frontiers in Pharmacology

ORIGINAL RESEARCH
published: 13 May 2015
doi: 10.3389/fphar.2015.00105

Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites

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

[1] *School of Chemistry, The University of Manchester, Manchester, UK,* [2] *The Manchester Institute of Biotechnology, The University of Manchester, Manchester, UK*

Edited by:
Sara Eyal,
Hebrew University of Jerusalem, Israel

Reviewed by:
Guanglong Jiang,
Indiana University School of Medicine, USA
Peng Hsiao,
Seattle Genetics Inc., USA

Correspondence:
Douglas B. Kell,
Bioanalytical Sciences Group, Faculty of Engineering and Physical Sciences, School of Chemistry and Manchester Institute of Biotechnology, The University of Manchester, 131 Princess St., Manchester M1 7DN, UK
dbk@manchester.ac.uk;
http://dbkgroup.org;
Twitter;@dbkell

Specialty section:
This article was submitted to Drug Metabolism and Transport, a section of the journal Frontiers in Pharmacology

Received: *24 February 2015*
Accepted: *29 April 2015*
Published: *13 May 2015*

Citation:
O'Hagan S and Kell DB (2015) Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites.
Front. Pharmacol. 6:105.
doi: 10.3389/fphar.2015.00105

Background: A recent comparison showed the extensive similarities between the structural properties of metabolites in the reconstructed human metabolic network ("endogenites") and those of successful, marketed drugs ("drugs").

Results: Clustering indicated the related but differential population of chemical space by endogenites and drugs. Differences between the drug-endogenite similarities resulting from various encodings and judged by Tanimoto similarity could be related simply to the fraction of the bitstrings set to 1. By extracting drug/endogenite substructures, we develop a novel family of fingerprints, the Drug Endogenite Substructure (DES) encodings, based on the ranked frequency of the various substructures. These provide a natural assessment of drug-endogenite likeness, and may be used as descriptors with which to derive quantitative structure-activity relationships (QSARs).

Conclusions: "Drug-endogenite likeness" seems to have utility, and leads to a simple, novel and interpretable substructure-based molecular encoding for cheminformatics.

Keywords: drug transporters, cheminformatics, endogenites, metabolomics, encodings

Introduction

In a recent study (O'Hagan et al., 2015), motivated by the recognition that drugs do, and probably have to, hitchhike on metabolite transporters in order to get into cells (Dobson and Kell, 2008; Dobson et al., 2009a,b; Giacomini et al., 2010; Kell et al., 2011, 2013, 2015; Kell, 2013, 2015; Kell and Goodacre, 2014; Kell and Oliver, 2014), we have used the recent availability of a curated reconstruction of the human metabolic network, Recon2 (Swainston et al., 2013; Thiele et al., 2013), to ask the question as to how similar in structural terms marketed drugs are to the molecules (hereafter "endogenites") involved in endogenous human metabolism. While the results depended quite considerably on the exact 2D descriptor used to encode the structures, it was noted that for the commonly used MACCS166 descriptor (Durant et al., 2002; Todeschini and Consonni, 2009) in the implementation described (and see http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs_key_44.html), there was at least one endogenite with a Tanimoto similarity (TS) exceeding 0.5 for more than 90% of marketed drugs. As noted in those references (Durant et al., 2002; Todeschini and Consonni, 2009), the MACCS166 descriptor consists of a string of 166 binary elements representing the presence or absence of 166 (slightly arbitrary and not necessarily druglike) features. We note that not all the

MACCS keys represent substructures, some are rather simple, e.g., "has one or more element [x] atoms." Most of the cheminformatic tool kits (e.g., RDkit, CDKit) are implemented using SMARTS queries; these can only approximate the original MDL MACCS keys. In some cases the intended behavior of the key (query) was ambiguous, in other cases, a SMARTS query is unable to replicate the original MDL query as intended. Nevertheless, the various toolkit MACCS fingerprints are claimed to be sufficiently close to the original MDL versions. The 166 subset were based on the MDL MACCS key that were made public. The RDKit implementation is described at http://rdkit.org/Python_Docs/rdkit.Chem.MACCSkeys-pysrc.html.

It was concluded that 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" (O'Hagan et al., 2015), implying that the degree of endogenite-likeness could indeed be a useful chemical filter in drug discovery programmes. Others too have noted the general "natural metabolite-likeness" of drugs (e.g., Feher and Schmidt, 2003; Karakoc et al., 2006; Gupta and Aires-De-Sousa, 2007; Dobson et al., 2009b; Khanna and Ranganathan, 2009, 2011; Peironcely et al., 2011; Zhang et al., 2011; Chen et al., 2012; Walters, 2012; Hamdalla et al., 2013; Manallack et al., 2013), often using supervised methods of machine learning, though in our own work (O'Hagan et al., 2015), especially to avoid the dangers of overtraining (Broadhurst and Kell, 2006), we purposely confined ourselves to using unsupervised methods only. We also noted (O'Hagan et al., 2015) that a rather smaller fraction of molecules in typical drug discovery libraries obeyed the rule.

Partly for reasons of space, however, the previous study (O'Hagan et al., 2015) left a considerable number of questions rather open. These included, for instance, which fingerprint method might be most "suitable" (and whether "better" ones existed), whether similarity measures should be based on a suitable fusion of the results from using different fingerprints (e.g., Ginn et al., 2000; Hert et al., 2004; Whittle et al., 2006; Gardiner et al., 2009; Chen et al., 2010; Medina-Franco et al., 2011; Willett, 2013a,b), which substructures were most important in determining endogenite-likeness, which parts of metabolite space were most fully populated by drugs, whether results differed markedly if we used other clustering methods, and so on. The purpose of the present paper is to develop and provide some of these analyses. It is concluded that drugs are indeed like metabolites when viewed in a variety of orthogonal ways, and that the substructures found within endogenites and marketed drugs provide a novel and useful means of encoding chemical structures in a simple and easy-to-understand manner. **Figure 1** gives an overview of the paper in the form of a "mind map" (Buzan, 2002).

Materials and Methods

Molecular Data

We used the same molecules for marketed drugs as before (O'Hagan et al., 2015); they were provided in their entirety as Supplementary files to that paper (O'Hagan et al., 2015) and are not reproduced here. The number of endogenites was lowered to 1057 to remove wildcards in lipids with variable chain lengths, since for some purposes we were here specifically interested in molecular weights, but the endogenites were otherwise identical too. Data for Maybridge fragments and Chembridge molecules were downloaded from their respective websites, and other data were downloaded as indicated in the text.

Software

We used the KNIME environment (Berthold et al., 2008; Mazanetz et al., 2012; Meinl et al., 2012) throughout, along with a variety of its cheminformatics toolkits such as CDK (Beisken et al., 2013) and RDKIT (Riniker et al., 2013). Details were as given previously (O'Hagan et al., 2015) (and note that the MACCS fingerprints there were not hashed; a correction has been appended at the journal). Quite a few of the nodes used R code, written by O'Hagan and incorporated into the "R

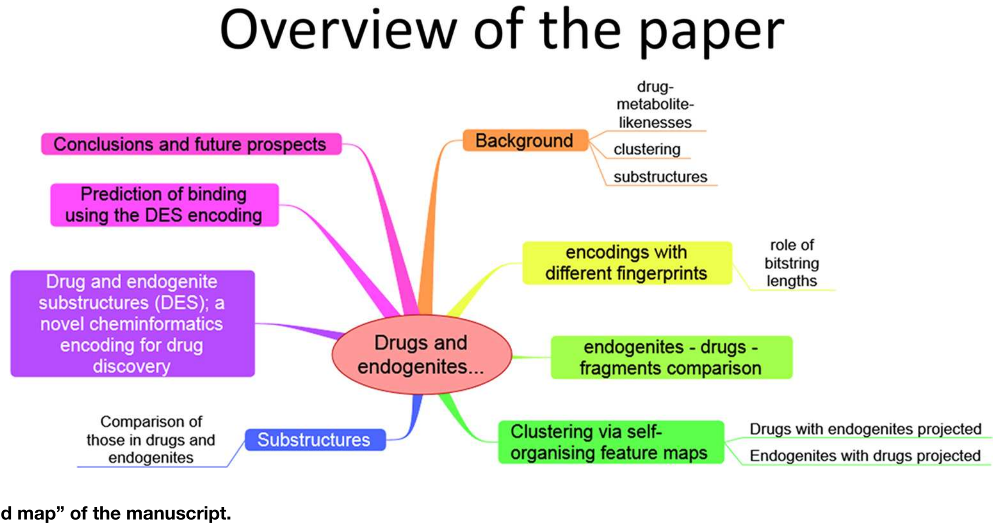

Overview of the paper
Conclusions and future prospects
Prediction of binding using the DES encoding
Drug and endogenite substructures (DES); a novel cheminformatics encoding for drug discovery
Comparison of those in drugs and endogenites
Substructures
Drugs and endogenites...
Background
drug-metabolite-likenesses
clustering
substructures
encodings with different fingerprints
role of bitstring lengths
endogenites - drugs - fragments comparison
Clustering via self-organising feature maps
Drugs with endogenites projected
Endogenites with drugs projected

FIGURE 1 | A "mind map" of the manuscript.

Snippet" KNIME node, with substructure counting via the RDKit Substructure Counter node.

Results and Discussion

Fingerprints

Even (as in O'Hagan et al., 2015) using just 2D fingerprints, the apparent closeness of drug and endogenite molecules to each other (as judged by their Tanimoto similarity coefficients) was differentially "rugged" (the hierarchical clustering showed many more small clusters for drugs than for metabolites), and could differ quite substantially depending on which fingerprint was used (see also e.g., Eckert and Bajorath, 2007; Leach and Gillet, 2007; Faulon and Bender, 2010; Koutsoukas et al., 2014; Maggiora et al., 2014; Medina-Franco and Maggiora, 2014). To explore this further, we decided to compare the drug and metabolite spaces, alone and with each other, using a modification of the approach. Because, of course, the nearest metabolite to itself has a TS of 1, we decided to proceed as follows:

1. For each querying molecule (whether a drug or an endogenite) rank the queried molecules (whether drug or endogenite) and determine the TS of the 90th percentile of closeness.
2. Do this for each fingerprint encoding.

