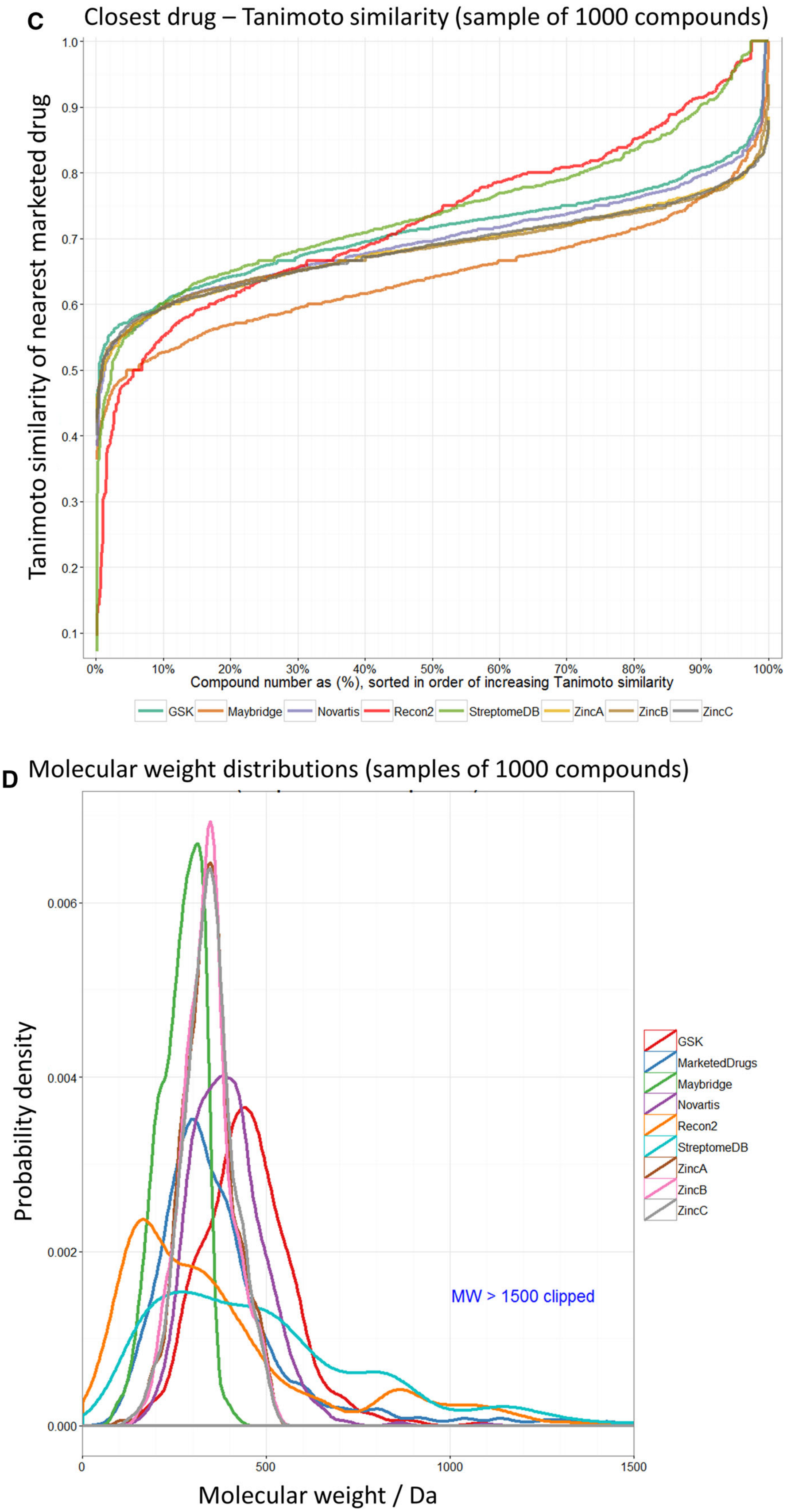


Fig. 5 Properties of drugs and drug fragments. **a** Heat map illustrating marketed drug-compound distances of 2,000 drug fragments selected randomly from a Maybridge library (the plot looks very similar for 15,000 fragments). **b** Heat map illustrating metabolite-compound distances of 2,000 drug fragments selected randomly from a Maybridge library (the plot looks very similar for 15,000 fragments). **c** Cumulative plots of nearest marketed drug-compound or marketed drug–fragment Tanimoto distances for various libraries. **d** Distribution of molecular weights for the various datasets used (Color figure online)

