JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

How Do Metabolites Differ from Their Parent Molecules and How Are They Excreted?

Johannes Kirchmair,[†] Andrew Howlett,[†] Julio E. Peironcely,[‡,§,⊥] Daniel S. Murrell,[†] Mark J. Williamson,[†] Samuel E. Adams,[†] Thomas Hankemeier,[§,⊥] Leo van Buren,[○] Guus Duchateau,[○] Werner Klaffke,[○] and Robert C. Glen*,[†]

[†]Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom

[‡]TNO Research Group Quality & Safety, P.O. Box 360, 3700 AJ Zeist, The Netherlands

[§]Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg, 2333 CC Leiden, The Netherlands

[⊥]Netherlands Metabolomics Centre, Einsteinweg, 2333 CL Leiden, The Netherlands

[○]Unilever Research & Development, Olivier van Noortlaan, 3133 AT Vlaardingen, The Netherlands

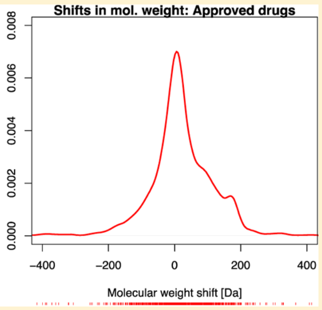
Supporting Information

What physicochemical property space do approved drugs, human metabolites and molecules related to TCM occupy?

How does metabolism alter the physicochemical property space from substrate to metabolic product?

What shifts in physicochemical property space do individual metabolic reactions introduce into molecules?

What are the physicochemical properties of metabolites found in the bile, faeces and urine?



ABSTRACT: Understanding which physicochemical properties, or property distributions, are favorable for successful design and development of drugs, nutritional supplements, cosmetics, and agrochemicals is of great importance. In this study we have analyzed molecules from three distinct chemical spaces (i) approved drugs, (ii) human metabolites, and (iii) traditional Chinese medicine (TCM) to investigate four aspects determining the disposition of small organic molecules. First, we examined the physicochemical properties of these three classes of molecules and identified characteristic features resulting from their distinctive biological functions. For example, human metabolites and TCM molecules can be larger and more hydrophobic than drugs, which makes them less likely to cross membranes. We then quantified the shifts in physicochemical property space induced by metabolism from a holistic perspective by analyzing a data set of several thousand experimentally observed metabolic trees. Results show how the metabolic system aims to retain nutrients/micronutrients while facilitating a rapid elimination of xenobiotics. In the third part we compared these global shifts with the contributions made by individual metabolic reactions. For better resolution, all reactions were classified into phase I and phase II biotransformations. Interestingly, not all metabolic reactions lead to more hydrophilic molecules. We were able to identify biotransformations leading to an increase of logP by more than one log unit, which could be used for the design of drugs with enhanced efficacy. The study closes with the analysis of the physicochemical properties of metabolites found in the bile, faeces, and urine. Metabolites in the bile can be large and are often negatively charged. Molecules with molecular weight >500 Da are rarely found in the urine, and most of these large molecules are charged phase II conjugates.

INTRODUCTION

Absorption, distribution, metabolism, and excretion (ADME) are fundamental properties which determine the disposition of a molecule in an organism. Absorption and distribution are dominated by the solubility, partitioning, size, and chemical stability of a molecule. These criteria are often decisive for the molecule to reach its pharmacological target. Metabolic processes which transform small organic molecules can lead to significantly altered bioactivity profiles, cause enhanced efficacy or a loss of activity, induce drug−drug interactions, or adverse and toxic effects. Clearance is a decisive criterion for the half-life of a molecule in the organism. Rapid clearance can cause therapeutic failure while incomplete or too slow clearance may result in the accumulation of a molecule and, consequently, off-target/toxic effects.

Understanding the ADME properties of molecules is therefore an imperative for the development of drugs,