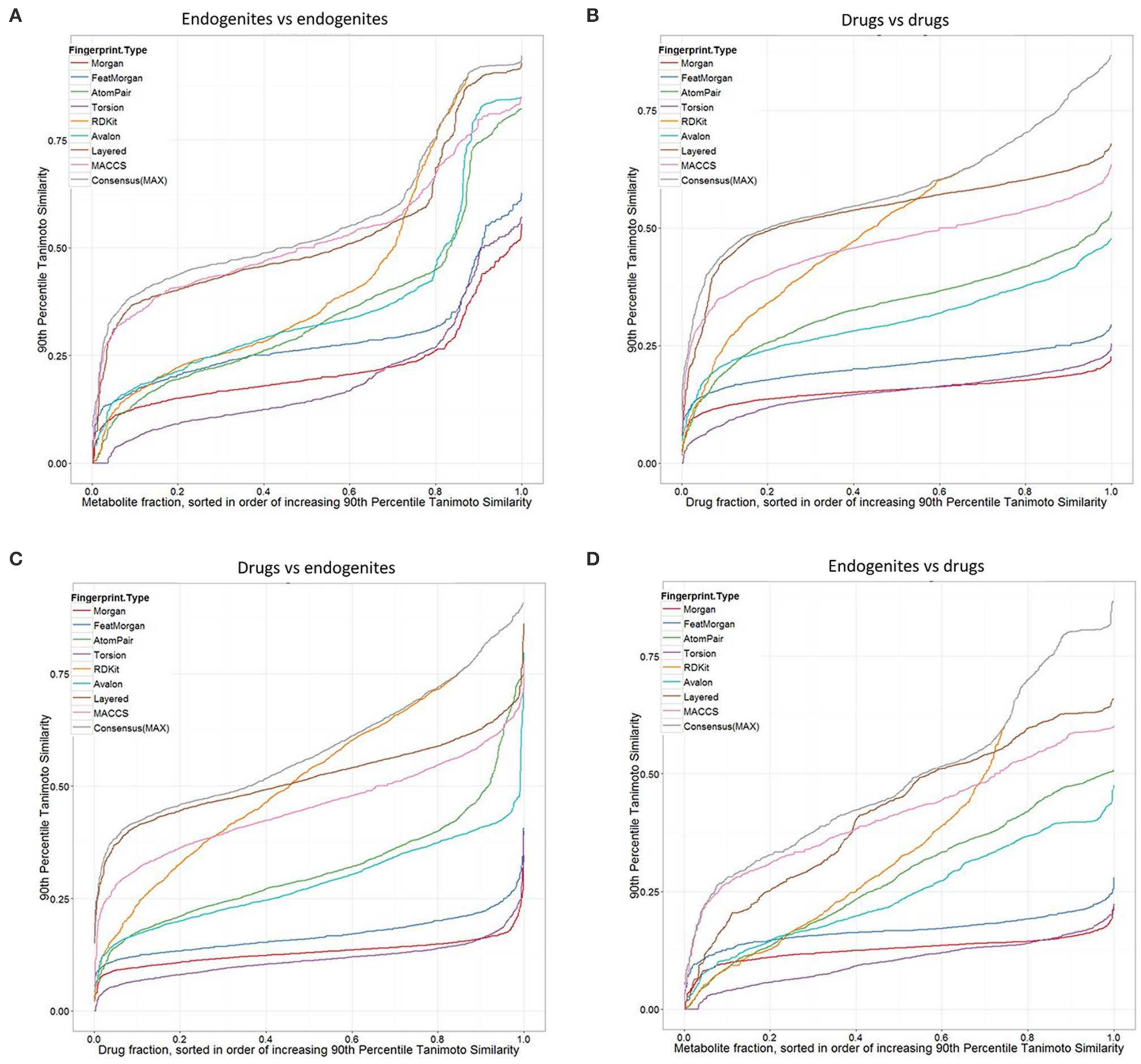


FIGURE 2 | Tanimoto similarities ranked according to the query target closest at the 90th percentile, for different fingerprint types. (A) Endogenites vs endogenites. **(B)** Drugs vs drugs. **(C)** Drugs (query) vs. endogenites. **(D)** Endogenites (query) vs. drugs. Consensus (MAX) refers to the "maximum" TS (Gardiner et al., 2009) among the different fingerprints when only the nearest metabolite is returned.

3. For each query molecule and each queried molecule, find the maximum value of the TS among the eight fingerprints tested.
4. Plot the TS of the 90th percentile of the queried molecule against the fraction of the querying molecules tested.

Considering first the endogenites (as compared to each other), we see (**Figure 2A**) that the RDKIT encoding shows the greatest similarities for metabolites that are *ranked* as being the most similar, but that MACCS and Layered encoding preserve the

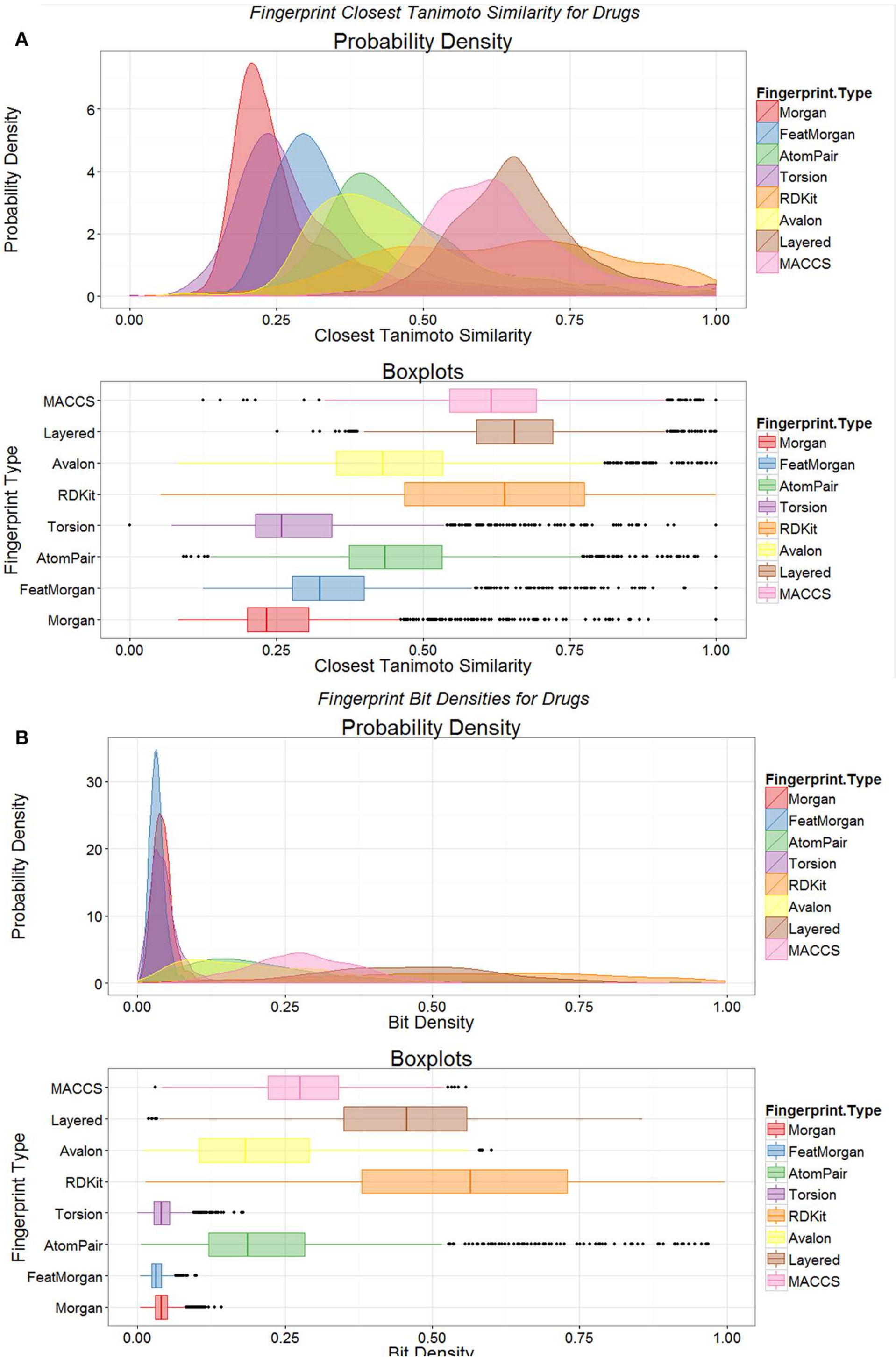


FIGURE 3 | Distributions of fingerprint properties of drugs. (A) Distributions of Tanimoto similarities between drugs and endogenites using eight different encodings, shown as probability densities (upper) and boxplots (lower); the boxplots show the median and interquartile range, with the end of the "whiskers" being at 1.5 times the interquartile range, and with extreme examples being given as dots. **(B)** Variation of the probability density of the number of bits set to 1 in the various encodings in **(A)**.

greater appearances of similarity as the overall similarities decrease. Using these encodings, 40–50% of molecules still had molecules whose TS at the 90 th percentile was 0.5 or above. By contrast (**Figure 2B**), these fractions were uniformly lower for drugs vs drugs, consistent with the rather spikier or "patchy" population of the normalized chemical space relative to that of endogenites (many of which, especially CoA and steroid/sterol derivatives, share many structural similarities) (O'Hagan et al., 2015). The drug-endogenite comparison (**Figure 2C**, with the drugs being the query molecules) gives data broadly similar to those shown in **Figure 2A** of O'Hagan et al. (2015) where closeness to only the very nearest metabolite was plotted, consistent with a view that a querying drug is more commonly close in structural terms not just to a single endogenite but to many such that occupy that part of endogenite space. **Figure 2** also shows the data for the "maximum" TS (Gardiner et al., 2009) among the different fingerprints when only the nearest metabolite is returned. Finally, the complementary endogenite-drug comparison, with the endogenite being the query molecule, shows similar but complementary behavior (**Figure 2D**). One conclusion, given the fact that more than 90% of marketed drugs are seen to be similar to at least some metabolites, and that one might therefore wish to use this as a filter in the analysis of candidate drug libraries, is that for these kinds of comparisons the MACCS, RDKit, Layered or "maximum" fingerprint choice is most likely to return such a result.

