ELSEVIER

'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries

Paul D. Dobson, Yogendra Patel and Douglas B. Kell

School of Chemistry and The Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK

Present drug screening libraries are constrained by biophysical properties that predict desirable pharmacokinetics and structural descriptors of 'drug-likeness' or 'lead-likeness'. Recent surveys, however, indicate that to enter cells most drugs require solute carriers that normally transport the naturally occurring intermediary metabolites and many drugs are likely to interact similarly. The existence of increasingly comprehensive summaries of the human metabolome allows the assessment of the concept of 'metabolite-likeness'. We compare the similarity of known drugs and library compounds to naturally occurring metabolites (endogenites) using relevant cheminformatics molecular descriptor spaces in which known drugs are more akin to such endogenites than are most library compounds.

Introduction

The search for pharmaceutically active drugs with desirable properties and negligible side effects can be considered as a multi-objective optimisation problem over an enormous search space of 'possible' drugs [1,2]. It is usual to start the search by looking for hits and then leads [3], because, according to Oprea *et al.* [4], 'lead structures exhibit, on the average, less molecular complexity (less MW, less number of rings and rotatable bonds), are less hydrophobic (lower cLogP and LogD), and less druglike' than actual drugs (see also [5]). The process of optimising a lead into a drug with favourable ADMET properties [6,7] results in more complex structures [8] and system approaches [9–13] that consider not only a molecular target but also biochemical networks may be of value in understanding why.

In seeking to narrow the search space of chemically diverse candidate compounds, cheminformatic methods are used to constrain the compounds screened such that they tend to display 'lead-likeness' [4,14–16] or 'drug-likeness' [16–23] (and even 'CNS-likeness' [24]). The same concepts hold true for drugs with multiple intended targets (promiscuous drugs [25] or poly-pharmacology [2,26]).

The most common cheminformatic filter used to constrain pharmaceutical drug libraries is Lipinski and colleagues' celebrated 'rule of five' (Ro5) [27]. This states that poor absorption or permeation of a compound is more probable when there are more than five hydrogen-bond donors, the molecular mass is above 500 Da, the lipophilicity is high (*clogP* > 5) and when the sum of nitrogen and oxygen atoms is greater than 10. Other rules or filters consider generic and calculable properties such as the number of rotatable bonds and the polar surface area [28,29] or the ligand efficiency [30–33], and a 'rule of three' has been proposed [34] for fragment-based lead discovery (see e.g. [35,36]). It was recognised explicitly in the original review [27] that the Lipinski rules do not normally cover drugs that are derived from natural products [37,38], in which transporters are clearly involved in their disposition and it is, in fact, probable that this involvement of carrier molecules holds true for most other compounds too [39–41].

Descriptors such as those of Lipinski and colleagues [27] are essentially biophysical rather than structural in nature, and despite the widespread use of these measures it is not completely obvious how they should be understood mechanistically, given the enormous structural diversity of both drugs and libraries. Clearly, if drugs are mainly transported by carriers, this gives a ready explanation of why general descriptors are not normally going to be entirely effective in individual cases [41]; it also promotes the view that we need to understand the specificities for existing and candidate drugs of known drug transporters much better than we do now at a mechanistic level. Indeed, it is considered difficult to design for the use of active transporters because transporter selectivity is not well understood [42]. This said, at

Corresponding author: Dobson, P.D. (paul.dobson@manchester.ac.uk)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2008.10.011

least for those uptake transporters that are known, it is reasonable to assume that they normally exist to transport common biochemical compounds involved in primary (intermediary) metabolism and the knowledge of these molecules may provide useful constraints for identifying other potential substrates by direct structural comparison or by SAR.

A drug may also interact with its target or targets in a manner that emulates native substrate binding (e.g. [43]). This will also constrain drugs towards metabolite space, given that the number of known protein folds is comparatively restricted and that once found they appear to be conserved in evolution [44,45]. Similarly, drugs may regulate activity allosterically at a native control site and this will also constrain such drugs towards metabolite-likeness. The guiding principle is that interacting with targets in any native-like manner imposes native-like constraints, and this will manifest itself in the drug discovery process by a tendency to develop towards the regions of chemical space occupied by metabolites, although other concerns in the drug development pipeline will also impinge upon the final drug structure(s) chosen.

Much of post-genomic drug discovery has concentrated on proteins and proteinaceous drug targets [19,46]. In recent years, however, we have witnessed the development of freely available, curated, reconstructed genome-scale metabolite networks (e.g. [47–51]) and of databases of human and other metabolites [52–59], where here the term 'metabolite' is used to refer to small molecule components of primary metabolism and not the products of the reaction of drugs with drug metabolising enzymes. To this end, we shall sometimes use the term 'endogenite' in this article to describe these endogenous, naturally occurring molecules. Nobeli and colleagues [60,61] have produced a very interesting summary of some of the properties of the known metabolome of *Escherichia coli* in particular (and we note that many microbially derived gut metabolites may also influence their human host (e.g. [62])). In a similar vein, the existence of databases of endogenite molecules allows us to ask the question as to whether existing drugs, that is those that have been successful in passing through the various phases of drug discovery to the marketplace, are more metabolite-like (i.e. endogenite-like) than are the typical contents of pharmaceutical screening libraries. To address this, known drugs and library compounds, representing the sorts of pre-drugs that might be screened in hit discovery, are compared to human metabolites in a variety of appropriate molecular descriptor spaces. We find that drugs are indeed considerably more similar to endogenous metabolites than are library compounds, and conclude that endogenite-likeness might be a useful filter in the design and analysis of pharmaceutical libraries for drug discovery.

Related comparisons between metabolites and other types of molecules have been considered previously. Gupta and Aires-de-Sousa [63] compared the distributions in chemical space of metabolites drawn from KEGG and compounds from the supplier library ZINC [64], concluding that discriminatory features include hydroxyl groups, aromatic systems and molecular weight when combined with other global descriptors.

In the major analysis of Karakoc *et al.* [65] relationships between drugs, drug-like compounds, antimicrobials, and human and bacterial metabolites were considered. One result finds that bacterial metabolites and antibiotics are highly similar and this mirrors the similarity between human metabolites and drugs we observe. There is also the suggestion, however, that human metabolites form a distinct class of molecules that are unlike bacterial metabolites, drugs or drug-like molecules, and they occupy a separate region of chemical space. This seemingly counter-intuitive result does not concur with that presented here. The set of 5333 metabolites used here is much larger (compared to 1104), giving far greater coverage of 'metabolite space' and so more fully represents the total diversity of human metabolites. Moreover, the redundancy measures of Karakoc *et al.* only removed exact duplicate molecules and this allows highly similar molecules to remain within the set. Inevitably this biases the set's properties towards the properties of over-represented molecules. Indeed their own analysis indicates the over-abundance of scaffolds drawn from sugar- and nucleotide-like molecules. Through the application of clustering to choose representative molecules for multiple represented 'types' of molecules the influence of redundancy within our sets is negated. We suggest that the differences observed between human metabolites and other classes actually reflect the construction of their human metabolite set and not a fundamental difference between the properties of human metabolites and other classes of molecule, and this is reinforced by our analysis. Also note that we do not claim that all human metabolites are similar to drugs; many clearly are not. If the human metabolite set of Karakoc *et al.* contains many of these (sugar scaffolds are prevalent among their metabolites but not their drugs) then this could also underpin the differences observed.