Received: October 12, 2012
Published: January 25, 2013

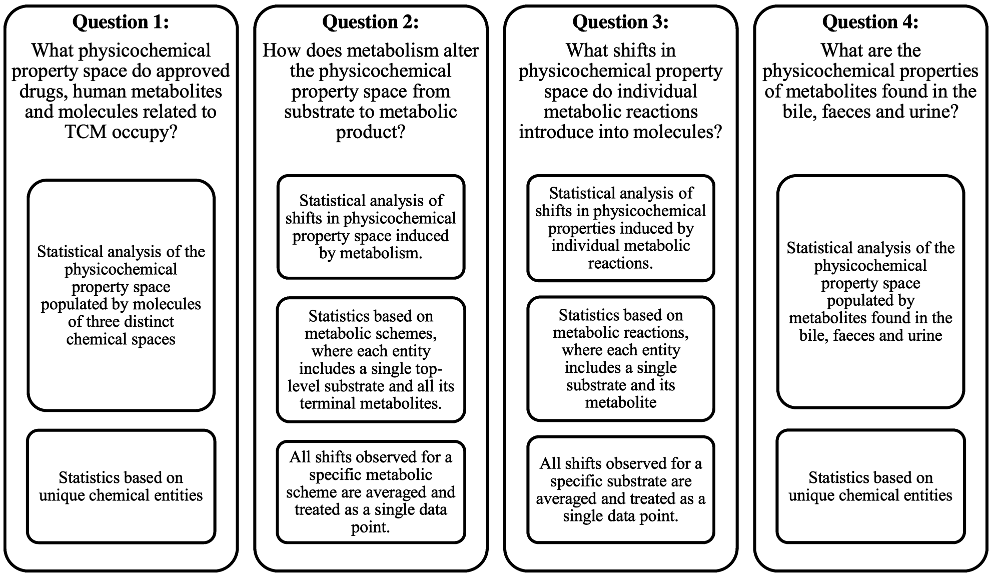

Question 1:
What physicochemical property space do approved drugs, human metabolites and molecules related to TCM occupy?
Statistical analysis of the physicochemical property space populated by molecules of three distinct chemical spaces
Statistics based on unique chemical entities
Question 2:
How does metabolism alter the physicochemical property space from substrate to metabolic product?
Statistical analysis of shifts in physicochemical property space induced by metabolism.
Statistics based on metabolic schemes, where each entity includes a single top-level substrate and all its terminal metabolites.
All shifts observed for a specific metabolic scheme are averaged and treated as a single data point.
Question 3:
What shifts in physicochemical property space do individual metabolic reactions introduce into molecules?
Statistical analysis of shifts in physicochemical properties induced by individual metabolic reactions.
Statistics based on metabolic reactions, where each entity includes a single substrate and its metabolite
All shifts observed for a specific substrate are averaged and treated as a single data point.
Question 4:
What are the physicochemical properties of metabolites found in the bile, faeces and urine?
Statistical analysis of the physicochemical property space populated by metabolites found in the bile, faeces and urine
Statistics based on unique chemical entities

Figure 1. Overview of the different questions addressed in this study and the analysis methods used.

nutritional supplements, cosmetics and agrochemicals. Experimental and theoretical techniques to probe ADME properties are advancing rapidly,[1,2] and industry has responded to the high attrition rates related to these properties by introducing screening methods to address poor bioavailability and unfavorable metabolism. A key strategy is to identify problems at an early stage by including ADME profiling in initial compound design. These efforts are supported by a plethora of computational approaches for ADME prediction.[3] However, despite this, failure rates are still very high. It is therefore of interest to expand our knowledge of ADME processes in order to understand and improve compound design.

In this work, we have addressed four important questions pertinent to the design and development of new molecules with favorable ADME properties (Figure 1):

Question 1. What physicochemical property space do approved drugs, human metabolites, and molecules related to traditional Chinese medicine (TCM) occupy? This question has been addressed earlier for chemicals,[4−9] drugs,[2,4,5,7,9−21] leads,[9,13−15,22] metabolites,[9,15,16,18,22−25] natural products,[6−8,15,17,22,26] peptides,[27] pesticides,[28] toxins,[16] and, very recently, TCM molecules.[29−31] In the current work, we aim to directly compare the distinct physicochemical property spaces occupied by approved drugs (represented by the Approved Drugs subset of DrugBank[32]), human metabolites (represented by the Human Metabolome Database (HMDB)[33]), and TCM molecules (represented by the TCM Database@Taiwan[34]) using curated and diversified data sets, which serve as a basis for investigating questions 2−4 (see below).

Question 2. How does metabolism alter the physicochemical property space from substrate to metabolic product? The aim of this analysis is to determine the changes in physicochemical properties introduced by the metabolic system.

Question 3. What shifts in physicochemical property space do individual metabolic reactions introduce into molecules? In contrast to question 2, here we analyze individual phase I and phase II reactions and quantify their effects on physicochemical properties.

Question 4. What are the physicochemical properties of metabolites found in the bile, faeces, and urine? We analyze this question for the three investigated chemical spaces based on experimental data taken from the Accelrys Metabolite Database (AMDB).[35] Knowing the excretion route of a compound's metabolites can be particularly important, e.g., when addressing specific clinical situations such as bacterial infections of the urinary bladder and for the treatment of patients with, e.g., renal impairment.

■ MATERIALS AND METHODS

An overview of the data processing and analysis workflow is provided in Supporting Information Figure 1, and a summary of the data sets and their sizes as used in the following statistical analyses is provided in Supporting Information Table 1. For details on computer hardware and software, see the Supporting Information.

Question 1. DrugBank, Human Metabolome Database (HMDB), and TCM (traditional Chinese medicine) Database@Taiwan were used as representative data sets to define the physicochemical property space populated by approved drugs, human metabolites, and TCM molecules. DrugBank[32] version 3.0 contains more than 6500 drug molecules. A subset of more than 1400 drugs approved by the FDA is available from DrugBank; this was used in the current investigation. HMDB[33] version 2.5 contains more than 8500 human metabolites. TCM Database@Taiwan[34] version 1.0 contains about 40 000 molecules collected from the literature on TCM. The data sets were prepared following to the protocol reported in the Supporting Information. Physicochemical property descriptors were calculated using Molecular Operating Environment (MOE) version 2010.10.[36] This procedure included various structure preparation steps such as adding hydrogens, assigning protonation states, and calculating a low energy conformation (see the Supporting Information). Any strong bias of the data sets toward certain molecular scaffolds was balanced by clustering for structural diversity using a Pipeline Pilot workflow and selecting the cluster center molecules for analysis only (see the Supporting Information).