Another way of looking at such data is to compare the *distributions* of the nearest Tanimoto similarities between marketed drugs and metabolites for the different encodings (**Figure 3A**). It is clear from such a plot (**Figure 3A**) that not

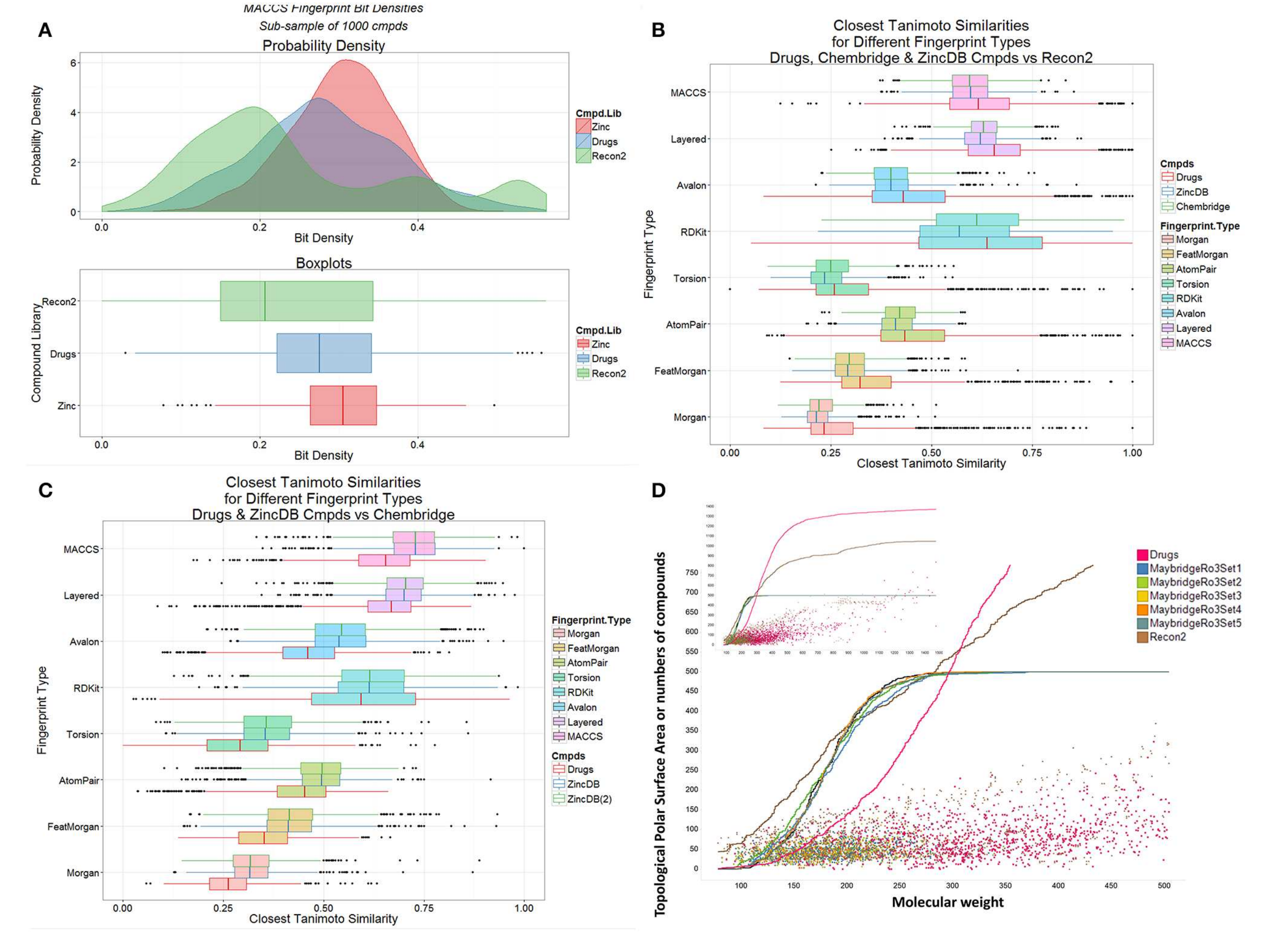
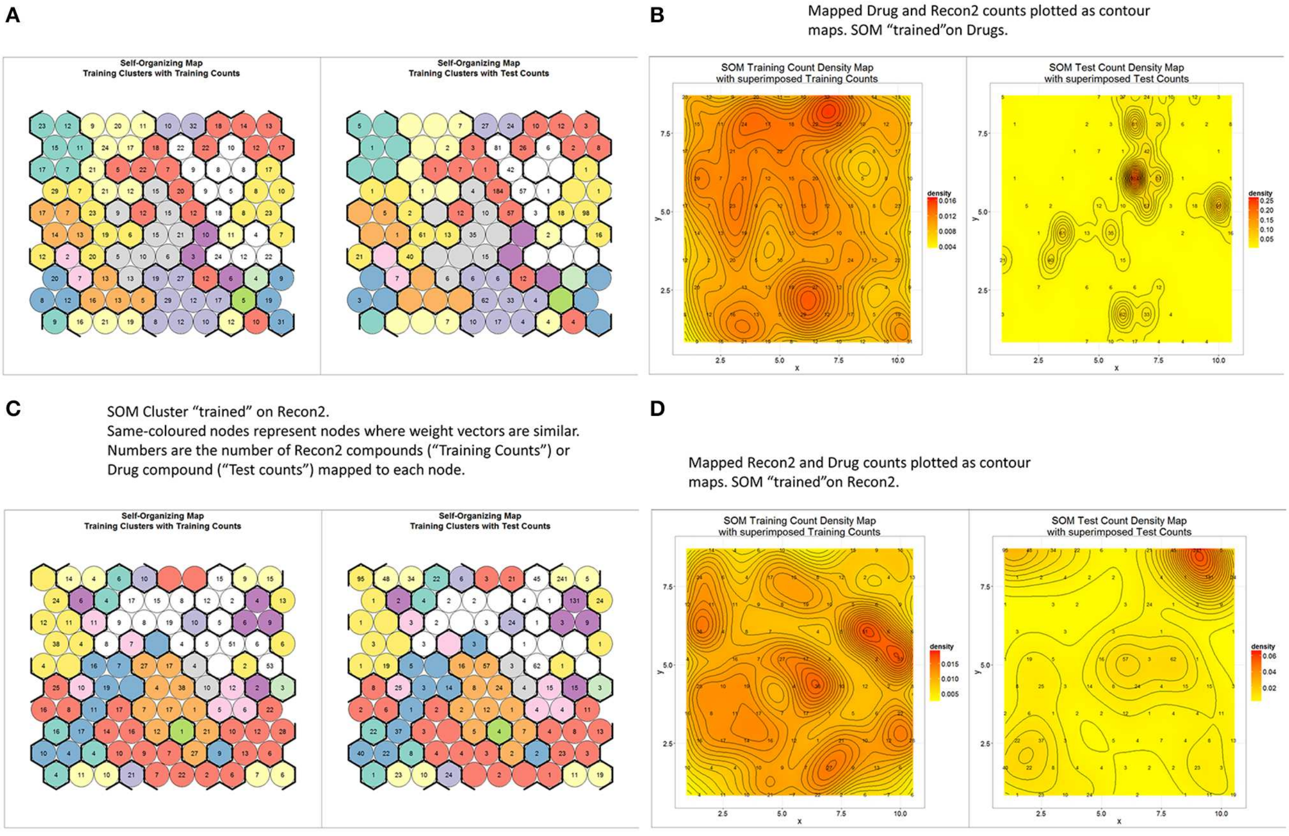


FIGURE 4 | Differences between marketed drugs, Recon2 and library compounds. (A) Variation of bit density for the three classes of compound (based on sampling 1000 of each from the three classes). **(B)** Variation of Tanimoto similarity to Recon2 for eight encodings of marketed drugs and library compounds (from Chembridge and from the ZINC database). In each case drugs are more similar to metabolites than are library compounds. **(C)** Variation of Tanimoto similarity of Chembridge library compounds to two subsets [ZINCDB and ZINCDB(2)] of ZINC database compounds and to marketed drugs. In each case library compounds are more similar to each other than to marketed drugs. **(D)** Topological polar surface area and molecular weight distributions of drugs, Recon2 compounds and five "rule-of-3"-compliant (Congreve et al., 2003) libraries of 500 fragments each that are sold for drug screening purposes. The inset is scaled to show all marketed drugs.

only is the closeness of the "nearest" metabolite different for the different encodings but that the encodings cover metabolite space differentially. At least for the Morgan and Feat Morgan encodings, that resemble ECFP and FCFP (Landrum et al., 2011), this can be ascribed in part to the much smaller number of bits in the encoding that have the value 1 (**Figure 3B**), since the value for the TS is partly a function of this (Flower, 1998; Godden et al., 2000; Holliday et al., 2002, 2003; Wang et al., 2007; Al Khalifa et al., 2009). [In a similar vein, we also looked at the use of a strategy that doubles the length of the bitstring encoding by adding its complement (Knuth, 1986), such that 50% of the bits are 1 and 50% 0. This was not beneficial, as the high density of zeroes in the original merely doubled the number of similar bits (data not shown).]

We also observed previously that the distribution of metabolite- (endogenite-) likenesses differed significantly between marketed drugs and (many of) the kinds of molecules typically found in drug discovery libraries. A convenient way of encoding these is simply to look at the distribution of bitstring densities (of 1 s) for the appropriate encoding between the molecules (Flower, 1998). Thus, **Figure 4A** shows that these differ very significantly for random samples taken from Recon2, from marketed drugs, and from the ZINC (Irwin et al., 2012) databases, with drug candidates typically being less like metabolites than are drugs (see also Chen et al., 2012; Walters, 2012), regardless of the database used (**Figures 4B,C**). The distributions of topological polar surface area (TPSA) and molecular weight (see Abad-Zapatero et al., 2010, 2014) are shown (**Figure 4D**) for endogenites (Recon2), for marketed drugs, and for 5 libraries of small molecule "fragments" (Maybridge "Ro3"-compatible, Congreve et al., 2003, libraries). For a given molecular weight, endogenites are typically significantly more polar than are marketed drugs or fragments, especially for lower molecular weights. Thus, when compounds are ranked by molecular weight (MW), the median MW for drugs, endogenites and fragments are 335, 291, and 179–185 (depending on the library). For these molecules the TPSA values are 69, 124, and 30–69Å^2. A noteworthy point (see also Gopal and Dick, 2014), however, is that fully one quarter of marketed drugs are not in fact larger than typical fragments (**Figure 4D**); indeed when ranked by increasing molecular mass, the 500th marketed drug (of 1383) has a MW of just 297.