Finally, Ganesan [38] has very recently compared natural products and synthetic molecules released as drugs, in terms of their 'Lipinski-likeness', commenting that (only) 'half of the 24 natural products lie in what can be called the ''Lipinski universe''.'

In this article, the metabolites are drawn from human-specific databases and genome-scale metabolic reconstructions, and are greater in number than in previous studies, although many of the carbohydrates and especially lipids [66,67] that might usefully be considered metabolites in this context still remain undetermined. This said, it emerges that the types of drugs that exhibit virtually no metabolite-likeness in our analysis are very atypical and will probably remain so even when the missing metabolites are included.

Comparing drugs and library compounds to metabolites

To assess the relationship between drugs and metabolites we compare against a background of compounds of the kinds that typically make up screening collections for hit discovery, which we refer to as library compounds. These represent pre-drugs and can be considered as starting points for drug discovery and development. During these processes candidate drugs are selected and modified to enhance properties favourable to drug action and our hypothesis suggests that this optimisation process drives such starting molecules towards the regions of chemical space occupied by metabolites, because of the necessity to participate in native-like reactions (including those with transporter molecules).

In the analysis we therefore distinguish metabolites (endogenites), drugs and library compounds (Table 1). The molecules retrieved from source databases contained duplicate records and over-represented structural types. Thus, to avoid [68] biasing the

TABLE 1

Sources of drug, metabolite and pre-drug structures.

Class	Source	Compounds	Total
Metabolite	HMDB [58]	2835	5333 (5560)
Metabolite	Palsson [94]	806	
Metabolite	BioCyc [95]	772	
Metabolite	BiGG [96]	698	
Metabolite	Edinburgh [48,51]	2048	
Drug	DrugBank [97]	4152	7330 (8002)
Drug	KEGG Drug [98]	4435	
Pre-drug	Zinc [64]	62,390	

'Compounds' is the number of unique structures after washing and filtering. 'Total' is the total number of unique compounds for each class. The figures in brackets are before the semi-automatic correction of annotation errors in sources.

analysis towards common structural forms, the compounds were clustered and cluster centres used as representatives. Physicochemical distributions are shown in Fig. 1a–d. Figure 1a illustrates (i) that the distribution of the number of atoms among metabolites is markedly different from that of drug and library compounds, but (ii) that the similarity between drugs and library compounds suggests that screening sets do cover drug space in a sensible manner, at least with respect to atom number.

The distribution of *clogD* [69], a calculated value of lipophilicity able to account for charged species, is shown in Fig. 1b. Positive values indicate a preference for a hydrophobic environment and negative values a preference for a hydrophilic environment. The difference between metabolites and drugs is clear, and again the library compound distribution conforms to the drug distribution and suggests appropriate representation of existing drug space in the screening set. That drugs and libraries have similar distributions is expected because considerations of lipophilicity have played a major role in designing drugs and libraries with useful bioavailability, with the Ro5 being particularly influential, despite the fact that only approximately half of the marketed drugs obey it [70]. The fact, however, that metabolites are in general much more hydrophilic than are both drugs and library compounds is very