Question 2. AMDB version 2011.2[35] comprises 103 908 experimentally confirmed biotransformations organized in 14 272 metabolic schemes/trees (Figure 2). The individual

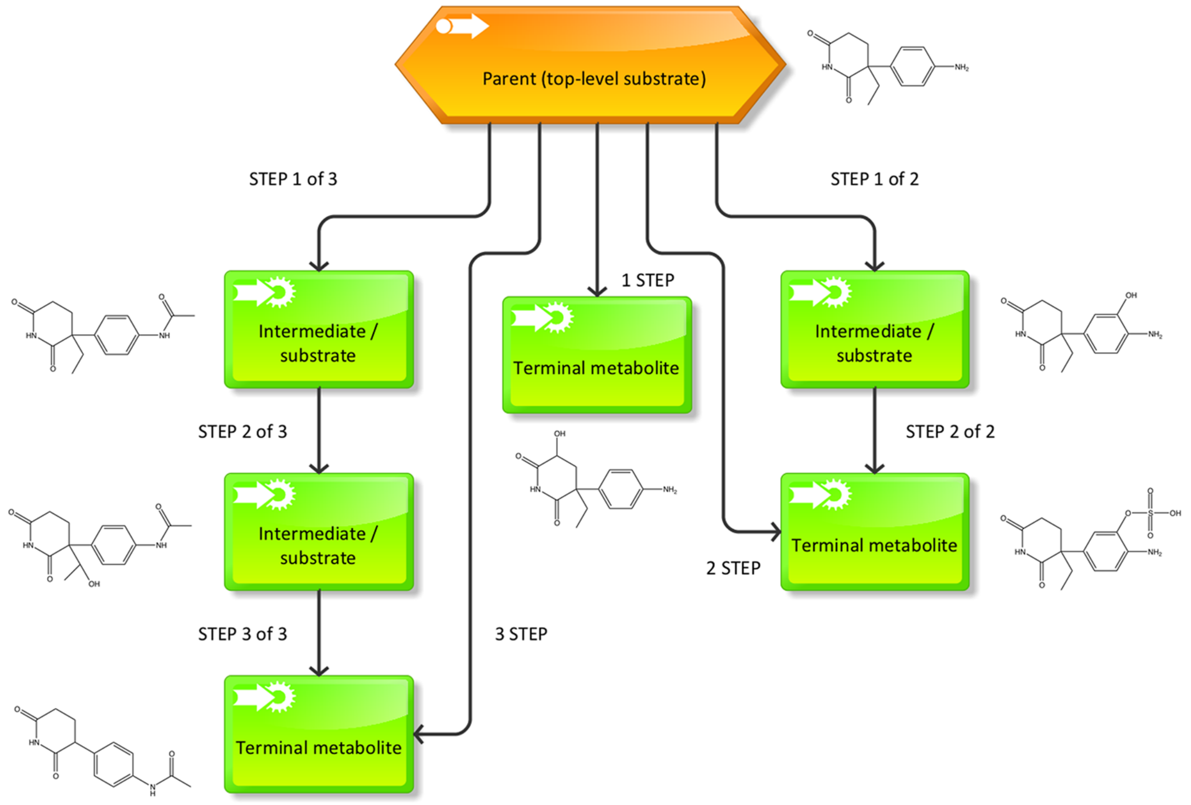


Figure 2. Definition of a metabolic scheme. A metabolic scheme consists of a single top-level substrate (parent molecule; root of the metabolic scheme) and all of its terminal metabolites (molecules for which no further metabolization has been recorded). It may also contain one or several intermediate metabolites (molecules, which are themselves substrates that are further metabolized). All metabolic schemes used in this study are taken from the AMDB and are experimental data extracted from primary literature. STEP denotes the sequence of a metabolic reaction and its step size, such as "1 STEP" (metabolite resulting from a single reaction step), "3 STEP" (metabolite which is a result of three reaction steps), or "2 of 3" (intermediate metabolite which is further metabolized). Intermediate metabolites are any molecules where $x < y$ applies to "STEP x of y" (such as "STEP 1 of 2"). Terminal metabolites are molecules where $x = y$ applies to "STEP x of y" (such as "STEP 2 of 2") or this expression is equal to "1 STEP".

database records hold a single chemical reaction together with annotations such as enzymes catalyzing this reaction, excretion routes of the metabolites, literature references, etc. All records were exported as an MDL RD file, which was converted into a MDL SD file using an in-house Java application. Some filters were applied to remove invalid metabolic schemes from the data set (see the Supporting Information). Physicochemical property descriptors were calculated as described for question 1. To assign chemical spaces, a lookup of all top-level substrates of the AMDB was performed in DrugBank, HMDB, and TCM Database@Taiwan using InChI notations (see the Supporting Information). If present, the metabolic schemes were assigned to the respective chemical space(s). A diverse set of metabolic schemes was generated using the clustering approach described for question 1. The top-level substrates served as references for the individual metabolic schemes.