We also looked to see whether metabolites that were known substrates (from the Recon2 map) for known transporters (see also Sahoo et al., 2014; Kell et al., 2015) exhibited any greater likelihood to be those with the nearest TS to the query drug;



Mapped Drug and Recon2 counts plotted as contour maps. SOM "trained"on Drugs.

SOM Cluster "trained" on Recon2.
Same-coloured nodes represent nodes where weight vectors are similar.
Numbers are the number of Recon2 compounds ("Training Counts") or Drug compound ("Test counts") mapped to each node.

Mapped Recon2 and Drug counts plotted as contour maps. SOM "trained"on Recon2.

FIGURE 5 | Relationships between endogenite and marketed drug spaces as judged by self-organizing feature maps trained on marketed drugs as encoded with the MACSS encoding. (A) A self-organizing map with 100 nodes and 10 clusters, trained to convergence (3000 iterations) (left), along with a projection of endogenites onto the trained network (right). **(B)** The data in A replotted as a contour plot. **(C)** A plot as in **(A)** but the network was trained using the Recon2 endogenite data. **(D)** Contour plot of the data in **(C)**.

no significant evidence for or against this was found (data not shown), and of course they may be, and may need to be, endogenite-like at their targets too.

Clustering Using Self-organizing Maps

Teuvo Kohonen's Self Organizing (Feature) Map (Kohonen, 1989, 2000; Oja and Kaski, 1998) is a well-known unsupervised learning method of clustering data according to a measure of their similarity. It was therefore of interest to see how "drug" and "endogenite" spaces were organized when represented as such a map. To this end, we used the MACCS encoding for marketed drugs, with 10 × 10 nodes and 10 clusters (numbers chosen to give a reasonable but not excessive degree of clustering, given the number of drugs). **Figure 5A** (left side) shows the distribution of the different numbers of drugs as clustered (by color, based on the similarity of their weight vectors) into the different nodes (circles), while the right hand side of the same figure represents a projection of Recon2 metabolites as projected onto the trained network. The number of circles for each cluster varies quite significantly, from 2 to 15, while the heterogeneous distribution of metabolites shows clearly that some parts of drug space are much less close to multiple metabolites than are others (e.g., the "orange"- and "lemon"-colored clusters). This is especially obvious when the data are displayed as a contour map (**Figure 5B**). In the converse approach, we trained a self-organizing map (SOM) on Recon2; in this case (**Figure 5C**) the number of nodes per cluster varied from 1 to 21, showing again that metabolite space has some significantly larger clusters than does drug space, while the projection of drugs onto metabolite space (**Figure 5D**) shows a highly significant clustering into a particular area of metabolite space, consistent with the finding that there was a significant preference for some metabolites (O'Hagan et al., 2015).

Substructural Basis for Drug-endogenite Likenesses

Our previous analyses of drug-endogenite likenesses looked at the molecules "as a whole." However, it is obvious that some

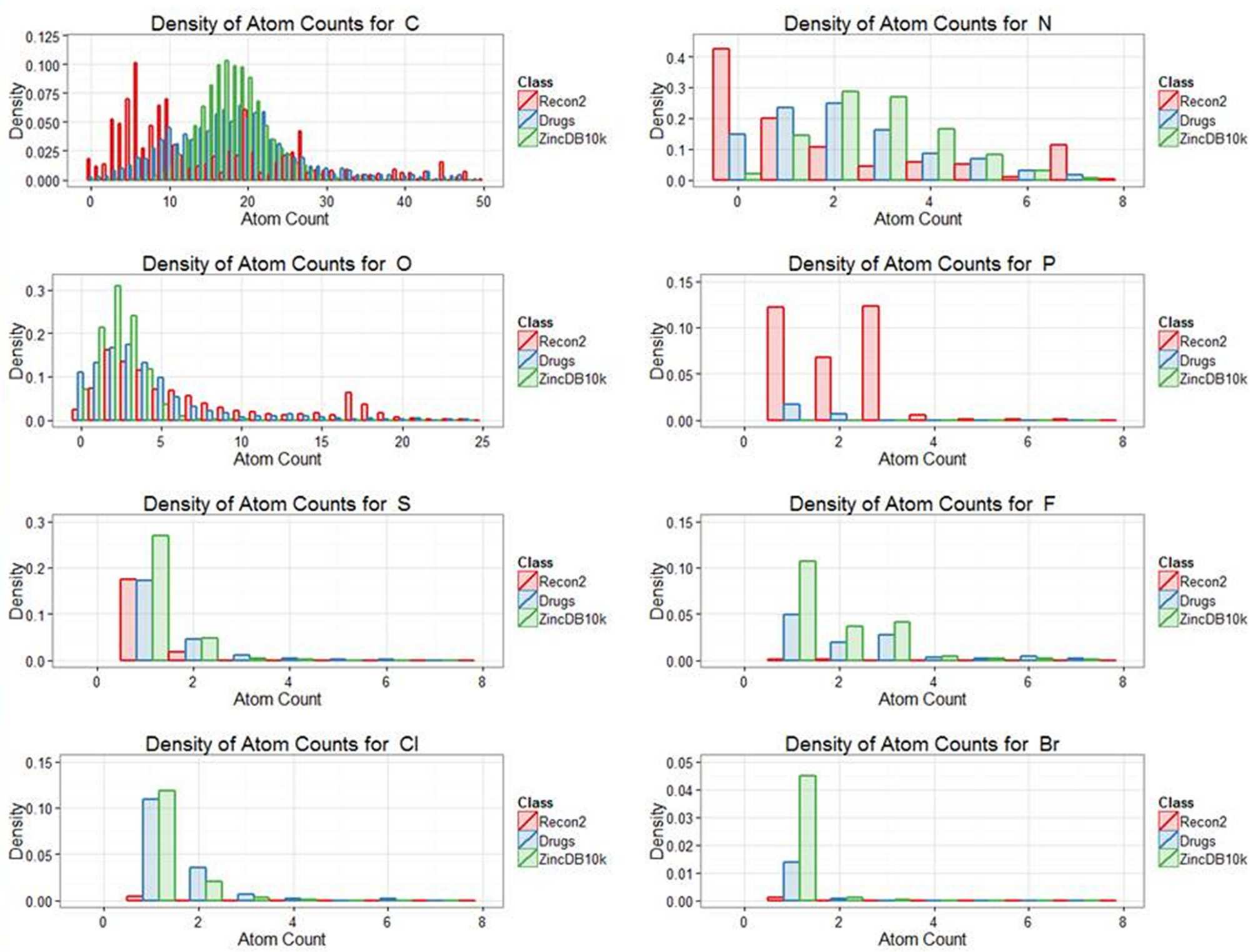


FIGURE 6 | Distribution of the frequency of appearance different atoms between the classes endogenites, marketed drugs and of 10,000 randomly chosen molecules from the ZINC database. To maximize visibility, numbers are not plotted if off the ordinate scale.

substructures may be more common in endogenites than in marketed drugs and vice versa, a simple example being the recognition that human endogenites do not contain halogen atoms while various drugs do (e.g., of the 1381 marketed drugs, 148 of them contain at least one fluorine atom). Thus, **Figure 6** shows the distribution of atom types for the three classes drugs, endogenites, and library compounds.

Starting arguably with (Bemis and Murcko, 1996, 1999), a number of papers have analyzed the frequency of occurrence in FDA-approved, marketed drugs of various substructures, including heterocycles (Vitaku et al., 2014), rings (Aldeghi et al., 2014; Taylor et al., 2014), the chronological (and relatively recent) appearance of S and F in drugs (Ilardi et al., 2014), and even metallodrugs (Mjos and Orvig, 2014). Papers also exist in which fingerprinting methods have been used to *distinguish* drugs from metabolites (e.g., Khanna and Ranganathan, 2009, 2011; Peironcely et al., 2011; Walters, 2012; Hamdalla et al., 2013). However, while Chen et al. (2012) did note that human metabolites and natural products tended to have fewer terminal rings than do marketed drugs, no one has compared the substructures found in marketed drugs with those found in the human endogenites represented in Recon2, which is what we now do here.

Using the Indigo substructure analyser in KNIME, we extracted relevant substructures from both endogenites and marketed drugs, and ranked them according to the normalized frequency of their appearances. The top 60 substructures in each clade are shown in **Figure 7**, while all are illustrated diagrammatically in the inset to **Figure 7A**, with the full Table of data being supplied as Supplementary Information. It is clear from **Figures 7A,B** that while there are indeed some clear similarities between drugs (blue) and endogenites (red) (**Figure 7A**), with a greater frequency of more substructures in drugs (**Figure 7B**), there are also some substantial differences (**Figure 7C**) in the frequency of various substructures between endogenites and present marketed drugs (those substructures that occur frequently in drugs are sometimes referred to as "privileged," Tounge and Reynolds, 2004; Costantino and Barlocco, 2006; Schnur et al., 2006). It is probably also worth noting that in some sense substructures may be related to the fragments that have proved so useful in drug screening (e.g., Hall et al., 2014), and that proposals exist that one might concentrate on those that are metabolite-like (Davies et al., 2009) or natural-product-like (Over et al., 2013).