Fig. 5 continued



et al. 2010), from the 5,697 compounds in the Novartis antimalarial collection (Guiguemde et al. 2010) (note that these last two are in fact 'hits' or actives), for 3 subsets of 1,000 molecules from ZINC (Irwin et al. 2012), and of 1,000 from the ~2,400 natural products molecules in StreptomedB (Lucas et al. 2013). We also checked to ensure that we are not biased systematically towards an appearance of metabolite-likeness by say differences in distributions of molecular weights in the different sets, and Fig. 5d shows that we are not, in that a propensity to metabolite-likeness does not seem to follow systematically the MW distribution of the libraries. It is interesting to note that the Novartis and GSK compounds, selected from a very much larger set on the basis of their bioactivity, were even slightly more 'drug-like' than were those from Recon 2 at the left-hand end, though Recon 2 was most drug-like overall (note how it and the streptomycete secondary metabolites 'pull away' from the other curves beyond 50 %, Fig. 5c), and it seems that no such 'MACCS artefact' contributes to the 'rule of 0.5'. Interestingly, Maybridge tends to contain a rather greater diversity of structures relative to human metabolites, but it is possible that the libraries might be enriched further for possible drugs if they were to include a greater degree of metabolite-likeness. It will obviously be of future interest to determine which fragments or compounds are enriched in molecules that happen to possess particular bioactivities.

4 Discussion and conclusions

While both drug and drug target spaces are evidently very heterogeneous (e.g. Adams et al. 2009; Hopkins et al. 2014; Medina-Franco and Maggiora 2014; Paolini et al. 2006), and that is reflected in the analyses presented here, it is highly desirable to be able to find properties that are well represented in marketed (and hence effective and successful) drugs. Given the complexity of drug space, finding a simple mnemonic or rule that has utility is to be welcomed. Indeed, the original 'rule of 5' paper states (Lipinski et al. 1997) "This analysis led to a simple mnemonic which we called the 'rule of 5' because the cutoffs for each of the four parameters were all close to 5 or a multiple of 5….The 'rule of 5' states that: poor absorption or permeation are more likely when: there are more than 5 H-bond donors (expressed as the sum of OHs and NHs); The MWT is over 500; the Log P is over 5 (or M Log P is over 4.15); there are more than 10 H-bond acceptors (expressed as the sum of Ns and Os); compound classes that are substrates for biological transporters are exceptions to the rule." This famous 'rule of 5' (Lipinski et al. 1997) has been highly influential in this regard, but only about 50 % of orally administered new chemical entities actually obey it (Overington et al. 2006; Zhang and Wilkinson 2007) (and see Hopkins et al. 2014); indeed half of recent 'new chemical entities' are natural products (Newman and Cragg 2012), that do not obey the Ro5 either. The (also very effective) 'rule of three' (Congreve et al. 2003) applies solely to leads and not drugs. While improving drug effectiveness is probably best addressed using combinations of molecules (e.g. Small et al. 2011), we have shown that when encoded using the public MDL MACCS keys, more than 90 % of individual marketed drugs obey a 'rule of 0.5' mnemonic, elaborated here, to the effect that a successful drug is likely to lie within a Tanimoto distance of 0.5 of a known human metabolite. While this does not mean, of course, that a molecule obeying the rule is likely to become a marketed drug for humans, it does mean that a molecule that fails to obey the rule is statistically most unlikely to do so. We note that this highlighting of the utility of 'metabolite-likeness' as a concept in drug discovery in systems pharmacology is just a first step, as the availability of Recon2 for such analyses open up many new avenues that we do not discuss here. The present analysis has necessarily been retrospective, as we have applied it to existing and successful (i.e. presently marketed) drugs. However, we consider that this rule, and the concept of the utility of metabolite-likeness more generally, may well have significant prospective value in reversing a current trend in medicinal chemistry (Chen et al. 2012; Walters et al. 2011) that runs in a direction precisely opposite to that of metabolite-likeness.

Acknowledgments D. B. K. conceived the study, N. S. contributed to and extracted the data from Recon2, S. O'. H. constructed all of the workflows, J. H. performed the multi-objective clustering, and all authors wrote the paper. D. B. K. thanks the University of Manchester and the Biotechnology and Biological Sciences Research Council (BBSRC) for financial support. N. S. thanks the BBSRC for funding under Grant BB/K019783/1.