Question 3. MetaPrint2D version 1.0[37,38] is open software that allows the identification and prediction of sites of metabolism and metabolite structures using a statistical model trained on biotransformations. The software was used to classify biotransformations of the AMDB into phase I and phase II reactions as well as into specific reaction types. MetaPrint2D discriminates between 96 reaction (sub)types. For the current analysis these reaction types were harmonized and merged for the sake of clarity (see Supporting Information Table 2). Some filters were applied to remove invalid metabolic schemes from the data set (see the Supporting Information). Descriptor values were taken from question 2. All reactions of the individual metabolic schemes were assigned the chemical space classification of their respective top-level substrate (see Question 2). The biotransformations were diversified with respect to their substrate structures following the protocol described for question 1.

Question 4. All metabolites present in the AMDB, for which data on their excretion route are available from this database, were considered for analysis. Descriptor values were taken from question 2. All terminal metabolites were assigned to the chemical space of their respective top-level substrate (see Question 2). The three sets of metabolites were diversified individually following the protocol described for question 1.

■ RESULTS

What Physicochemical Property Space Do Approved Drugs, Human Metabolites, and Molecules Related to Traditional Chinese Medicine Occupy (Question 1)? Here we attempt to characterize the distribution of key physicochemical parameters for the ADME properties of approved drugs, human metabolites, and TCM molecules. Diverse subsets were generated from cluster center molecules identified by clustering for maximum diversity (see Materials and Methods). All values reported herein refer to these diversified data sets, except where indicated. The most interesting findings

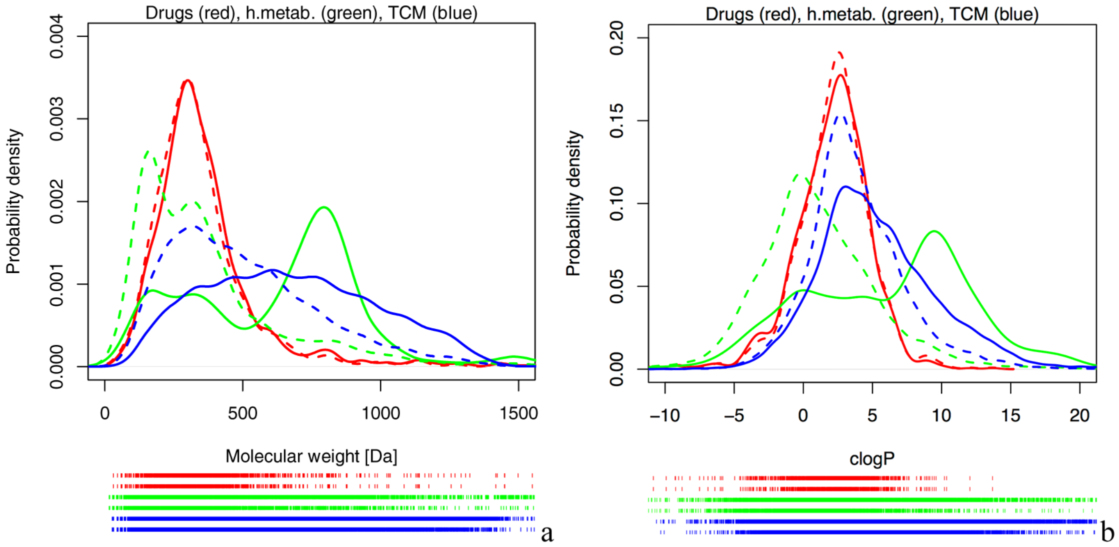


Figure 3. MW and clogP probability density for Approved Drugs (red), HMDB (green), and TCM Database@Taiwan (blue), for the nonclustered (continuous lines) and clustered data sets (dashed lines). For each color, the upper rung (value indicator bar located below the graph) represents the nonclustered and the lower rung represents the clustered data sets. (a) TCM molecules are on average substantially larger than approved drugs and also than human metabolites. Areas of high MW are much more densely populated by TCM molecules than by approved drugs. (b) In contrast to approved drugs, human metabolites show a highly populated range of negative clogP values, which documents their hydrophilicity. TCM molecules cover a wide range of clogP values, from hydrophilic to highly hydrophobic.