Use of Drug/endogenite Substructure Presence as an Encoding Strategy

While some encodings, such as MACCS (Durant et al., 2002), use the presence or absence of particular substructures as the basis for their binary scoring, the substructures so chosen are somewhat arbitrary (or at least not necessarily based on any knowledge of the structures of marketed drugs nor endogenites). Armed with the substructures of **Figures 7A–C** (Supplementary Information) we used each of the substructures found (whether in endogenites, drugs or both) as a 1419-bit presence/absence encoding, on the basis that these substructures ought at least to form the basis of useful drug molecules in the future, as they must include or contribute to the concept of "drug-likeness" (Muegge, 2003; Lipinski, 2004; Oprea et al., 2007; Abad-Zapatero et al.,

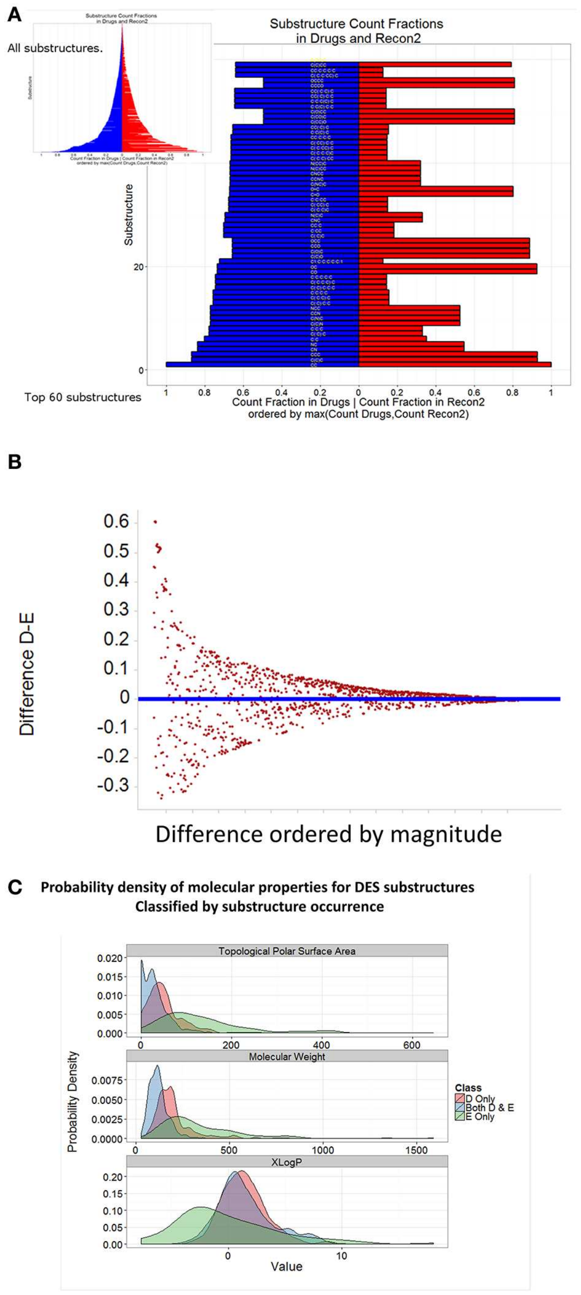


FIGURE 7 | Frequency of representation of different substructures in endogenites and marketed drugs. Self-organizing maps were run as in **Figure 5** for 10 separate occasions. For each SOM node, using the MCS (maximum common scaffold) analyser from Indigo within KNIME, we extracted all substructures for each SOM node; this was performed 10 times, and duplicates removed. **(A)** Substructures were ranked according to their frequency of appearance in either drugs or endogenites, normalized to the total number of either. **(B)** Difference plot of the data in **(A)**. **(C)** Distribution of three properties of drug and endogenite substructures.

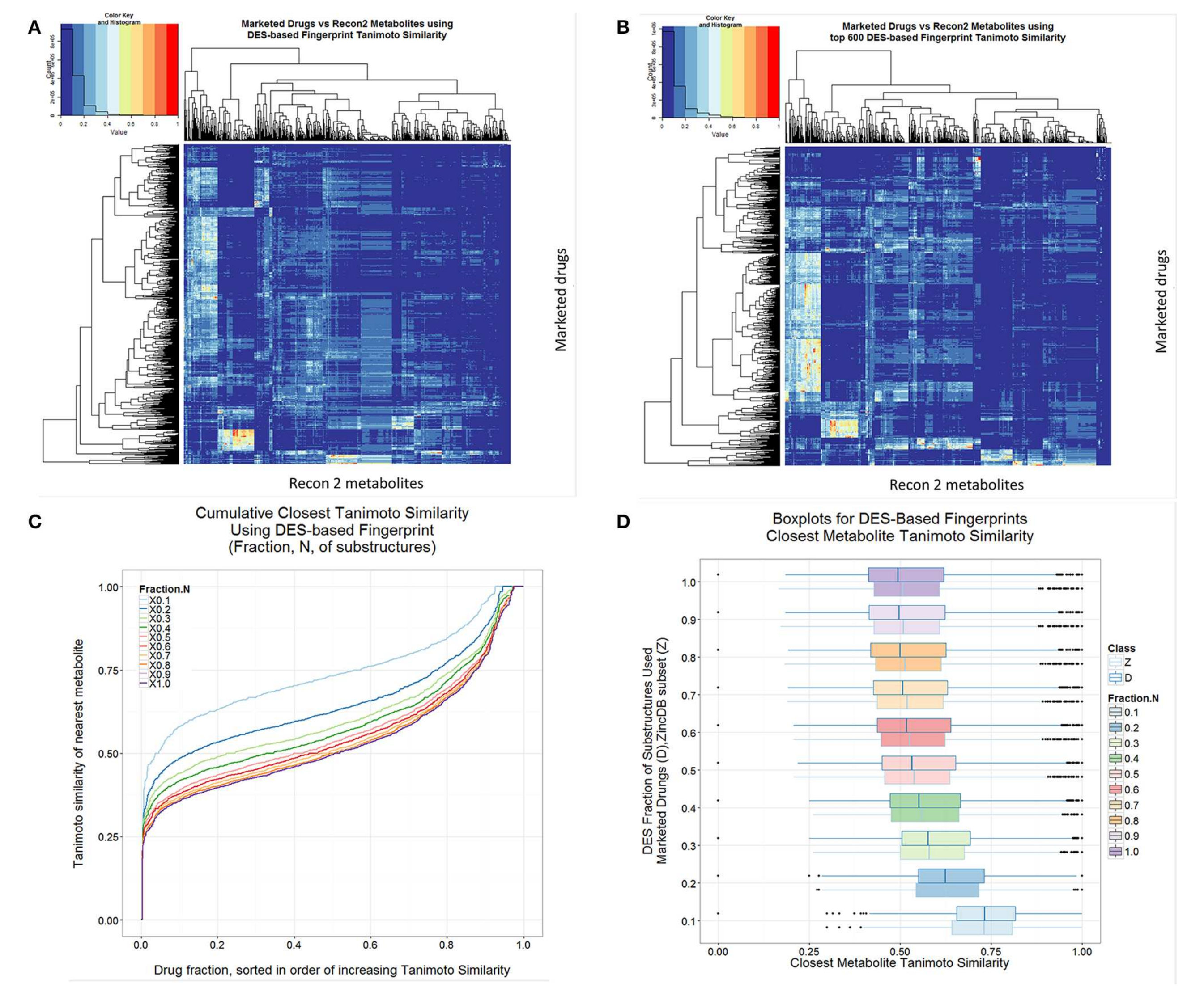


FIGURE 8 | Drug-endogenite similarities as judged using the DES encoding. (A) Heat map of endogenites vs. drugs with full DES encoding. **(B)** As A but with the most frequent 600 substructures ($^{1}\mathrm{DES}_{600}$). **(C)** Cumulative plot of endogenite-likeness using DES encodings based on the fraction of total substructures ordered (from right to left) from the most frequent to least frequent. **(D)** Boxplots of nearest Tanimoto similarities of drugs to endogenites or to ZINC database subsets as the fraction of the DES encoding was varied.

2010, 2014; Camp et al., 2012; Garcia-Sosa et al., 2012; Yusof and Segall, 2013), not least since approved drugs occupy only a rather particular subset of the chemical Universe (Ruddigkeit et al., 2012, 2013). We refer to this encoding as the Drug-Endogenite-Substructure (DES) encoding.

Given its origins and basis, the DES encoding is necessarily likely to indicate more clearly than many encodings the drug-metabolite similarities, and such data are given in **Figure 8**, both for the full set of substructures so extracted (**Figure 8A**) and for truncated versions decreased as per the ranking order in the full Supplementary Information (**Figures 8B–D**). In this case, it is clear that there are advantages in not being too comprehensive, and that using the DES encoding with the top 10% of drug-endogenite substructures results in a drug-endogenite similarity even greater than that found previously [1] using the MACCS encoding; this again would seem to reflect the fraction of bits set to 1 in the bitstring that results from the encoding. This is also true for molecules taken at random from the ZINC database (**Figure 8D**). The KNIME element that calculates the bitstring from the molecular structure encoded in SMARTS strings was mainly written in R, and is provided as Supplementary File 2 (Scaffold2DES-Fingerprint.7z).

Given the supplementary information it is possible to cut substructures from both the most and least frequently found substructures in the list. We suggest that these encodings might also be useful for various purposes, and might usefully be referred to as $^{X}\mathrm{DES}_{Y}$ where X and Y are numbers referring to the first and last of the substructures used. [We note that one might also use something like an evolutionary algorithm for subset selection (e.g., Broadhurst et al., 1997) and other kinds of optimization (Kell and Lurie-Luke, 2015), but as noted above we have chosen to avoid supervised methods for these purposes here.]

A common use of these kinds of encodings is in the calculation of quantitative structure-activity relationships (Geldenhuys et al., 2006; Tropsha, 2010; Stålring et al., 2011; Warr, 2011; Ruusmann et al., 2014). We assessed the ability of the DES and other encodings to predict the binding of various drugs to three candidate targets, using data taken from the internet. Thus, **Figure 9A** shows the out-of-bag prediction from a random forest-based (Breiman, 2001; Svetnik et al., 2003; Knight et al., 2009) QSAR using data on the dopamine D2 receptor downloaded from http://www.bindingdb.org/. In this case we used a random forest learner that was based on the "ensemble tree learner" KNIME node and the full DES encodings, and compared it with the other encodings. The DES encoding was of comparable utility to the other encodings used, although we note that these are log-log plots and that the slope of the lines are rather less than unity, so there would be inaccuracy in linear plots (Kell et al., 2011, 2013; Kell and Oliver, 2014). **Figure 9B** shows the same QSAR, using only the fractions of the DES encodings indicated. Clearly one can learn very effectively using just the commonest 20% of substructures. **Figures 9C,D** show a similar analysis for factor Xa inhibition (Fontaine et al., 2005) using data downloaded from http://www.cheminformatics.org/datasets/, while **Figure 9E** split the data (as did the original authors) into training (out of bag predictions) and test sets as is arguably preferable (Broadhurst and Kell, 2006; Kell and Oliver, 2014). Lastly here (**Figure 9F**), those data were also split into two output classes based on whether the molecule was a "good" or "poor" inhibitor for factor Xa; obviously the DES encoding admits a highly accurate classifier.