Conflict of interest The authors declare that they have no financial or other conflicts of interest.

References

Adams, J. C., et al. (2009). A mapping of drug space from the viewpoint of small molecule metabolism. *PLoS Computational Biology, 5*, e1000474.

Altman, T., Travers, M., Kothari, A., Caspi, R., & Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics, 14*, 112. doi:10.1186/1471-2105-14-112.

Baldi, P., & Nasr, R. (2010). When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *Journal of Chemical Information and Modeling, 50*, 1205–1222. doi:10.1021/ci100010v.

Beisken, S., Meinl, T., Wiswedel, B., de Figueiredo, L. F., Berthold, M., & Steinbeck, C. (2013). KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinformatics, 14*, 257. doi:10.1186/1471-2105-14-257.

Bender, A. (2010). How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opinion on Drug Discovery, 5*, 1141–1151. doi:10.1517/17460441.2010.517832.

Bender, A., & Glen, R. C. (2004). Molecular similarity: A key technique in molecular informatics. *Organic & Biomolecular Chemistry, 2*, 3204–3218.

Berthold, M. R., et al. (2007). The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Studies in classification, data analysis, and knowledge organization (GfKL 2007)* (pp. 319–326). Heidelberg: Springer.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry, 4*, 90–98.

Brewer, C. A., MacEachren, A. M., Pickle, L. W., & Herrmann, D. (1997). Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers, 87*, 411–438. doi:10.1111/1467-8306.00061.

Caspi, R., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research, 42*, D459–D471. doi:10.1093/nar/gkt1103.

Chen, H. M., Engkvist, O., Blomberg, N., & Li, J. (2012). A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *Medchemcomm, 3*, 312–321.

Congreve, M., Carr, R., Murray, C., & Jhoti, H. (2003). A rule of three for fragment-based lead discovery? *Drug Discovery Today, 8*, 876–877.

de Matos, P., Adams, N., Hastings, J., Moreno, P., & Steinbeck, C. (2012). A database for chemical proteomics: ChEBI. *Methods in Molecular Biology, 803*, 273–296.

Degtyarenko, K., Hastings, J., de Matos, P., Ennis, M. (2009). ChEBI: An open bioinformatics and cheminformatics resource. *Current Protocols in Bioinformatics*. Chapter 14, Unit 14–9.

Dhanda, S. K., Singla, D., Mondal, A. K., & Raghava, G. P. S. (2013). DrugMint: A webserver for predicting and designing of drug-like molecules. *Biology Direct, 8*, 1–12. doi:10.1186/1745-6150-8-28.

Dobson, P. D., & Kell, D. B. (2008). Carrier-mediated cellular uptake of pharmaceutical drugs: An exception or the rule? *Nature Reviews Drug Discovery, 7*, 205–220.

Dobson, P., Lanthaler, K., Oliver, S. G., & Kell, D. B. (2009a). Implications of the dominant role of cellular transporters in drug uptake. *Current Topics in Medicinal Chemistry, 9*, 163–184.

Dobson, P. D., Patel, Y., & Kell, D. B. (2009b). “Metabolite-likeness” as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discovery Today, 14*, 31–40.

Duan, J., Dixon, S. L., Lowrie, J. F., & Sherman, W. (2010). Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling, 29*, 157–170. doi:10.1016/j.jmgm.2010.05.008.

Dunn, W. B., et al. (2014). Molecular phenotyping of a UK population: Defining the human serum metabolome. *Metabolomics, 1*, 18.

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences, 42*, 1273–1280.

Eckert, H., & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discovery Today, 12*, 225–233.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Sciences, 95*, 14863–14868.

Empfield, J. R., & Leeson, P. D. (2010). Lessons learned from candidate drug attrition. *IDrugs, 13*, 869–873.

Everitt, B. S. (1993). *Cluster analysis*. London: Edward Arnold.

Faulon, J.-L., & Bender, A. (Eds.). (2010). *Handbook of chemoinformatics algorithms*. London: CRC.

Feher, M., & Schmidt, J. M. (2003). Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *Journal of Chemical Information and Computer Sciences, 43*, 218–227.

Flower, D. R. (1998). On the properties of bit string-based measures of chemical similarity. *Journal of Chemical Information and Computer Sciences, 38*, 379–386.