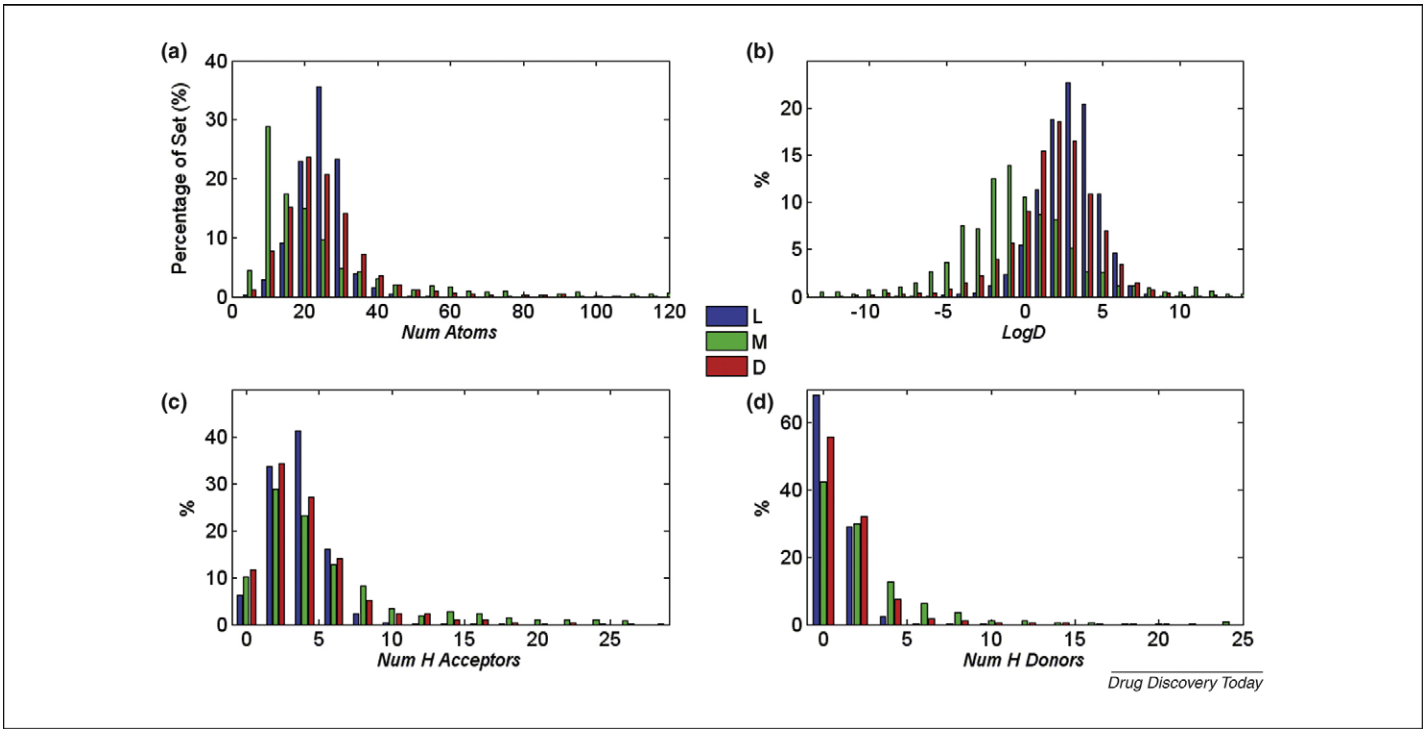

FIGURE 1

Histograms, normalised by class sizes, for different simple molecular properties. **(a)** Number of atoms (excluding hydrogens) over the drug, library and metabolite representative sets. Of note is the fact that above 40 atoms there are very few library compounds, although drugs and metabolites are still represented. **(b)** *clogD*, a lipophilicity calculation that takes charge into account. The distribution of library compounds is similar to that of drugs, but metabolites differ markedly. **(c)** Number of hydrogen bond acceptors. The number of hydrogen bond acceptors in Lipinski's rule of five is suggested to be not more than ten. The histogram illustrates that both known drugs and metabolites are still found above this number, although the library compounds are rare, suggesting that Lipinski and related measures of 'drug-likeness' have been used as a constraint in library design. **(d)** Number of hydrogen bond donors. The number of hydrogen bond donors in Lipinski's rule of five is suggested to be not more than five. As in (c), library compounds largely follow the rule, but there are many drugs and metabolites that do not.

noticeable, and is at least consistent with the requirement for specialised carriers to transfer them into and out of cells and between intracellular compartments.

Figure 1c and d shows the distributions of the numbers of hydrogen bond donors and acceptors. The Ro5 suggests that the number of hydrogen bond acceptors be not more than ten, and the number of hydrogen bond donors be not more than five. Both figures illustrate that whilst the library sets mostly follow these suggestions, there are numerous drugs (and metabolites) that do not. From this perspective, endogenites are considerably more like drugs than are library compounds.

Metabolite-likeness curves

Whilst physicochemical distributions provide a general overview of the relationships between types of molecules, it is their distributions in chemical space that are of most relevance. If the processes of drug discovery and development drive towards metabolite-likeness then this should manifest itself through considerable overlap between the distributions of drugs and (a subset of) metabolites, and one greater than seen between library compounds and metabolites. The notion of chemical space is abstract, but it can be represented and operated on by the techniques of cheminformatics that allow molecular similarity to be quantified.

The similarity of drug and library compounds to metabolites is here assessed by calculating the Tanimoto distance to the closest metabolite. A variety of molecular descriptors were computed, and similarities calculated using the Tanimoto coefficient [71]. The molecular descriptors used were connectivity fingerprints [72,73], paths [74,75], MDL Public Keys [76] and electrotopological state (E-state) keys [77,78].

Representative sets of drugs and library compounds were calculated in each space at thresholds that removed high-level redundancy, which will be described later. Redundancy within the metabolites was not addressed because it does not negatively affect the outcome because similarity is measured only with the closest metabolite.

Figure 2a–d shows the proportion of drugs and library compounds within a given distance to the closest metabolite, using the above four sets of molecular descriptors. For example, in Fig. 2a one can determine that 12% of drugs have a Tanimoto distance of 0.5 or less to their closest metabolite. By contrast, less than 2% of library compounds fall within the same threshold. Although the shapes of the curves vary for the different descriptors used, the drugs are consistently closer, often considerably so, to endogenous metabolites than are the contents of typical screening libraries. That the drug curves are consistently higher than the curves for library compounds, in a variety of descriptor spaces covering various ways of assessing molecular similarity, indicates that successful, marketed drugs are indeed much more like metabolites than are the typical library compounds.

Molecular similarity can be represented in different ways. Because of this, metabolite-likeness is calculated in several molecular descriptor spaces that capture different aspects of structure, and so illustrate that metabolite-likeness is not simply an artefact of a particular descriptor but a general phenomenon in each of the chemical spaces assessed. It appears to be generally true that drugs that are very close to metabolites are typically analogues of the native substrate of their targets. For example, Fig. 3 illustrates how the closest metabolite differs in the various spaces using the example query of atorvastatin (Lipitor). Different metabolites are retrieved in each space, and whilst there are features common to the query and each of the retrieved structures, the connectivity fingerprint-retrieved structure particularly recalls the native product structure (mevalonate) of the main atorvastatin target HMG-CoA reductase (although we note that statins can exhibit many pleiotropic effects, see e.g. [79–81]). Generally, the closest metabolite to a drug is quantifiably more similar than in the example shown in Fig. 3.

The valyl-ester prodrug of ganciclovir (valganciclovir), which is taken up by peptide transporters [82] of solute carrier family 15 [83], retrieves nucleoside-like metabolites that more closely resemble the active drug than do the valine modification that one might expect, although the relative contributions of the large drug and small valine probably bias molecular similarity measures towards the drug, and maximal common substructure methods may be of use. Another type of prodrug modification couples bile acids and drugs [84], including those designed to target the human apical sodium-dependent bile acid transporter (hASBT) [85], which transports bile acids including chenodeoxycholate, deoxycholate, cholate and ursodeoxycholate. By coupling bile acids via valine to acyclovir, enhanced uptake was observed *in vitro* and *in vivo*, most successfully for the prodrug acyclovir valylchenodeoxycholate, which lead to a twofold increase in acyclovir bioavailability in rats. Using acyclovir valylchenodeoxycholate as the drug query in the metabolite search all spaces retrieve bile acids (taurochenodeoxycholate) or intermediates in bile acid biosynthesis (choloyl-CoA). This emphasises that metabolite-likeness can be because of drugs mimicking metabolites in a pathway as opposed to those interacting with a specific target.

Dissimilar drugs

Whilst drugs are generally more similar to metabolites than to library compounds, certain drugs do not conform to this trend. The fraction that does not depends upon how one chooses to define the boundary between 'similar' and 'dissimilar'. In the connected fingerprint space a realistic choice for the limit of molecular similarity is a Tanimoto distance between 0.7 and 0.8, equating to 50–80% of drugs being metabolite-like. An illustrative selection of these 'remote' compounds is shown in Fig. 4, but clear trends towards particular types of drug or structural classes are not immediately discernable, although many remote compounds are heavily halogenated or sulphurated.

Discussion

That the processes of drug discovery and development lead largely to regions of chemical space already occupied by metabolites, although a novel discovery, is both expected from the arguments given in the introduction and observed experimentally in our analyses. This has major implications for future library design, which might beneficially take account of the structures and properties of endogenous metabolites now that usefully complete structural metabolomes are available. Of course, further efforts to elucidate the measured metabolome are ongoing, but it is of note that many metabolites observed experimentally have yet to be identified chemically [86–88], particularly lipids [66,89,90].