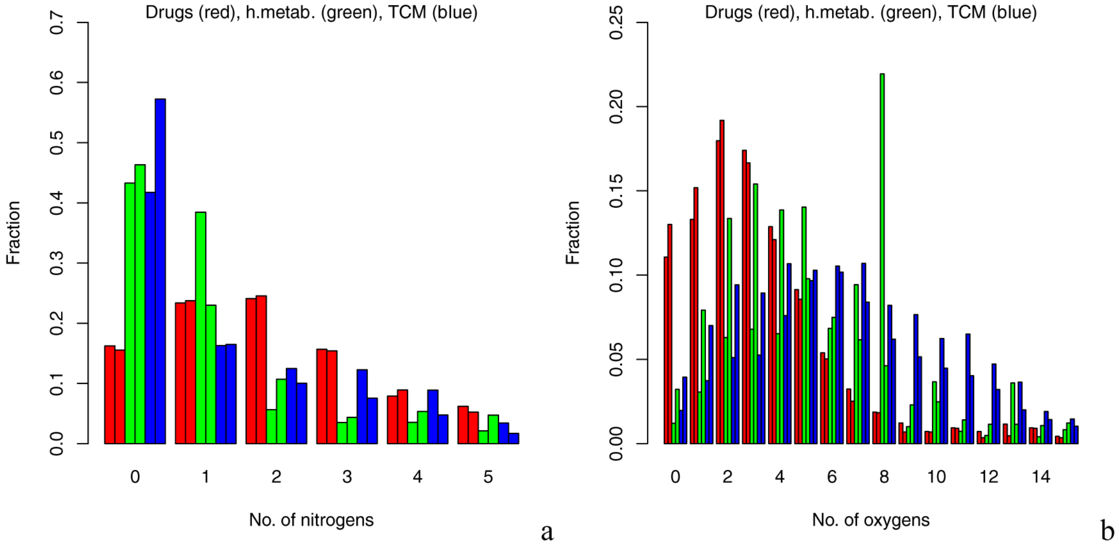


Figure 4. Number of (a) nitrogen and (b) oxygen atoms per molecule for approved drugs (red), human metabolites (green), and TCM molecules (blue). For each color, the left bar represents the nonclustered, and the right bar, the clustered data set. The proportion of molecules without any nitrogen atom is much higher for human metabolites (46%) and TCM molecules (57%), compared to approved drugs (16%, all values reported for the clustered data sets), which is in agreement with earlier reports.[16] Maxima are observed at $N = 2$ per molecule for approved drugs and at $N = 0$ for human metabolites and TCM molecules. (b) For approved drugs, a maximum is found for two oxygen atoms per molecule. Approved drugs with more than five oxygen atoms are rare. The nonclustered HMDB data set shows a maximum at eight oxygen atoms per molecule. This accumulation caused by phospholipids is neutralized by clustering for maximum diversity.

are summarized in the following section. Additional data on a variety of physicochemical descriptors are provided in Supporting Information Table 3.

Basic Molecular Properties and Elementary Composition. MW and lipophilicity are two key properties that often correlate with a molecule's ADME characteristics, including cell permeability and clearance.[20,39−41] The lipophilicity of a molecule typically dominates the volume of distribution, which is the result of a combination of interdependent physicochemical properties such as plasma protein binding, membrane permeation, and (plasma to) tissue distribution.[20,42,43] It is critical for a variety of related pharmacological and pharmacokinetic aspects such as metabolism, half-life, volume of distribution, and clearance. Recently it has been pointed out that in vivo concentrations of metabolites can be decisively affected by physicochemical parameters such as hydrophobicity and charge.[44]

Lipophilicity can be estimated by logP, the partition coefficient between two solvents, usually water and octanol. Hann and Keserü[2] have defined the sweet spot for drug-like molecules to be located at MW ∼ 400 Da with logP ∼ 3 and the favorable chemical space to be situated within a MW range

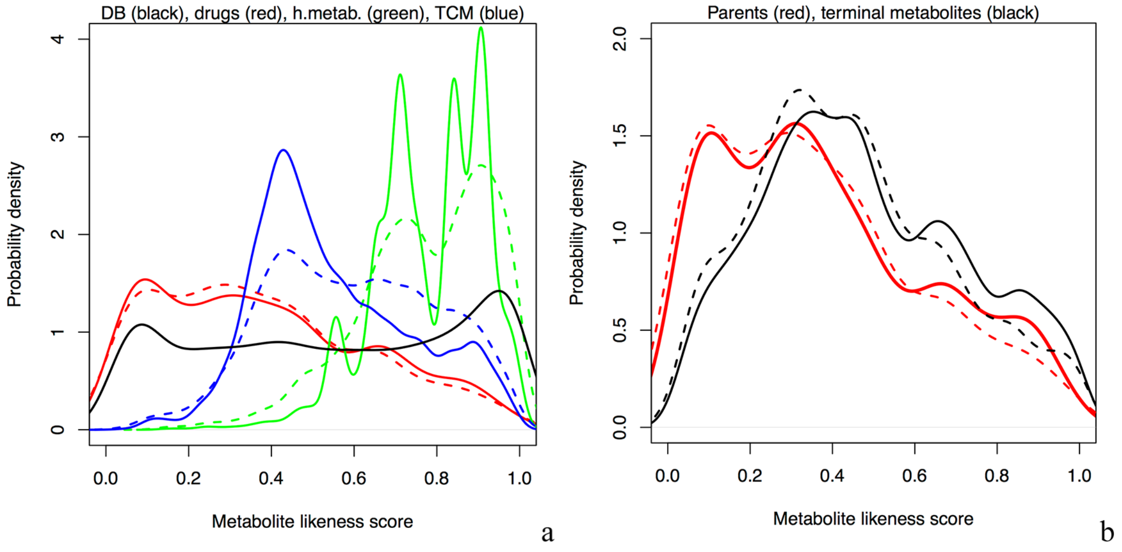


Figure 5. Metabolite likeness calculated with a random forest model employing MDL Public Keys. (a) Probability densities calculated for all DrugBank molecules (black), approved drugs (red), human metabolites (green), and TCM molecules (blue), for nonclustered (continuous lines) and clustered data sets (dashed lines). The model can clearly identify human metabolites. The peak observed for the DrugBank data set at ∼0.95 is a result of the presence of molecules that are themselves human metabolites or close analogues thereof. (b) Shift in metabolite likeness introduced by metabolism (top-level substrates in red, terminal metabolites in black, line style as for part a).

of 250−500 Da and a logP range of 2−4. Hence, these two parameters, which are known to be correlated to a certain extent (see below), significantly constrain the optimum chemical space of interest in drug discovery.