Gamo, F. J., et al. (2010). Thousands of chemical starting points for antimalarial lead identification. *Nature, 465*, 305–310.

Gasteiger, J. (2003). *Handbook of Chemoinformatics: From data to knowledge*. Weinheim: Wiley/VCH.

Giacomini, K. M., & Huang, S. M. (2013). Transporters in drug development and clinical pharmacology. *Clinical Pharmacology and Therapeutics, 94*, 3–9. doi:10.1038/clpt.2013.86.

Giacomini, K. M., et al. (2010). Membrane transporters in drug development. *Nature Reviews Drug Discovery, 9*, 215–236.

Gozalbes, R., & Pineda-Lucena, A. (2011). Small molecule databases and chemical descriptors useful in chemoinformatics: An overview. *Combinatorial Chemistry & High Throughput Screening, 14*, 548–558.

Guiguemde, W. A., et al. (2010). Chemical genetics of *Plasmodium falciparum*. *Nature, 465*, 311–315.

Gupta, S., & Aires-de-Sousa, J. (2007). Comparing the chemical spaces of metabolites and available chemicals: Models of metabolite-likeness. *Molecular Diversity, 11*, 23–36.

Hamdalla, M. A., Mandoiu,. I. I., Hill, D. W., Rajasekaran, S., & Grant, D. F. (2013). BioSM: Metabolomics tool for identifying endogenous mammalian biochemical structures in chemical structure space. *Journal of Chemical Information and Modeling, 53*, 601–612. doi:10.1021/ci300512q.

Handl, J., & Knowles, J. (2007). An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation, 11*, 56–76.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21*, 3201–3212.

Hastings, J., et al. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Research, 41*, D456–D463. doi:10.1093/nar/gks1146.

Haug, K., et al. (2013). MetaboLights-an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research, 41*, D781–D786. doi:10.1093/Nar/Gks1004.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., & Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nature Biotechnology, 32*, 40–51. doi:10.1038/nbt.2786.

Herrgård, M. J., et al. (2008). A consensus yeast metabolic network obtained from a community approach to systems biology. *Nature Biotechnology, 26*, 1155–1160.

Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J., & Shoichet, B. K. (2009). Quantifying biogenic bias in screening libraries. *Nature Chemical Biology, 5*, 479–483.

Holdgate, G. A. (2007). Thermodynamics of binding interactions in the rational drug design process. *Expert Opinion on Drug Discovery, 2*, 1103–1114. doi:10.1517/17460441.2.8.1103.

Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C., & Reynolds, C. H. (2014). The role of ligand efficiency metrics in drug discovery. *Nature Reviews Drug Discovery, 13*, 105–121. doi:10.1038/nrd4163.

Huttunen, K. M., Raunio, H., & Rautio, J. (2011). Prodrugs–from serendipity to rational design. *Pharmacological Reviews, 63*, 750–771.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., & Coleman, R. G. (2012). ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling, 52*, 1757–1768. doi:10.1021/ci3001277.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research, 40*, D109–D114. doi:10.1093/nar/gkr988.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research, 42*, D199–D205. doi:10.1093/nar/gkt1076.

Karakoc, E., Sahinalp, S. C., & Cherkasov, A. (2006). Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *Journal of Chemical Information and Modeling, 46*, 2167–2182.

Karp, P. D., & Caspi, R. (2011). A survey of metabolic databases emphasizing the MetaCyc family. *Archives of Toxicology, 85*, 1015–1033. doi:10.1007/s00204-011-0705-2.

Kell, D. B. (2012). Scientific discovery as a combinatorial optimisation problem: How best to navigate the landscape of possible experiments? *BioEssays, 34*, 236–244.

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 Journal, 280*, 5957–5980.

Kell, D. B., Dobson, P. D. (2009). The cellular uptake of pharmaceutical drugs is mainly carrier-mediated and is thus an issue not so much of biophysics but of systems biology. In M. G. Hicks, & C. Kettner (Eds.), *Proceedings of International Beilstein Symposium on Systems Chemistry* (pp. 149–168). Berlin: Logos. http://www.beilstein-institut.de/Bozen2008/Proceedings/Kell/Kell.pdf.

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 Discovery Today, 18*, 218–239.

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 Discovery Today, 16*, 704–714.

Kell, D. B., & Goodacre, R. (2014). Metabolomics and systems pharmacology: Why and how to model the human metabolic network for drug discovery. *Drug Discovery Today, 19*, 171–182.