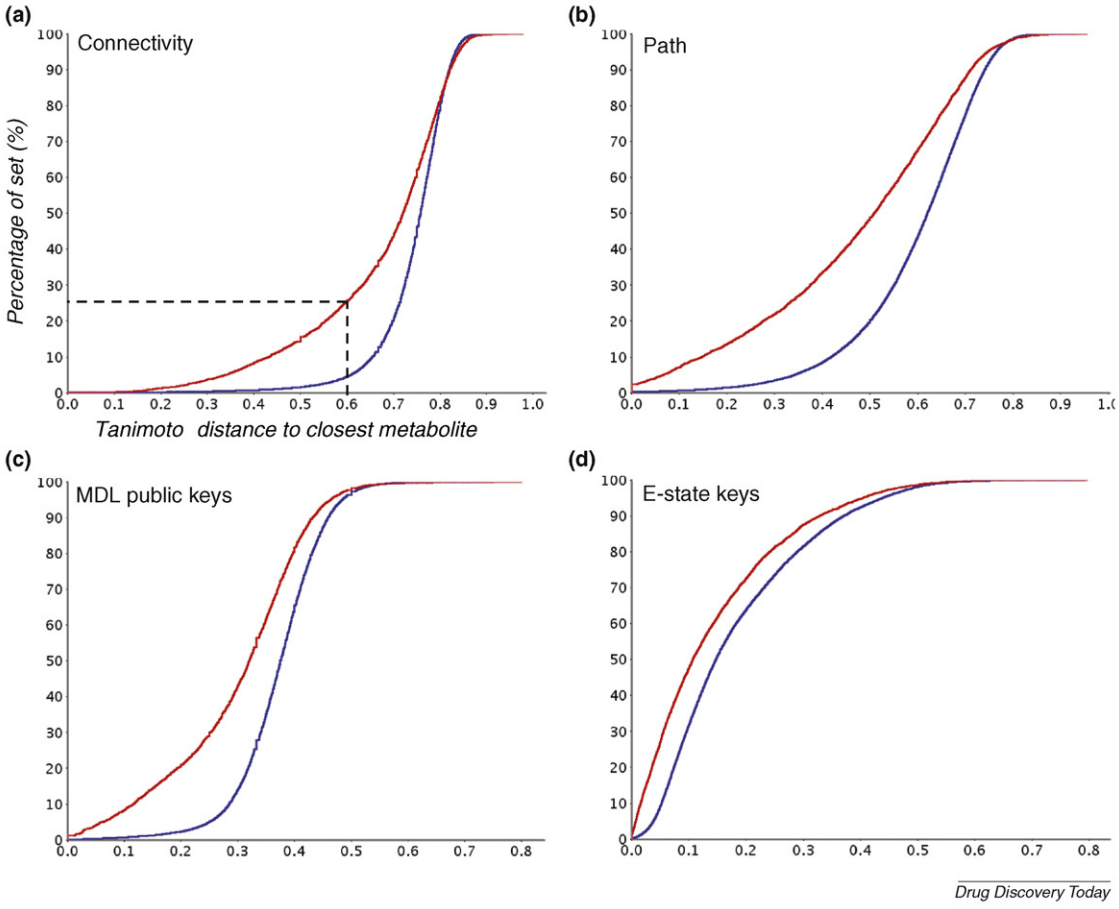

FIGURE 2

A comparison of drug and library distances to closest metabolites in various molecular descriptor spaces. **(a)** Connectivity fingerprint space. The proportion of the representative sets that lie within the specified Tanimoto distance can be seen for each class; the dashed line illustrates that 25% of drugs have a distance to their closest metabolite of less than or equal to 0.6, but fewer than 5% of library compounds are found within the same threshold closeness. **(b)** Path descriptor space. In this space some drugs and metabolites are equivalent, these being 2% of drugs with a Tanimoto distance to their closest metabolite of 0. Only 3.5% of library compounds are within a distance of 0.2 of a metabolite, compared with 22% of drugs. **(c)** MDL Public Keys space. The closeness of both kinds of compound to the nearest metabolite (endogenite) using these descriptors is numerically rather smaller than that using the extended connectivity fingerprint space (a) but again the distances are considerably smaller for endogenites than for library compounds. Only 2% of library compounds are within a Tanimoto distance of 0.2 of a metabolite, compared with 20% of drugs. **(d)** E-state keys space. The closeness of both kinds of compound to the nearest metabolite (endogenite) using these descriptors is numerically rather smaller than in (a) and (c), but the distances to closest endogenites are again considerably lower for drugs than for library compounds. Only 6% of library compounds are within a Tanimoto distance of 0.05 of a metabolite, compared with 23% of metabolites.

Understanding why drugs are more similar to metabolites than library compounds are is complex. High levels of similarity imply considerable structural equivalence between drug and metabolite molecules, whereas lower similarity levels might suggest that something more general has been derived that relates the molecules, such as similar physicochemical properties, but it is difficult to generalise any such interpretation over the set. The conclusion that drugs are more like metabolites than library molecules, as demonstrated using multiple molecular descriptor spaces, however, is clear and will, we believe, be of considerable value in shaping future drug discovery efforts. Unlike drugs, metabolites have not been through a human-guided development process that considers important factors such as developability, desirable PK/PD properties and other concerns that dictate how hits become leads and then drugs. This is, however, information that should be inductively learned by a well-constructed model of drug-likeness [19]. Conversely, such models are largely bounded by the limits of existing information and only describe the types of drugs already known. The search for new types of drug that exist outside the current drug space (see e.g. [70]) might usefully begin in the unexplored regions of metabolite space, particularly given that drugs probably do not cover the whole of metabolite space, our knowledge of which continues to grow. Furthermore, the spaces in which drugs exhibit enhanced metabolite-likeness will be of use in predicting drug–metabolite interactions.

For problems of molecular similarity it is important to consider the scope of descriptors and the extent to which they are able to capture useful relationships between molecules. As with any other kind of clustering, where utility is the most significant criterion [91], molecular similarity has no innate or absolute meaning and the appropriateness of one descriptor ahead of another is largely a subjective choice, and this problem of representation is well