Finally, to show the generality of the utility of the new encodings (**Figure 10**), we used the various encodings to devise quantitative structure-activity relationships for two datasets from the ChEMBL bioactivity database (Bento et al., 2014), here using partial least squares (Wold et al., 2001) and the regression error characteristic (Bi and Bennett, 2003; Mittas and Angelis, 2010) to indicate that reasonable predictions could be obtained by methods other than random forests.

Conclusions

The concept of drug-endogenite likenesses continues to appear to have utility, and substructure analyses of drugs and endogenites (for which we provide all the data) show both similarities and differences that have led us to implement here a simple substructure-based cheminformatics encoding

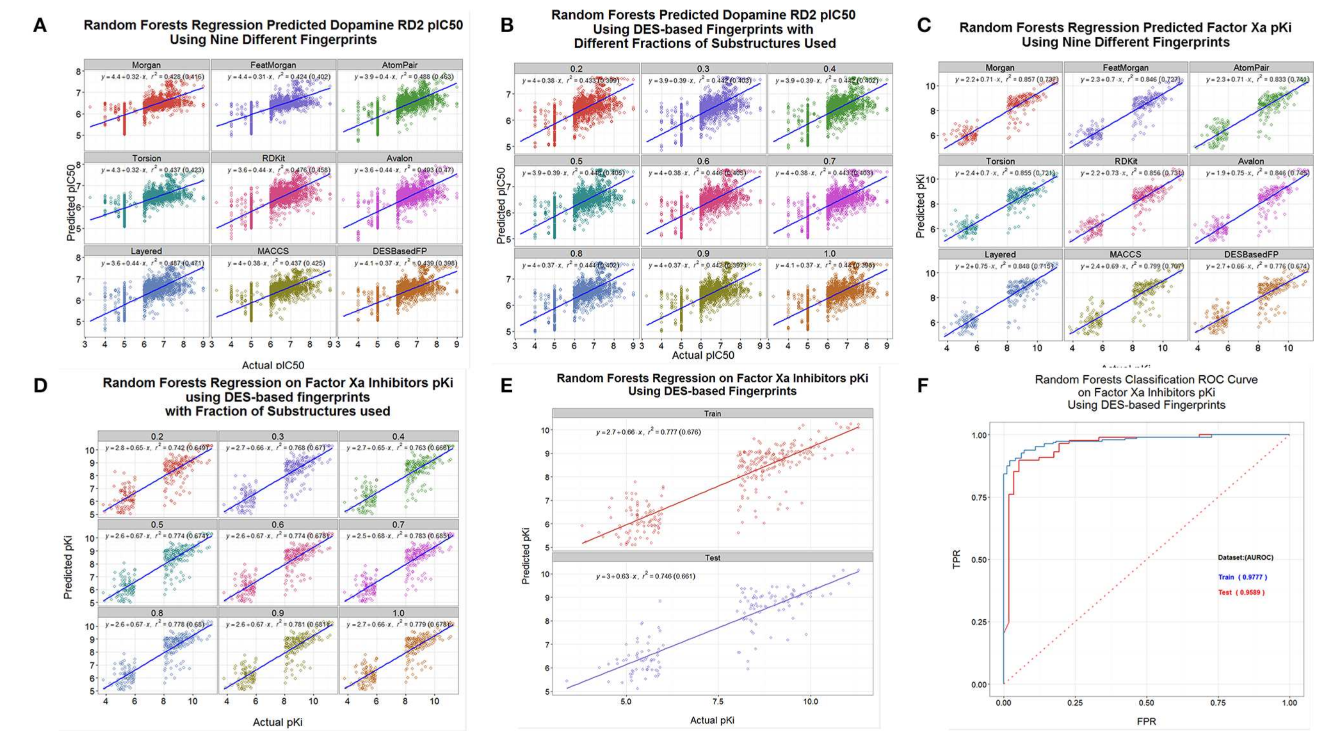


FIGURE 9 | QSAR and classifier analyses of drug binding using various encodings of drug structures. (A) A random forest model was learned using the data for drug binding to the dopamine D2 receptor at http://www.bindingdb.org/bind/ByMonomersTargets.jsp?nBindingData=9349&submit=Search. The out-of-bag predictions were made after 2000 trees were added. **(B)** Same as **(A)** save that we used only the fractions of the DES encodings indicated. **(C)** Same as **(A)** save that the data were for factor Xa inhibition (Fontaine et al., 2005) using data downloaded from http://www.cheminformatics.org/datasets/. **(B)** Same as **(C)** save that we used only the fractions of the DES encodings indicated. **(E)** Same as **(C)** save that data were split into training (out of bag predictions) and test sets as per the data at http://www.cheminformatics.org/datasets/. **(F)** Classification of data (using a Receiver Operator Characteristic curve) from **(C)** to **(D)** based on whether the molecule was a "good" or "poor" inhibitor.

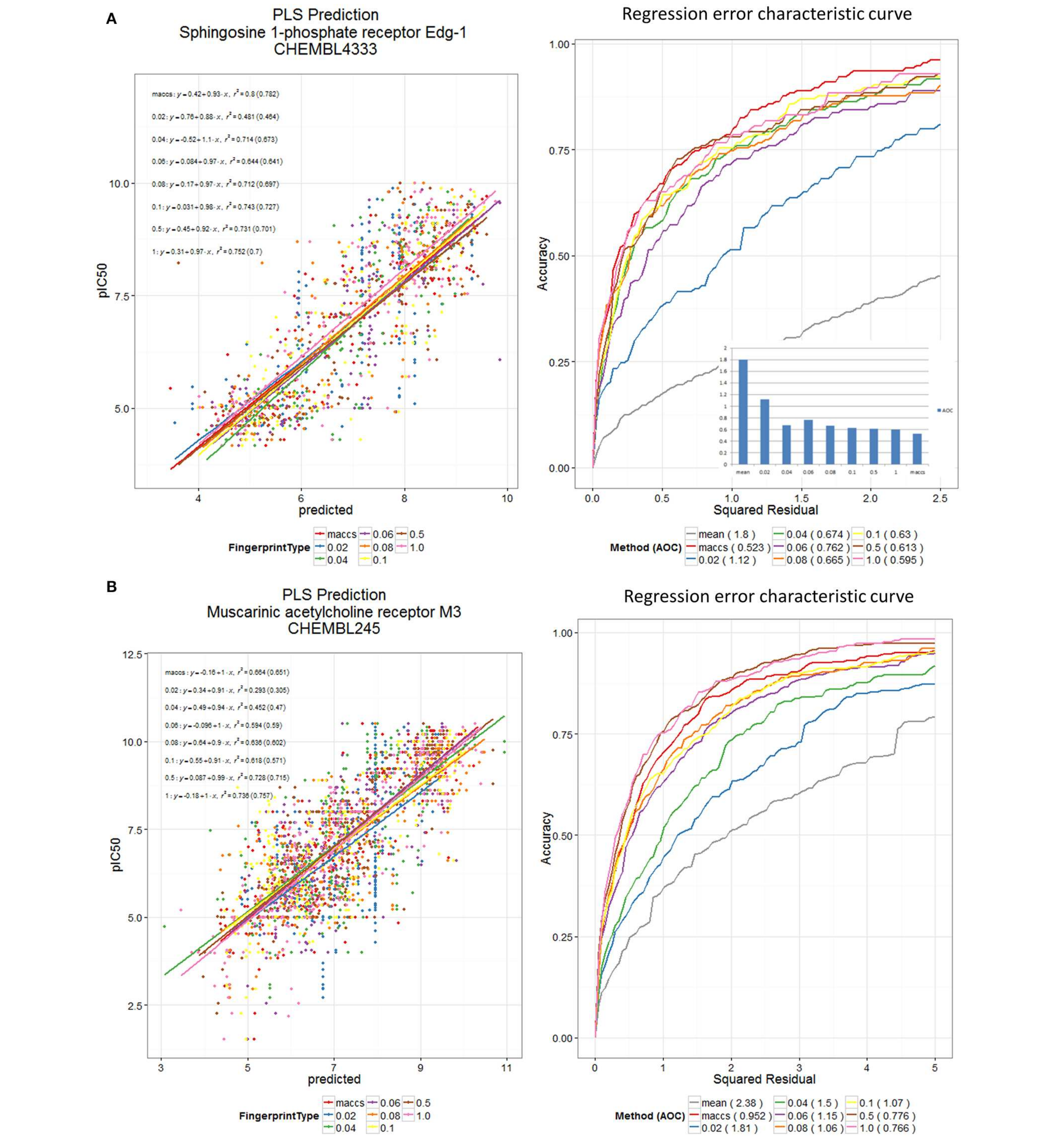


FIGURE 10 | DES and MACCS encodings predict receptor binding in ChEMBL datasets. PLS Prediction of **(A)** the ChEMBL 4333 and **(B)** the ChEMBL 245 dataset pIC50 values using DES Fingerprints (for various fractions, N, of scaffolds used) and MACCS Fingerprints. Scaffolds were sorted according to maximum of their (frequency of occurrence in Drugs, frequency of occurrence in Recon2 metabolites). Datasets were split 60:40 into training and test sets, and training data were pre-processed using a low variance filter, and a correlation filter prior to PLS (5 latent variables). The test data were used for plotting the scatter plot and the REC curve. PLS was carried out using the R plsdepot package using the Knime R Integration and Scripting Nodes. The REC curve plot also shows the curve for using the mean value as predictor; this is taken as a reference worst-case method.

family, DES, that has a clear and interpretable basis. We note a strong tendency for the Tanimoto similarity metric to favor bitstrings (and hence encodings that lead to them) that are highly populated with ones, and this will bear further analysis. However, we anticipate that variants of the DES encoding may provide useful filters for assessing drug- and endogenite-likenesses and for other cheminformatics purposes.