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

Khanna, V., & Ranganathan, S. (2011). Structural diversity of biologically interesting datasets: A scaffold analysis approach. *Journal of Cheminformatics, 3*, 30. doi:10.1186/1758-2946-3-30.

Knight, C. G., et al. (2009). Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. *Nucleic Acids Research, 37*, e6.

Knox, C., et al. (2011). DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research, 39*, D1035–D1041.

Kola, I. (2008). The state of innovation in drug development. *Clinical Pharmacology and Therapeutics, 83*, 227–230.

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery, 3*, 711–715.

Koutsoukas, A., et al. (2013). How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *Journal of Chemical Information and Modeling*. doi:10.1021/ci400469u.

Landrum, G., Lewis, R., Palmer, A., Stiefl, N., & Vulpetti, A. (2011). Making sure there's a "give" associated with the "take": Producing and using open-source software in big pharma. *Journal of Cheminformatics, 3*(Suppl1), O3.

Lanthaler, K., et al. (2011). Genome-wide assessment of the carriers involved in the cellular uptake of drugs: A model system in yeast. *BMC Biology, 9*, 70.

Law, V., et al. (2014). DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research*. doi:10.1093/nar/gkt1068.

Li, P., Oinn, T., Soiland, S., & Kell, D. B. (2008a). Automated manipulation of systems biology models using libSBML within Taverna workflows. *Bioinformatics, 24*, 287–289. doi:10.1093/bioinformatics/btm578.

Li, P., et al. (2008b). Performing statistical analyses on quantitative data in Taverna workflows: An example using R and maxdBrowse to identify differentially expressed genes from microarray data. *BMC Bioinformatics, 9*, 334.

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. *Advanced Drug Delivery Reviews, 23*, 3–25.

Lucas, X., et al. (2013). StreptomeDB: A resource for natural compounds isolated from *Streptomyces* species. *Nucleic Acids Research, 41*, D1130–D1136. doi:10.1093/nar/gks1253.

Maggiora, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry, 57*, 3186–3204. doi:10.1021/jm401411z.

Maldonado, A. G., Doucet, J. P., Petitjean, M., & Fan, B. T. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity, 10*, 39–79.

Mazanetz, M. P., Marmon, R. J., Reisser, C. B. T., & Morao, I. (2012). Drug discovery applications for KNIME: An open source data mining platform. *Current Topics in Medicinal Chemistry, 12*, 1965–1979.

McGregor, M. J., & Pallai, P. V. (1997). Clustering of large databases of compounds: Using the MDL ''keys'' as structural descriptors. *Journal of Chemical Information and Computer Sciences, 37*, 443–448. doi:10.1021/Ci960151e.

Medina-Franco, J. L., & Maggiora, G. M. (2014). Molecular similarity analysis. In J. Bajorath (Ed.), *Chemoinformatics for drug discovery* (pp. 343–399). Hoboken: Wiley.

Meinl, T., Jagla, B., Berthold, M. R. (2012). Integrated data analysis with KNIME. Open source software in life science research: Practical solutions in the pharmaceutical industry and beyond, pp. 151–171. doi:10.1533/9781908818249.

Muchmore, S. W., Debe, D. A., Metz, J. T., Brown, S. P., Martin, Y. C., & Hajduk, P. J. (2008). Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *Journal of Chemical Information and Modeling, 48*, 941–948.

Newman, D. J., & Cragg, G. M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of Natural Products, 75*, 311–335. doi:10.1021/Np200906s.

Ohno, K., Nagahara, Y., Tsunoyama, K., & Orita, M. (2010). Are there differences between launched drugs, clinical candidates, and commercially available compounds? *Journal of Chemical Information and Modeling, 50*, 815–821. doi:10.1021/ci100023s.

Ooi, H. S., Schneider, G., Lim, T. T., Chan, Y. L., Eisenhaber, B., & Eisenhaber, F. (2010). Biomolecular pathway databases. *Methods and Molecular Biology, 609*, 129–144. doi:10.1007/978-1-60327-241-4_8.

Oprea, T. I. (2004). *Chemoinformatics in drug discovery*. Weinheim: Wiley/VCH.

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? *Journal of Computer-Aided Molecular Design, 21*, 113–119.

Oprea, T. I., Davis, A. M., Teague, S. J., & Leeson, P. D. (2001). Is there a difference between leads and drugs? A historical perspective. *Journal of Chemical Information and Computer Sciences, 41*, 1308–1315.