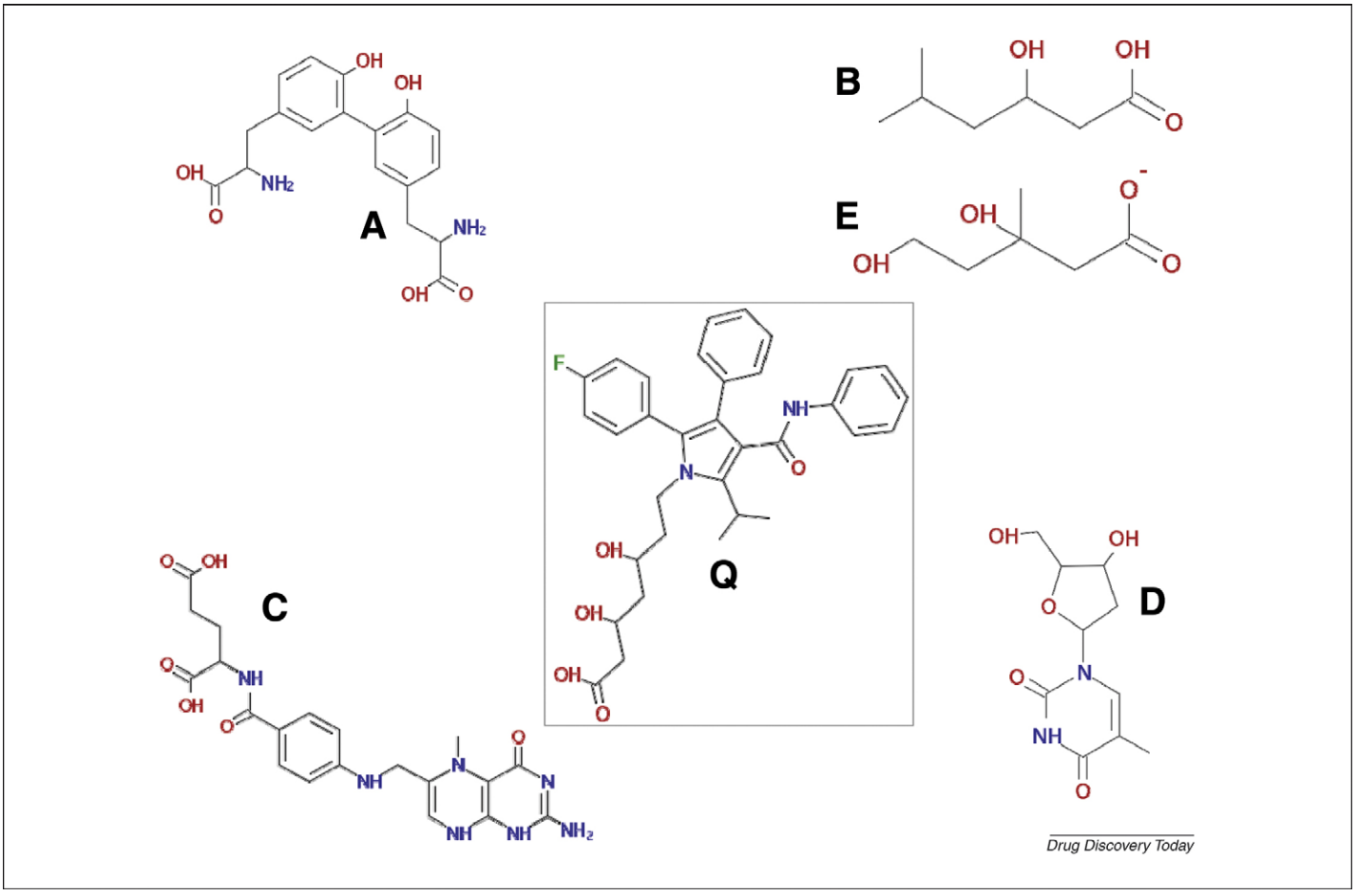

FIGURE 3

An example of metabolite-likeness. The query atorvastatin (Q), and its closest metabolites from E-state key (**a**; dityrosine), connectivity (**b**; 3-hydroxyisoheptanoic acid), path (**c**; 5-methyldihydrofolic acid) and MDL Public Key spaces (**d**; thymidine). Note the similarity of the connectivity-retrieved metabolite to the HMG-CoA reductase product mevalonate (E).

known [92]. In practice, certain descriptors are more useful than others in common tasks, such as learning and retrieval, and the ability of the chosen descriptors to capture relevant chemical information underpins their utility in such tasks. In approaching the question of metabolite-likeness through different spaces, differing views of molecular similarity are afforded, and in all there is clear evidence that drugs exhibit a high level of endogenite-likeness. Future approaches may consider utilising data fusion or consensus approaches [93,94] to combine multiple descriptors into a new descriptor that can draw information from all input spaces to define a still more useful space optimised particularly for assessing the metabolite-likeness of drugs.

A drug that displays a high level of similarity with its closest metabolite is likely to interact with the same target(s) as the metabolite in a native-like manner. This underpins much of the rationalisation of metabolite-likeness, but interactions with non-target molecules are also clearly important. Owing to the increasingly acknowledged role of transporters in drug uptake, the relevance of similarity to native transporter substrates (metabolites and digestion products) will be of crucial importance in drug delivery [39–41]. The promiscuous nature of certain drug transporters, however, appears to impose far fewer constraints than would be expected for drug–target reactions, which are typically more specific. This said, promiscuity is a strong function of lipophilicity, especially for certain chemical classes such as bases [5], and the greater hydrophilicity of metabolites may help users of 'endogenite-likeness' as a filter to avoid unwanted promiscuity. Among the solute carrier family of transporters are certain families that exhibit extremely broad substrate specificity, particularly the peptide (SLC15), organic anion (SLCO) and organic anion/cation/zwitterion (SLC22) transporters, which are known to transport many drugs and other xenobiotics [41].

Conclusion

The space of potential biologically relevant pharmacophores is enormous, even large libraries populate it only sparsely, and attrition remains severe. Consequently it is desirable to develop 'filters' that help to bias the drug discovery search in our favour. Lead-likeness, drug-likeness and the Ro5 have all been used to advantage, but as our knowledge of systems biology grows there is a need to move towards more mechanistic approaches [12,95]. This will also require prediction of where and how drugs interact with metabolism, which can be addressed by cheminformatic methods to assess molecular similarity between putative drugs and metabolites. This is a strategy in which we begin to understand those features of candidate hits and leads that interact not only