The MW of approved drugs follows a normal distribution with a mean value of 350 Da (Figure 3a). The clustered HMDB data set comes close to that (376 Da); however, the mean MW for the nonclustered HMDB is substantially higher (667 Da). This is caused by an accumulation of acylglycerols, in particular triglycerides, polyphosphates and oligosaccharides (see peak at ∼800 Da). TCM molecules are significantly heavier than molecules of the other data sets. Differences between the nonclustered (mean MW 678 Da) and the clustered data set (mean MW 503 Da) are smaller than for the HMDB data sets and are caused by the accumulation of flavonoids polyphenols, anthraquinones, glycosylated steroids, and a variety of macrocyclic structures.

For TCM molecules, the mean clogP value (3.9) is much higher than for approved drugs (2.2); see Figure 3b. For the clustered HMDB data set the mean clogP value is 1.0, indicating better aqueous solubility than approved drugs. There is a strong bias introduced by acylglycerols and similar molecules in the nonclustered HMDB data set, raising the mean clogP value to 6.0. MW and clogP distributions are also displayed as scatter plots in Supporting Information Figure 2, confirming a high density of approved drugs located in the proximity of the property sweet spot discussed earlier and defined by Hann and Keserü.[2] Also, the trend for larger drug and TCM molecules to be more lipophilic is apparent.

The accessible surface area (ASA) of polar atoms is often related to oral bioavailability and the ability of a molecule to cross cell membranes (see the Introduction and the work of Veber et al.[11]). It is hence straightforward that drugs, which need to cross cell membranes to reach their target, have on average a much smaller polar surface area (171 Å^2) than human metabolites (240 Å^2) and TCM molecules (212 Å^2), which are often produced at their area of function and hence not necessarily required to partition into other compartments.

The average approved drug contains more nitrogen atoms (2.3) than any of the other investigated chemical spaces (human metabolites 1.5, TCM molecules 1.0); see Figure 4a. In contrast to that, human metabolites (6.0) and TCM molecules (6.3) contain almost twice as many oxygen atoms per molecule than approved drugs (3.5); see Figure 4b. About 0.5 halogen atoms are on average present in an approved drug. Approved drugs with >3 halogen atoms are rare and most of them contain a trifluoromethyl group. Three-quarters of all approved drugs do not contain any halogen atoms.

Chemical Features. The mean number of hydrogen bond acceptors per molecule is lower for approved drugs (4.1) than for human metabolites (6.2) and TCM molecules (6.1; see Supporting Information Figure 3 for histogram plots of this and further descriptors discussed in this section). Approved drugs contain on average only about half the number of hydrogen bond donors (2.3) of human metabolites (4.7) and TCM molecules (3.9). The number of hydrogen bond donors and acceptors is related to the number of oxygen atoms per molecule.

The average number of acidic atoms per approved drug molecule is 0.8. This is double the incidence of acidic atoms contained in TCM molecules (0.4). The number is higher for human metabolites (2.1) due to the large number of glucuronidated molecules present. The mean value of basic atoms is higher for approved drugs (0.6) than for human metabolites (0.4), which is a consequence of the increased occurrence of nitrogen atoms in approved drugs. The majority of approved drugs and TCM molecules do not contain strongly acidic or basic atoms. With strong acids and bases deprotonated/protonated, the mean formal charge of approved drugs (+0.2) and TCM molecules (+0.2) is positive while it is negative for human metabolites (−0.8). 29% of all human metabolites have a formal charge of −1, which is a result of glucuronidation.

The number of hydrophobic atoms is related to the total number of heavy atoms and in part also to the MW. About two-thirds of all atoms of approved drugs (66%) and three-quarters

Table 1. Violation of Lipinski's Rule of Five

violation of rule	d[a]	d_c[b]	h[c]	h_c[d]	t[e]	t_c[f]
hydrogen bond donors ≤ 5	9%	8%	24%	28%	45%	28%
hydrogen bond acceptors ≤ 10	6%	5%	20%	14%	22%	14%
logP ≤ 5	11%	12%	58%	15%	47%	31%
MW ≤ 500 Da	14%	13%	66%	21%	68%	43%

[a] Approved drugs. [b] Approved drugs, clustered data set. [c] Human metabolites. [d] Human metabolites, clustered data set. [e] TCM molecules. [f] TCM molecules, clustered data set.

(74%) of all atoms of TCM molecules are hydrophobic. For human metabolites the percentage is 57%. The average number of aromatic bonds is much lower for human metabolites (2.6) than for drug molecules (7.8) and TCM molecules (8.0). For the latter, molecules with more than one aromatic ring are common. However, this is also related to the substantially higher MW of molecules of this chemical space.