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery, 5*, 993–996.

Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S., & Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nature Biotechnology, 24*, 805–815.

Papadatos, G., et al. (2010). Lead optimization using matched molecular pairs: Inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *Journal of Chemical Information and Modeling, 50*, 1872–1886. doi:10.1021/Ci100258p.

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

Rafols, I., et al. (2014). Big Pharma, little science? A bibliometric perspective on Big Pharma’s R&D decline. *Technological Forecasting and Social Change, 81*, 22–38. doi:10.1016/j.techfore.2012.06.007.

Riniker, S., & Landrum, G. A. (2013a). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics, 5*, 26. doi:10.1186/1758-2946-5-26.

Riniker, S., & Landrum, G. A. (2013b). Similarity maps—A visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics, 5*, 43. doi:10.1186/1758-2946-5-43.

Sastry, M., Lowrie, J. F., Dixon, S. L., & Sherman, W. (2010). Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *Journal of Chemical Information and Modeling, 50*, 771–784. doi:10.1021/ci100062n.

Saubern, S., Guha, R., & Baell, J. B. (2011). KNIME workflow to assess PAINS filters in SMARTS format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Molecular Informatics, 30*, 847–850. doi:10.1002/minf.201100076.

Sheridan, R. P., Feuston, B. P., Maiorov, V. N., & Kearsley, S. K. (2004). Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences, 44*, 1912–1928. doi:10.1021/ci049782w.

Sheridan, R. P., & Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today, 7*, 903–911.

Small, B. G., et al. (2011). Efficient discovery of anti-inflammatory small molecule combinations using evolutionary computing. *Nature Chemical Biology, 7*, 902–908.

Steinbeck, C., Han, Y. Q., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences, 43*, 493–500.

Stöter, M., Niederlein, A., Barsacchi, R., Meyenhofer, F., Brandl, H., & Bickle, M. (2013). Cell Profiler and KNIME: Open source tools for high content screening. *Methods and Molecular Biology, 986*, 105–122. doi:10.1007/978-1-62703-311-4_8.

Swainston, N., & Mendes, P. (2009). libAnnotationSBML: A library for exploiting SBML annotations. *Bioinformatics, 25*, 2292–2293.

Swainston, N., Mendes, P., & Kell, D. B. (2013). An analysis of a ‘community-driven’ reconstruction of the human metabolic network. *Metabolomics, 9*, 757–764.

Thiele, I., et al. (2013). A community-driven global reconstruction of human metabolism. *Nature Biotechnology, 31*, 419–425.

Todeschini, R., & Consonni, V. (2000). *Handbook of molecular descriptors*. Weinheim: Wiley-VCH Verlag GmbH.

van der Greef, J., & McBurney, R. N. (2005). Rescuing drug discovery: In vivo systems pathology and systems pharmacology. *Nature Reviews Drug Discovery, 4*, 961–967.

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. *Journal of Computer-Aided Molecular Design, 25*, 649–662.

Walters, W. P. (2012). Going further than Lipinski’s rule in drug design. *Expert Opinion on Drug Discovery, 7*, 99–107.

Walters, W. P., Green, J., Weiss, J. R., & Murcko, M. A. (2011). What do medicinal chemists actually make? A 50-year retrospective. *Journal of Medicinal Chemistry, 54*, 6405–6416. doi:10.1021/jm200504p.

Wang, Y., & Bajorath, J. (2010). Advanced fingerprint methods for similarity searching: Balancing molecular complexity effects. *Combinatorial Chemistry & High Throughput Screen, 13*, 220–228.

Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design, 26*, 801–804. doi:10.1007/s10822-012-9577-7.

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today, 11*, 1046–1053.

Wishart, D. S., et al. (2013). HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research, 41*, D801–D807. doi:10.1093/nar/gks1065.

Wunberg, T., et al. (2006). Improving the hit-to-lead process: Data-driven assessment of drug-like and lead-like screening hits. *Drug Discovery Today, 11*, 175–180.

Zhang, M. Q., & Wilkinson, B. (2007). Drug discovery beyond the ‘rule-of-five’. *Current Opinion in Biotechnology, 18*, 478–488.

Zhang, J., Lushington, G. H., & Huan, J. (2011). Characterizing the diversity and biological relevance of the MLPCN assay manifold and screening set. *Journal of Chemical Information and Modeling, 51*, 1205–1215.