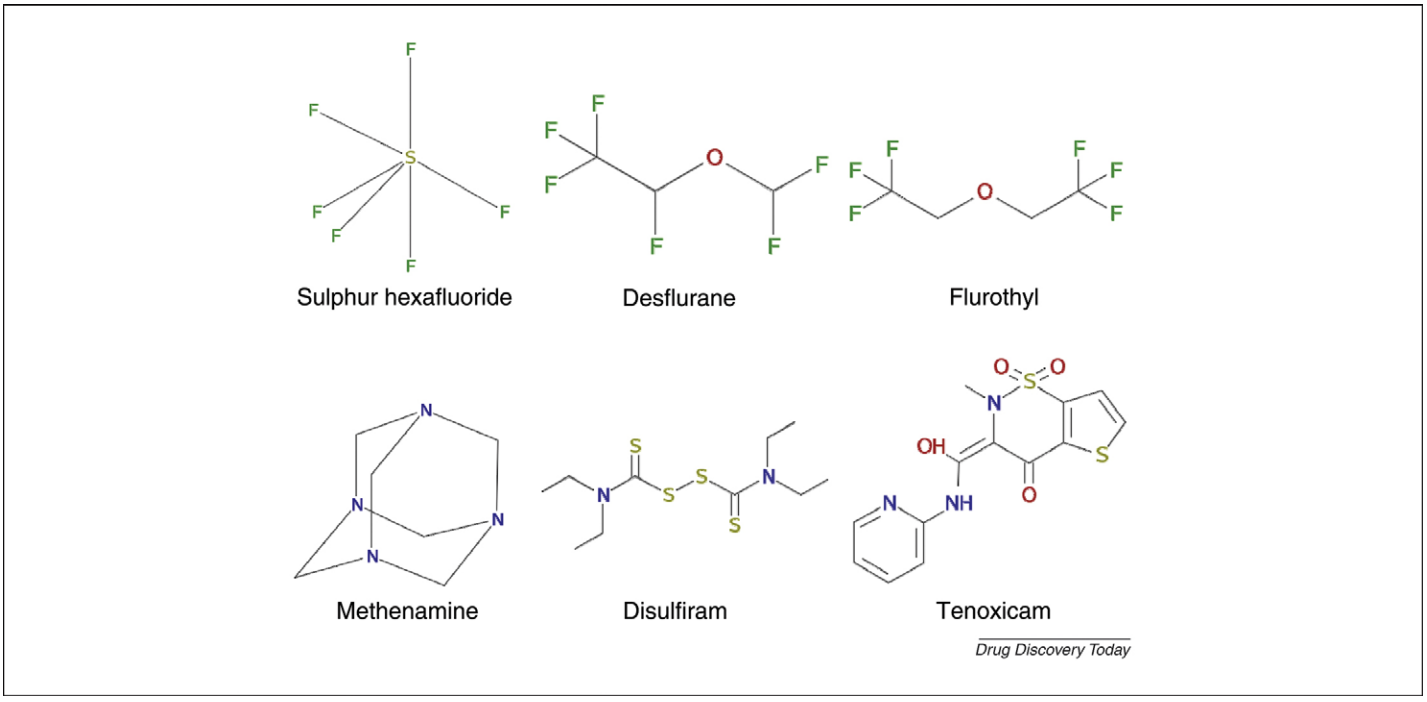

FIGURE 4

A selection of the 'drugs' that are not close to metabolites, including ultrasound contrast agents (sulphur hexafluoride and others, leaving aside a debate on whether these really constitute drugs), general anaesthetics (the structurally similar desflurane, roflurane and methoxyflurane), the convulsant flurothyl, an antibacteriurial (methenamine), the acetaldehyde dehydrogenase inhibitor disulfiram and the non-steroidal anti-inflammatory tenoxicam.

with specific biomolecules (e.g. kinases), but also with other parts of biochemical pathways, such as transporters.

Compound set sources

Three classes of compounds are defined: endogenous human metabolites ('metabolites' shown in Table 1), drugs and pre-drug compounds. Definitions are inherited from source databases; for a compound to be labelled as a human metabolite it need only be present in one of the source human metabolite databases or models. Sources are not considered if it is not possible to derive structures and human origin from the source. Source databases are listed in Table 1.

The pre-drug set is drawn from Zinc (http://zinc.docking.org). It was established that a random subset of 2.5% of Zinc that was clustered in extended connectivity fingerprint (diameter 4) space at a Tanimoto threshold of 0.6 produces sufficient clusters to assign >75% of the whole Zinc database at the same threshold.

The source databases contain misannotated drugs and metabolites, molecules that fall into both categories, and molecules that do not belong in either. Similar compounds from different classes potentially represent misannotations. A semi-automatic strategy to correct errors identified tight clusters in extended fingerprint space containing both drugs and metabolites, which were then examined manually for errors. Molecules were assigned to the classes: 'drug', 'metabolite', 'both' (such as thyroxine), or 'neither' (illicit drugs, food additives and pharmaceutical aids). Molecules that are both metabolites and drugs are considered solely as metabolites when in comparisons. The final set consists of 5333 metabolites, 7330 drugs and 62,390 library compounds.

Overview of protocol

A summary of the procedure to characterise the metabolite-likeness of query compounds is illustrated in Fig. 5.

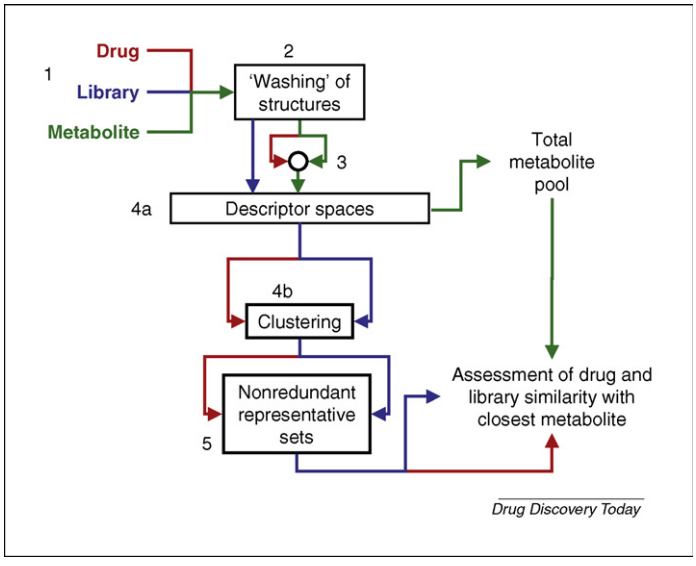

FIGURE 5

An overview of the data processing workflow. Drugs, libraries and metabolites are read (1) and standardised by the 'washing' algorithm (2). The circle indicates a manual check of drug and metabolite definitions (3). Drug and library compounds are clustered in different descriptor spaces (4) to remove redundancy. Cluster centres form representative sets (5). The similarity of representative set members to the *total* metabolite pool is then calculated.

Processing structures

Before the analysis all compounds were 'washed' in Pipeline Pilot [96]. Washing involved isolation of the largest fragment in the structure, the removal of salts and hydrogens and the standardisation of stereochemical and charge information using the Pipeline Pilot 'Standardize Molecule' component. Only compounds with more than three atoms were considered. The washing procedure is available as a Pipeline Pilot workflow at http://www.myexperiment.org/ and http://dbkgroup.org/.

Molecular descriptors

Four molecular descriptors were calculated using Pipeline Pilot.

Extended connectivity fingerprints [72,73] operate by identifying the substructural environment of each atom up to a diameter of 4. The descriptor is then the set of substructures in a molecule calculated thus, with similar molecules sharing more substructures than dissimilar molecules.