Metabolite Likeness. Molecules from the three distinct chemical spaces were evaluated regarding their metabolite likeness using the random forest-based model published earlier.[18] This model calculates a metabolite-likeness score and was derived from the direct comparison of the chemical space occupied by human metabolites (from HMDB) with nonmetabolites (from ZINC[45]). Employing MDL Public Keys as descriptors, this model was able to classify metabolites and nonmetabolites with high accuracy (correct classification of an external validation set of 457 molecules in 96% of all cases).[18] Figure 5a illustrates the distribution of metabolite-likeness scores obtained with this model for the individual data sets. A good separation between approved drugs and human metabolites is observed. The metabolite-likeness values for TCM molecules are situated in between the two other chemical spaces. The plot also includes the probability density calculated for the whole DrugBank data set. Most interestingly, a difference between the distribution of the approved drugs and the whole DrugBank data set can be observed: approved drugs include a lower proportion of molecules that obtain high metabolite likeness scores. This would suggest, surprisingly, that molecules based on structures distinct from human metabolites are more likely to be approved for therapeutic use. Investigation of the structures of DrugBank which obtain a high metabolite-likeness score however reveals that these molecules in fact are human metabolites or close analogues thereof, mainly sugars and peptides. An accumulation of the carboxylic acid functional group is observed, and these molecules are generally of low MW.

Compliance with the Rule of Five. 75% of all approved drugs comply with every one of the individual physicochemical criteria defined by the Rule of Five.[10] A further 12% of approved drugs show only one violation of these rules. In contrast to that, only 61% of all human metabolites comply with all criteria and a further 15% violate only one of the criteria. For TCM molecules, the respective ratios are 44% and 15%. In particular the rule for the MW not to exceed 500 Da is often violated, with a maximum for TCM molecules of 43% (Table 1). This confirms that ADME filters can be useful in the preclinical stage, in particular for the screening of TCM molecules that are required to be orally bioavailable.

Molecular Complexity. Drug molecules show a degree of flexibility that is comparable to TCM molecules (20% vs 16% rotatable bonds), while human metabolites show an extended degree of flexibility (31% rotatable bonds). The number of rotatable bonds resembles a normal distribution (see Supporting Information Figure 4 for histogram plots of molecular complexity descriptors), with two-thirds of all drug molecules and 55% of all human metabolites having fewer than six rotatable bonds.

Related to the number of aromatic bonds described above, the number of rings is higher for TCM molecules (4.8) than for approved drugs (2.6) and human metabolites (1.7). TCM molecules also have a 3.3 and 1.9 times higher number of chiral centers, respectively. 45% of all approved drugs are nonchiral; 69% of them have fewer than two chiral centers. While the incidence of chiral centers decreases rapidly for approved drugs, TCM molecules show a broad distribution. Approved drugs are in general of lower molecular complexity than human metabolites and TCM molecules.

The results reported herein for the original (nonclustered) drug, human metabolite, and TCM molecule data sets are in good agreement with earlier published reports (in particular, refs 16 and 31). Very recently, Shen et al.[30] published a study investigating the physicochemical property distributions of data sets derived from an in-house TCM database. It is not feasible to make direct comparisons in this case because Shen et al. applied a molecular weight filter prior to their statistical analysis.

We believe that the data selection, curation, and diversification protocol employed in the current study adds to the significance and interpretability of the obtained results. Instead of analyzing a mixed data set comprising approved and experimental drugs (DrugBank) as used in earlier studies, we focus on the Approved Drugs subset of DrugBank. This is to avoid any bias introduced by the consideration of molecules with a potentially unfavorable metabolic profile. Another key feature of the current study is the optimized protocol for generating structurally diverse, representative data sets in order to remove any strong bias of the data sets toward certain molecular scaffolds. The impact of this clustering approach is apparent, for example, in the distribution of oxygen molecules (Figure 4b), which shows an unusual peak at 8 oxygen atoms per molecule for the nonclustered HMDB data set as a result of the accumulation of phospholipids. This peak is not observable for the clustered HMDB data set; the clustering protocol neutralizes this bias. Even though in the work of Khanna and Ranganathan[16] a clustering protocol to increase structural diversity was employed, a significant bias toward specific molecular scaffolds appeared to remain.

How Does Metabolism Alter the Physicochemical Property Space from Substrate to Metabolic Product (Question 2)? Here, we quantify the global shifts of physicochemical properties of molecules induced by metabolism. The results reported in this section are based on the analysis of experimentally observed metabolic schemes. Each metabolic scheme includes a single top-level substrate and all its terminal metabolites. It may also include one or several intermediates, but these are not considered for this analysis.