Author Contributions

DBK and SO'H conceived of the study, participated in its design and coordination and helped to draft the manuscript. SO'H wrote the workflows. All authors read and approved the final manuscript.

Authors' Information

DBK is a Research Professor at the University of Manchester, a role to which he returned full time following a 0.8FTE 5-year secondment at Chief Executive of the Biotechnology and Biological Sciences Research Council. He was previously Director of the Manchester Centre for Integrative Systems Biology (www.mcisb.org). His interests include systems biology, chemical biology, pharmaceutical drug transporters, synthetic biology, and iron metabolism. His website is http://dbkgroup.org and he tweets as @dbkell. At Google Scholar his work has been cited more than 30,000 times, with an H-index of 90. SO'H has a Ph.D. in Chemistry from Warwick University, and following a period in industry is now a Computer Officer at the University of Manchester, specializing in cheminformatics, chemometrics, machine learning and the closed-loop automation of scientific instrumentation.

Acknowledgments

DBK thanks the Biotechnology and Biological Sciences Research Council for financial support (grant BB/M017702/1). We thank Dr Neil Swainston for extracting the subset of transporters from Recon 2. This is a contribution from the Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fphar.2015.00105/abstract

Additional Data Files

The following additional data are available with the online version of this paper. Additional data file 1 (VolcanoPlotData.xlsx) lists (in order of abundance) all of the substructures extracted from the endogenites and marketed drugs used herein, for which a truncated version is visualized as **Figure 7**. Additional datafile 2 (Scaffold2DES-Fingerprint.7z)—KNIME node elements for computing the DES encoding(s).

References

Abad-Zapatero, C., Champness, E. J., and Segall, M. D. (2014). Alternative variables in drug discovery: promises and challenges. *Future Med. Chem.* 6, 577–593. doi: 10.4155/fmc.14.16

Abad-Zapatero, C., Perisic, O., Wass, J., Bento, A. P., Overington, J., Al-Lazikani, B., et al. (2010). Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation. *Drug Discov. Today* 15, 804–811. doi: 10.1016/j.drudis.2010.08.004

Aldeghi, M., Malhotra, S., Selwood, D. L., and Chan, A. W. E. (2014). Two-and three-dimensional rings in drugs. *Chem. Biol. Drug Des.* 83, 450–461. doi: 10.1111/cbdd.12260

Al Khalifa, A., Haranczyk, M., and Holliday, J. (2009). Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J. Chem. Inf. Model.* 49, 1193–1201. doi: 10.1021/ci8004644

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

Bemis, G. W., and Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893. doi: 10.1021/jm9602928

Bemis, G. W., and Murcko, M. A. (1999). Properties of known drugs. 2. Side chains. *J. Med. Chem.* 42, 5095–5099. doi: 10.1021/jm9903996

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090. doi: 10.1093/nar/gkt1031

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2008). KNIME: the Konstanz Information Miner. *Stud. Class Data Anal.* 319, 326. doi: 10.1007/978-3-540-78246-9_38

Bi, J., and Bennett, K. P. (2003). "Regression error characteristic curves," in *Proceedings of 20th International Conference on Machine Learning*, eds T. Fawcett and N. Mishra (Washington, DC).

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Broadhurst, D., Goodacre, R., Jones, A., Rowland, J. J., and Kell, D. B. (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Anal. Chim. Acta* 348, 71–86. doi: 10.1016/S0003-2670(97)00065-2

Broadhurst, D., and Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2, 171–196. doi: 10.1007/s11306-006-0037-z

Buzan, T. (2002). *How to Mind Map.* London: Thorsons.

Camp, D., Davis, R. A., Campitelli, M., Ebdon, J., and Quinn, R. J. (2012). Drug-like properties: guiding principles for the design of natural product libraries. *J. Nat. Prod.* 75, 72–81. doi: 10.1021/np200687v

Chen, B. N., Mueller, C., and Willett, P. (2010). Combination rules for group fusion in similarity-based virtual screening. *Mol. Inform.* 29, 533–541. doi: 10.1002/minf.201000050

Chen, H. M., Engkvist, O., Blomberg, N., and 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. doi: 10.1039/C2MD00238H

Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003). A rule of three for fragment-based lead discovery? *Drug Discov. Today* 8, 876–877. doi: 10.1016/S1359-6446(03)02831-9

Costantino, L., and Barlocco, D. (2006). Privileged structures as leads in medicinal chemistry. *Curr. Med. Chem.* 13, 65–85. doi: 10.2174/092986706775197999

Davies, D. R., Mamat, B., Magnusson, O. T., Christensen, J., Haraldsson, M. H., Mishra, R., et al. (2009). Discovery of leukotriene A4 hydrolase inhibitors using metabolomics biased fragment crystallography. *J. Med. Chem.* 52, 4694–4715. doi: 10.1021/jm900259h

Dobson, P. D., and Kell, D. B. (2008). Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat. Rev. Drug Disc.* 7, 205–220. doi: 10.1038/nrd2438

Dobson, P. D., Patel, Y., and Kell, D. B. (2009b). "Metabolite-likeness" as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today* 14, 31–40. doi: 10.1016/j.drudis.2008.10.011

Dobson, P., Lanthaler, K., Oliver, S. G., and Kell, D. B. (2009a). Implications of the dominant role of cellular transporters in drug uptake. *Curr. Top. Med. Chem.* 9, 163–184. doi: 10.2174/156802609787521616

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42, 1273–1280. doi: 10.1021/ci010132r

Eckert, H., and Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* 12, 225–233. doi: 10.1016/j.drudis.2007.01.011

Faulon, J.-L., and Bender, A. (eds.). (2010). *Handbook of Chemoinformatics Algorithms*. London: CRC Press. doi: 10.1201/9781420082999

Feher, M., and Schmidt, J. M. (2003). Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 43, 218–227. doi: 10.1021/ci0200467

Flower, D. R. (1998). On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comp. Sci.* 38, 379–386. doi: 10.1021/ci970437z

Fontaine, F., Pastor, M., Zamora, I., and Sanz, F. (2005). Anchor-GRIND: filling the gap between standard 3D QSAR and the GRid-INdependent descriptors. *J. Med. Chem.* 48, 2687–2694. doi: 10.1021/jm049113+

Garcia-Sosa, A. T., Maran, U., and Hetényi, C. (2012). Molecular property filters describing pharmacokinetics and drug binding. *Curr. Med. Chem.* 19, 1646–1662. doi: 10.2174/092986712799945021

Gardiner, E. J., Gillet, V. J., Haranczyk, M., Hert, J. O., Holliday, J. D., Malim, N., et al. (2009). Turbo similarity searching: effect of fingerprint and dataset on virtual-screening performance. *Stat. Anal. Data Mining* 2, 103–114. doi: 10.1002/sam.10037

Geldenhuys, W. J., Gaasch, K. E., Watson, M., Allen, D. D., and Van Der Schyf, C. J. (2006). Optimizing the use of open-source software applications in drug discovery. *Drug Discov. Today* 11, 127–132. doi: 10.1016/S1359-6446(05)03692-5

Giacomini, K. M., Huang, S. M., Tweedie, D. J., Benet, L. Z., Brouwer, K. L., Chu, X., et al. (2010). Membrane transporters in drug development. *Nat. Rev. Drug Discov.* 9, 215–236. doi: 10.1038/nrd3028

Ginn, C. M. R., Willett, P., and Bradshaw, J. (2000). Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Des.* 20, 1–16. doi: 10.1023/A:1008752200506

Godden, J. W., Xue, L., and Bajorath, J. (2000). Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J. Chem. Inf. Comp. Sci.* 40, 163–166. doi: 10.1021/ci990316u

Gopal, P., and Dick, T. (2014). Reactive dirty fragments: implications for tuberculosis drug discovery. *Curr. Opin. Microbiol.* 21C, 7–12. doi: 10.1016/j.mib.2014.06.015

Gupta, S., and Aires-De-Sousa, J. (2007). Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. *Mol. Divers.* 11, 23–36. doi: 10.1007/s11030-006-9054-0

Hall, R. J., Mortenson, P. N., and Murray, C. W. (2014). Efficient exploration of chemical space by fragment-based screening. *Prog. Biophys. Mol. Biol.* 116, 82–91. doi: 10.1016/j.pbiomolbio.2014.09.007

Hamdalla, M. A., Mandoiu, II, Hill, D. W., Rajasekaran, S., and Grant, D. F. (2013). BioSM: metabolomics tool for identifying endogenous mammalian biochemical structures in chemical structure space. *J. Chem. Inf. Model.* 53, 601–612. doi: 10.1021/ci300512q

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2004). Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comp. Sci.* 44, 1177–1185. doi: 10.1021/ci034231b

Holliday, J. D., Hu, C. Y., and Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screen* 5, 155–166. doi: 10.2174/1386207024607338

Holliday, J. D., Salim, N., Whittle, M., and Willett, P. (2003). Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comp. Sci.* 43, 819–828. doi: 10.1021/ci034001x

Ilardi, E. A., Vitaku, E., and Njardarson, J. T. (2014). Data-mining for sulfur and fluorine: an evaluation of pharmaceuticals to reveal opportunities for drug design and discovery. *J. Med. Chem.* 57, 2832–2842. doi: 10.1021/jm401375q

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 52, 1757–1768. doi: 10.1021/ci3001277

Karakoc, E., Sahinalp, S. C., and Cherkasov, A. (2006). Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model.* 46, 2167–2182. doi: 10.1021/ci0601517

Kell, D. B. (2013). Finding novel pharmaceuticals in the systems biology era using multiple effective drug targets, phenotypic screening, and knowledge of transporters: where drug discovery went wrong and how to fix it. *FEBS J.* 280, 5957–5980. doi: 10.1111/febs.12268

Kell, D. B. (2015). What would be the observable consequences if phospholipid bilayer diffusion of drugs into cells is negligible? *Trends Pharmacol. Sci.* 36, 15–21. doi: 10.1016/j.tips.2014.10.005