Path fingerprints [75] describe a compound by all paths through the molecular graph, here up to length 4, using the Pipeline Pilot implementation of a Daylight-like path fingerprint [74]. In contrast to connectivity fingerprints, paths are not branched and therefore represent the molecule differently.

The MDL keys [76] are substructural features observed to be of utility in retrieval tasks such as database searching, and have also been used in learning problems. Of the full set of 960 useful substructures the definitions of 166 were released as the MDL Public Keys.

E-state indices [77,78] capture the electronic and topological properties of atoms. The indices capture both electronegativity and topological information for each atom in the molecule via electronic interactions with neighbouring atoms, and by distance on the molecular graph, in an index. The set of indices over all molecules forms a descriptor space.

Clustering and representative sets

Compound libraries are typically distributed such that certain regions of chemical space are more highly populated than others, reflecting the types of chemistries that are accessible and considered interesting, and it is still relatively rare in cheminformatics to consider the effects of this redundancy [68]. In consequence, global analyses that ignore this may be biased towards over-represented types and not reflective of the whole set, one result of which can be overstated performance on learning tasks, as has been suggested to have occurred in the previous analyses of drug-likeness predictors [97].

To avoid this problem the total library is sub-sampled by clustering to remove redundancy, with cluster centres used as representative compounds. Using the Pipeline Pilot component 'Cluster Molecules', based upon maximal dissimilarity partitioning, clusters are derived by the imposition of a distance threshold, where the threshold specifies the maximum distance from a molecule to its cluster representative. Representative sets of drug and library compounds were produced by combining both sets and clustering, with closest compounds to the cluster centre from each cluster representing that cluster in the final set. Note that for clusters containing both drug and library compounds one of each class is selected so as to represent each class even when overlap occurs within the distance threshold.

Determining appropriate Tanimoto values for the threshold is non-trivial. Jónsdóttir *et al.* and Frimurer *et al.* [68,97] suggested a threshold similarity value of 0.85, which captures high levels of similarity between compounds, although this depends upon the molecular descriptors chosen. In all descriptors it generates very tight clusters, many of which are singletons, and only very high levels of redundancy are addressed. Here the centres of diverse clusters covering 70% of the total set are selected as representatives. For clusters containing more than one class the closest member to the cluster centre is selected as the class representative. Representative sets in the different descriptor spaces are summarised in Table 2. The representative sets of Table 2 are available in the supplementary material as SD files.

TABLE 2

Sizes of representative sets in each of the descriptor spaces following clustering to remove similarity.

Descriptor space	Drug	Library
Connectivity fingerprint	5723	44,275
Path fingerprint	5835	44,318
MDL Public Keys	5813	44,325
E-state	6028	44,419

Property distributions

Counts and physicochemical properties were calculated using Pipeline Pilot, which implements the *clogD* method of Csizmadia *et al.* [69], here calculated at the default pH of 7.4.

Acknowledgements

Our interest in pursuing these issues has been helped considerably by grant BB/D007747/1 from the BBSRC, together with attendant funding from GSK. We thank Scott Summerfield and Phil Jeffrey of GSK for their support and interest, and Karin Lanthaler and Steve Oliver for useful discussions. PD and YP also thank BBSRC for current funding under the ONDEX SABR and SCIBS schemes. DBK also thanks the EPSRC and RSC for financial support, and the Royal Society/Wolfson Foundation for a Research Merit Award. This is a contribution from the BBSRC- and EPSRC-funded Manchester Centre for Integrative Systems Biology (www.mcisb.org/).