Kell, D. B., Dobson, P. D., Bilsland, E., and 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 Discov. Today* 18, 218–239. doi: 10.1016/j.drudis.2012.11.008

Kell, D. B., Dobson, P. D., and Oliver, S. G. (2011). Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only. *Drug Discov. Today* 16, 704–714. doi: 10.1016/j.drudis.2011.05.010

Kell, D. B., and Goodacre, R. (2014). Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. *Drug Discov. Today* 19, 171–182. doi: 10.1016/j.drudis.2013.07.014

Kell, D. B., and Lurie-Luke, E. (2015). The virtue of innovation: innovation through the lenses of biological evolution. *J. R. Soc. Interface* 12, 20141183. doi: 10.1098/rsif.2014.1183

Kell, D. B., and Oliver, S. G. (2014). How drugs get into cells: tested and testable predictions to help discriminate between transporter-mediated uptake and lipoidal bilayer diffusion. *Front. Pharmacol.* 5:231. doi: 10.3389/fphar.2014.00231

Kell, D. B., Swainston, N., Pir, P., and Oliver, S. G. (2015). Membrane transporter engineering in industrial biotechnology and whole-cell biocatalysis. *Trends Biotechnol.* 33, 237–246. doi: 10.1016/j.tibtech.2015.02.001

Khanna, V., and Ranganathan, S. (2009). Physicochemical property space distribution among human metabolites, drugs and toxins. *BMC Bioinformatics* 10:S10. doi: 10.1186/1471-2105-10-S15-S10

Khanna, V., and Ranganathan, S. (2011). Structural diversity of biologically interesting datasets: a scaffold analysis approach. *J. Cheminform.* 3, 30. doi: 10.1186/1758-2946-3-30

Knight, C. G., Platt, M., Rowe, W., Wedge, D. C., Khan, F., Day, P., et al. (2009). Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. *Nucleic Acids Res.* 37, e6. doi: 10.1093/nar/gkn899

Knuth, D. E. (1986). Efficient balanced codes. *IEEE Trans. Inf. Theory* 32, 51–53. doi: 10.1109/TIT.1986.1057136

Kohonen, T. (1989). *Self-Organization and Associative Memory*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-88163-3

Kohonen, T. (2000). *Self-organising Maps*. Berlin: Springer.

Koutsoukas, A., Paricharak, S., Galloway, W. R., Spring, D. R., Ijzerman, A. P., Glen, R. C., et al. (2014). How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inf. Model.* 54, 230–242. doi: 10.1021/ci400469u

Landrum, G., Lewis, R., Palmer, A., Stiefl, N., and Vulpetti, A. (2011). Making sure there's a "give" associated with the "take": producing and using open-source software in big pharma. *J. Cheminform.* 3, O3. doi: 10.1186/1758-2946-3-S1-O3

Leach, A. R., and Gillet, V. J. (2007). *An Introduction to Chemoinformatics, Revised Edn.* Dordrecht: Springer. doi: 10.1007/978-1-4020-6291-9

Lipinski, C. A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1, 337–341. doi: 10.1016/j.ddtec.2004.11.007

Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *J. Med. Chem.* 57, 3186–3204. doi: 10.1021/jm401411z

Manallack, D. T., Dennis, M. L., Kelly, M. R., Prankerd, R. J., Yuriev, E., and Chalmers, D. K. (2013). The acid/base profile of the human metabolome and natural products. *Mol. Inform.* 32, 505–515. doi: 10.1002/minf.201200167

Mazanetz, M. P., Marmon, R. J., Reisser, C. B. T., and Morao, I. (2012). Drug discovery applications for KNIME: an open source data mining platform. *Curr. Top. Med. Chem.* 12, 1965–1979. doi: 10.2174/156802612804910331

Medina-Franco, J. L., and Maggiora, G. M. (2014). "Molecular similarity analysis," in *Chemoinformatics for Drug Discovery*, ed J. Bajorath (Hoboken, NY: Wiley), 343–399.

Medina-Franco, J. L., Yongye, A. B., Perez-Villanueva, J., Houghten, R. A., and Martínez-Mayorga, K. (2011). Multitarget structure-activity relationships characterized by activity-difference maps and consensus similarity measure. *J. Chem. Inf. Model.* 51, 2427–2439. doi: 10.1021/ci200281v

Meinl, T., Wiswedel, B., and Berthold, M. (2012). "Workflow tools for managing biological and chemical data," in *Computational Approaches in Chemiformatics and Bioinformatics*, eds R. Guha and A. Bender (New York, NY: Wiley), 179–209.

Mittas, N., and Angelis, L. (2010). Visual comparison of software cost estimation models by regression error characteristic analysis. *J. Syst. Softw.* 83, 621–637. doi: 10.1016/j.jss.2009.10.044

Mjos, K. D., and Orvig, C. (2014). Metallodrugs in medicinal inorganic chemistry. *Chem. Rev.* 114, 4540–4563. doi: 10.1021/cr400460s

Muegge, I. (2003). Selection criteria for drug-like compounds. *Med. Res. Rev.* 23, 302–321. doi: 10.1002/med.10041

O'Hagan, S., Swainston, N., Handl, J., and Kell, D. B. (2015). A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics* 11, 323–339. doi: 10.1007/s11306-11014-10733-z

Oja, E., and Kaski, S. (eds.). (1998). *Kohonen Maps*. Amsterdam: Elsevier.

Oprea, T. I., Allu, T. K., Fara, D. C., Rad, R. F., Ostopovici, L., and Bologa, C. G. (2007). Lead-like, drug-like or "Pub-like": how different are they? *J. Comput. Aided Mol. Des.* 21, 113–119. doi: 10.1007/s10822-007-9105-3

Over, B., Wetzel, S., Grütter, C., Nakai, Y., Renner, S., Rauh, D., et al. (2013). Natural-product-derived fragments for fragment-based ligand discovery. *Nat. Chem.* 5, 21–28. doi: 10.1038/nchem.1506

Peironcely, J. E., Reijmers, T., Coulier, L., Bender, A., and Hankemeier, T. (2011). Understanding and classifying metabolite space and metabolite-likeness. *PLoS ONE* 6:e28966. doi: 10.1371/journal.pone.0028966

Riniker, S., Fechner, N., and Landrum, G. A. (2013). Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. *J. Chem. Inf. Model.* 53, 2829–2836. doi: 10.1021/ci400466r

Ruddigkeit, L., Blum, L. C., and Reymond, J. L. (2013). Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* 53, 56–65. doi: 10.1021/ci300535x

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J. L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52, 2864–2875. doi: 10.1021/ci300415d

Ruusmann, V., Sild, S., and Maran, U. (2014). QSAR DataBank - an approach for the digital organization and archiving of QSAR model information. *J. Cheminform.* 6, 25. doi: 10.1186/1758-2946-6-25

Sahoo, S., Aurich, M. K., Jonsson, J. J., and Thiele, I. (2014). Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease. *Front. Physiol.* 5:91. doi: 10.3389/fphys.2014.00091

Schnur, D. M., Hermsmeier, M. A., and Tebben, A. J. (2006). Are target-family-privileged substructures truly privileged? *J. Med. Chem.* 49, 2000–2009. doi: 10.1021/jm0502900

Stålring, J. C., Carlsson, L. A., Almeida, P., and Boyer, S. (2011). AZOrange - High performance open source machine learning for QSAR modeling in a graphical programming environment. *J. Cheminform.* 3, 28. doi: 10.1186/1758-2946-3-28

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

Swainston, N., Mendes, P., and Kell, D. B. (2013). An analysis of a 'community-driven' reconstruction of the human metabolic network. *Metabolomics* 9, 757–764. doi: 10.1007/s11306-013-0564-3

Taylor, R. D., Maccoss, M., and Lawson, A. D. G. (2014). Rings in drugs. *J. Med. Chem.* 57, 5845–5859. doi: 10.1021/jm4017625

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., et al. (2013). A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31, 419–425. doi: 10.1038/nbt.2488

Todeschini, R., and Consonni, V. (2009). *Molecular Descriptors for Cheminformatics.* Weinheim: WILEY-VCH Verlag GmbH. doi: 10.1002/9783527628766

Tounge, B. A., and Reynolds, C. H. (2004). Defining privileged reagents using subsimilarity comparison. *J. Chem. Inf. Comp. Sci.* 44, 1810–1815. doi: 10.1021/ci049854j

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol. Informat.* 29, 476–488. doi: 10.1002/minf.201000061

Vitaku, E., Smith, D. T., and Njardarson, J. T. (2014). Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among U.S. FDA approved pharmaceuticals. *J. Med. Chem.* 57, 10257–10274. doi: 10.1021/jm501100b

Walters, W. P. (2012). Going further than Lipinski's rule in drug design. *Exp. Opin. Drug Disc.* 7, 99–107. doi: 10.1517/17460441.2012.648612

Wang, Y. A., Eckert, H., and Bajorath, J. (2007). Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem* 2, 1037–1042. doi: 10.1002/cmdc.200700050

Warr, W. A. (2011). Some trends in Chem(o)informatics. *Meth. Mol. Biol.* 672, 1–37. doi: 10.1007/978-1-60761-839-3_1

Whittle, M., Gillet, V. J., Willett, P., and Loesel, J. (2006). Analysis of data fusion methods in virtual screening: theoretical model. *J. Chem. Inf. Model.* 46, 2193–2205. doi: 10.1021/ci049615w

Willett, P. (2013a). Combination of similarity rankings using data fusion. *J. Chem. Inf. Model.* 53, 1–10. doi: 10.1021/ci300547g

Willett, P. (2013b). Fusing similarity rankings in ligand-based virtual screening. *Comput. Struct. Biotechnol. J.* 5: e201302002. doi: 10.5936/csbj.201302002

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab Syst.* 58, 109–130. doi: 10.1016/S0169-7439(01)00155-1

Yusof, I., and Segall, M. D. (2013). Considering the impact drug-like properties have on the chance of success. *Drug Discov. Today* 18, 659–666. doi: 10.1016/j.drudis.2013.02.008

Zhang, J., Lushington, G. H., and Huan, J. (2011). Characterizing the diversity and biological relevance of the MLPCN assay manifold and screening set. *J. Chem. Inf. Model.* 51, 1205–1215. doi: 10.1021/ci1003015

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.