References

1 Lipinski, C. and Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature* 432, 855–861

2 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815

3 Bleicher, K.H. *et al.* (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* 2, 369–378

4 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315

5 Leeson, P.D. and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* 6, 881–890
6 Hodgson, J. (2001) ADMET – turning chemicals into drugs. *Nat. Biotechnol.* 19, 722–726
7 Gleeson, M.P. (2008) Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* 51, 817–834
8 Lipinski, C.A. (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249
9 Noble, D. *et al.* (1999) Biological simulations in drug discovery. *Drug Discov. Today* 4, 10–16
10 Butcher, E.C. (2005) Can cell systems biology rescue drug discovery? *Nat. Rev. Drug Discov.* 4, 461–467
11 Cho, C.R. *et al.* (2006) The application of systems biology to drug discovery. *Curr. Opin. Chem. Biol.* 10, 294–302
12 Kell, D.B. (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov. Today* 11, 1085–1092
13 Kell, D.B. (2007) The virtual human: towards a global systems biology of multiscale, distributed biochemical network models. *IUBMB Life* 59, 689–695
14 Rishton, G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* 8, 86–96
15 Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* 8, 255–263
16 Wunberg, T. *et al.* (2006) Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* 11, 175–180
17 Gillet, V.J. *et al.* (1998) Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 38, 165–179
18 Wagener, M. and van Geerestein, V.J. (2000) Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inform. Comput. Sci.* 40, 280–292
19 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730
20 Walters, W.P. and Murcko, M.A. (2002) Prediction of 'drug-likeness'. *Adv. Drug Deliv. Rev.* 54, 255–271
21 Muegge, I. (2003) Selection criteria for drug-like compounds. *Med. Res. Rev.* 23, 302–321
22 Ajay, (2002) Predicting drug-likeness: why and how? *Curr. Top. Med. Chem.* 2, 1273–1286
23 Biswas, D. *et al.* (2006) A simple approach for indexing the oral druglikeness of a compound: discriminating druglike compounds from nondruglike ones. *J. Chem. Inf. Model* 46, 1394–1401
24 Reichel, A. (2006) The role of blood–brain barrier studies in the pharmaceutical industry. *Curr. Drug Metab.* 7, 183–203
25 Hopkins, A.L. *et al.* (2006) Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* 16, 127–136
26 Hopkins, A. *et al.* (2005) Chemical tools for indications discovery. *Ann. Rep. Med. Chem.* 40, 339–348
27 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25
28 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623
29 Baurin, N. *et al.* (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* 44, 643–651
30 Hopkins, A.L. *et al.* (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431
31 Rees, D.C. *et al.* (2004) Fragment-based lead discovery. *Nat. Rev. Drug Discov.* 3, 660–672
32 Abad-Zapatero, C. and Metz, J.T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* 10, 464–469
33 Reynolds, C.H. *et al.* (2007) The role of molecular size in ligand efficiency. *Bioorg. Med. Chem. Lett.* 17, 4258–4261
34 Congreve, M. *et al.* (2003) A rule of three for fragment-based lead discovery? *Drug Discov. Today* 8, 876–877
35 Ciulli, A. and Abell, C. (2007) Fragment-based approaches to enzyme inhibition. *Curr. Opin. Biotechnol.* 18, 489–496
36 Hubbard, R.E. *et al.* (2007) Informatics and modeling challenges in fragment-based drug discovery. *Curr. Opin. Drug Discov. Dev.* 10, 289–297
37 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
38 Ganesan, A. (2008) The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol.* 12, 306–317
39 Sai, Y. and Tsuji, A. (2004) Transporter-mediated drug delivery: recent progress and experimental approaches. *Drug Discov. Today* 9, 712–720
40 Sai, Y. (2005) Biochemical and molecular pharmacological aspects of transporters as determinants of drug disposition. *Drug Metab. Pharmacokinet.* 20, 91–99
41 Dobson, P.D. and Kell, D.B. (2008) Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat. Rev. Drug Discov.* 7, 205–220
42 Cheng, A.C. *et al.* (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75
43 Hajduk, P.J. *et al.* (2005) Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* 48, 2518–2525
44 Greene, L.H. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* 35, D291–297 (database issue)
45 Marsden, R.L. *et al.* (2006) Exploiting protein structure data to explore the evolution of protein function and biological complexity. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 361, 425–440
46 Pang, Y.P. (2007) In silico drug discovery: solving the "target-rich and lead-poor" imbalance using the genome-to-drug-lead paradigm. *Clin. Pharmacol. Ther.* 81, 30–34
47 Duarte, N.C. *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1777–1782
48 Ma, H. *et al.* (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* 3, 135
49 Feist, A.M. and Palsson, B.O. (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol* 26 (6), 659–667
50 Herrgard, M.J. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol* 26 (10), 1155–1160
51 Ma, H. and Goryanin, I. (2008) Human metabolic network reconstruction and its impact on drug discovery and development. *Drug Discov. Today* 13, 402–408
52 Jenkins, H. *et al.* (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* 22, 1601–1606
53 Brooksbank, C. *et al.* (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* 33, D46–53 (database issue)
54 Smith, C.A. *et al.* (2005) METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* 27, 747–751
55 Kanehisa, M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–357 (database issue)
56 Spasic, I. *et al.* (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinformatics* 7, 281
57 Wheeler, D.L. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35, D5–12 (database issue)
58 Wishart, D.S. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35, D521–526 (database issue)
59 Spasić, I. *et al.* (2008) Facilitating the development of controlled vocabularies for metabolomics with text mining. *BMC Bioinformatics* 9, S5
60 Nobeli, I. *et al.* (2003) A structure-based anatomy of the *E. coli* metabolome. *J. Mol. Biol.* 334, 697–719
61 Nobeli, I. and Thornton, J.M. (2006) A bioinformatician's view of the metabolome. *Bioessays* 28, 534–545
62 Ley, R.E. *et al.* (2008) Evolution of mammals and their gut microbes. *Science* 320, 1647–1651
63 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
64 Irwin, J.J. and Shoichet, B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model* 45, 177–182
65 Karakoc, E. *et al.* (2006) Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model* 46, 2167–2182
66 Fahy, E. *et al.* (2005) A comprehensive classification system for lipids. *J. Lipid Res.* 46, 839–861
67 Sud, M. *et al.* (2007) LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* 35, D527–532 (database issue)
68 Jónsdóttir, S.Ó. *et al.* (2005) Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics* 21, 2145–2160
69 Csizmadia, F. *et al.* (1997) Prediction of distribution coefficient from structure. 1. Estimation method. *J. Pharm. Sci.* 86, 865–871

70 Zhang, M.Q. and Wilkinson, B. (2007) Drug discovery beyond the 'rule-of-five'. *Curr. Opin. Biotechnol.* 18, 478–488
71 Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* 11, 1046–1053
72 Hert, J. *et al.* (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* 2, 3256–3266
73 Rogers, D. *et al.* (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* 10, 682–686
74 James, C.A. and Weininger, D. (1995) *Daylight Theory Manual*. Daylight Chemical Information Systems, Inc.
75 Flower, D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* 38, 379–386
76 Durant, J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42, 1273–1280
77 Hall, L.H. and Kier, L.B. (2000) The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.* 40, 784–791
78 Kier, L.B. and Hall, L.H. (2001) Database organization and searching with E-state indices. *SAR QSAR Environ. Res.* 12, 55–74
79 Wierzbicki, A.S. *et al.* (2003) The lipid and non-lipid effects of statins. *Pharmacol. Ther.* 99, 95–112
80 Liao, J.K. and Laufs, U. (2005) Pleiotropic effects of statins. *Annu. Rev. Pharmacol. Toxicol.* 45, 89–118
81 Wang, C.Y. *et al.* (2008) Pleiotropic effects of statin therapy: molecular mechanisms and clinical results. *Trends Mol. Med.* 14, 37–44
82 Sugawara, M. *et al.* (2000) Transport of valganciclovir, a ganciclovir prodrug, via peptide transporters PEPT1 and PEPT2. *J. Pharm. Sci.* 89, 781–789
83 Daniel, H. and Kottra, G. (2004) The proton oligopeptide cotransporter family SLC15 in physiology and pharmacology. *Pflugers Arch.* 447, 610–618
84 Sievaen, E. (2007) Exploitation of bile acid transport systems in prodrug design. *Molecules* 12, 1859–1889
85 Tolle-Sander, S. *et al.* (2004) Increased acyclovir oral bioavailability via a bile acid conjugate. *Mol. Pharm.* 1, 40–48
86 O'Hagan, S. *et al.* (2005) Closed-loop, multiobjective optimisation of analytical instrumentation: gas-chromatography–time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Anal. Chem.* 77, 290–303
87 Smith, C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787
88 O'Hagan, S. *et al.* (2007) Closed-loop, multi-objective optimisation of two-dimensional gas chromatography (GC × GC–tof-MS) for serum metabolomics. *Anal. Chem.* 79, 464–476
89 Cotter, D. *et al.* (2006) LMPD: LIPID MAPS proteome database. *Nucleic Acids Res.* 34, D507–510 (database issue)
90 Yetukuri, L. *et al.* (2008) Informatics and computational strategies for the study of lipids. *Mol. Biosyst.* 4, 121–127
91 Everitt, B.S. (1993) *Cluster Analysis*. Edward Arnold
92 Maggiora, G.M. and Shanmugasundaram, V. (2004) Molecular similarity measures. *Methods Mol. Biol.* 275, 1–50
93 Ginn, C.M.R. *et al.* (2000) Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Des.* 20, 1–16
94 Holliday, J.D. *et al.* (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
95 Kell, D.B. (2006) Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher lecture. *FEBS J.* 273, 873–894
96 Hassan, M. *et al.* (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol. Divers.* 10, 283–299
97 Frimurer, T.M. *et al.* (2000) Improving the odds in discriminating ''drug-like'' from ''non drug-like'' compounds. *J. Chem. Inf. Comput. Sci.* 40, 1315–1324
98 Kanehisa, M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36 (Database issue), D480–D484