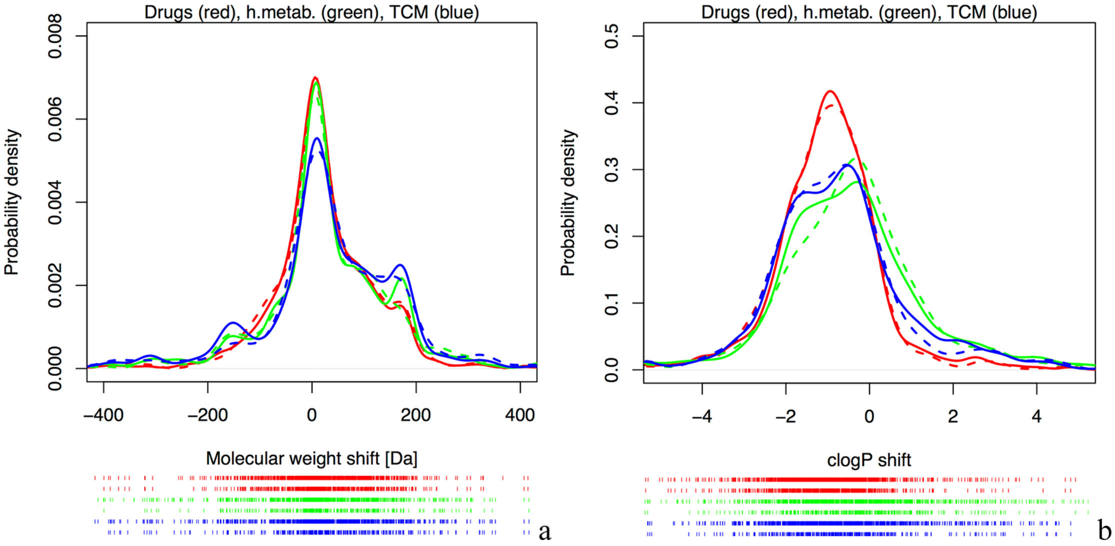


Figure 6. Shifts in physiochemical property space from a metabolic scheme-based perspective. Drug molecules (red), human metabolites (green), and TCM molecules (blue), for nonclustered (continuous lines) and clustered data sets (dashed lines) data sets. For each color, the upper rung represents the nonclustered and the lower rung represents the clustered data sets. (a) The main peak observed in the MW shift probability density (~16 Da) is to a large part a result of hydroxylation reactions. The two minor peaks at ~170 and ~−170 Da are related to glucuronidation and deglycosylation reactions, respectively. The latter is not present for approved drugs. (b) Terminal metabolites of approved drugs and TCM molecules have on average an about one log unit lower clogP compared to their parent molecules. This shift is much less pronounced for human metabolites.

Metabolic schemes are assigned to one or multiple chemical spaces according to the coverage of the top-level substrate by the representative data sets (Approved Drugs from the DrugBank, HMDB, and TCM Database@Taiwan). The amount of experimental data available for specific molecular scaffolds in general reflects its biological and pharmacological relevance. Due to this fact a certain bias is to be expected when overlapping the chemical space-specific data sets with the AMDB. We investigated this bias by comparing the MW and clogP distributions of the approved drugs, human metabolites, and TCM data sets with the data sets obtained from the overlapping regions of data. There is little to no bias observable for approved drugs (Supporting Information Figure 5a and d)—they are apparently very well covered by experimental data from the AMDB. Also human metabolites with MW < 500 Da are well covered, and, in particular, we do not see a problematic bias of the clustered data sets (Supporting Information Figure 5b and e). However, a significant bias is introduced for TCM molecules, where AMDB shows a better coverage of molecules with MW < 500 Da and lower clogP (Supporting Information Figure 5c and f). This does not necessarily constrain the merit of the data set since the data produced represent the fraction of molecules most interesting to pharmaceutical and related research.

After the assignment of metabolic schemes to chemical spaces, all individual shifts between the top-level substrate and its terminal metabolites are averaged and treated as individual data points for statistical analysis. Diverse subsets of metabolic schemes were generated using the top-level substrates as reference structures for clustering for maximum diversity (see Materials and Methods). All values reported herein refer to these diversified data sets, except where indicated. The most interesting results are summarized in the following section; more detail is provided in Supporting Information Table 4.

Basic Molecular Properties and Elementary Composition. Metabolism induces an average increase of the MW by 15−36 Da in the course of biotransformations leading from the top-level substrate to its terminal metabolites. Probability density plots allow the identification of major contributors to these shifts (Figure 6a).

Metabolism increases the average polarity and hence aqueous solubility of molecules. For approved drugs and TCM molecules, a negative shift of about one log unit in clogP is observed for approved drugs and TCM molecules. Human metabolites behave somewhat differently in this aspect. The average shift of clogP is smaller (−0.3 log units) and a noticeable portion of molecules shows an increase in hydrophobicity (Figure 6b). This may be a result of mechanisms for the retention of nutrients/micronutrients and their metabolites, whereas, on the other hand, excretion mechanisms ensure the rapid elimination of non-nutrients (xenobiotics) and their metabolites. An increase in the population of molecules with a low clogP value and higher MW is observed from inspection of the scatter plots in Supporting Information Figure 6.

While the number of nitrogen atoms is not substantially altered by metabolism, an increase of oxygen atoms is observed. For approved drugs, this increase is 1.7 oxygen atoms per molecule. This is comparable to TCM molecules (1.6), while human metabolites gain only 1.1 oxygen atoms on average.

Chemical Features. Two to three hydrogen bonding features are added by metabolism on average, which is a consequence of the introduction of oxygen atoms in the form of hydroxy groups, by hydroxylation and glucuronidation. The latter is also the main contributor to the increase of acidic atoms in terminal metabolites. On average, 0.7−1.0 acidic atom are introduced. No significant changes in the number of basic atoms are observed. This results in a net decrease of the formal charge by 0.4−0.5. Minor changes in the number of aromatic bonds are identified for approved drugs, which on average lose 0.5 aromatic bonds.

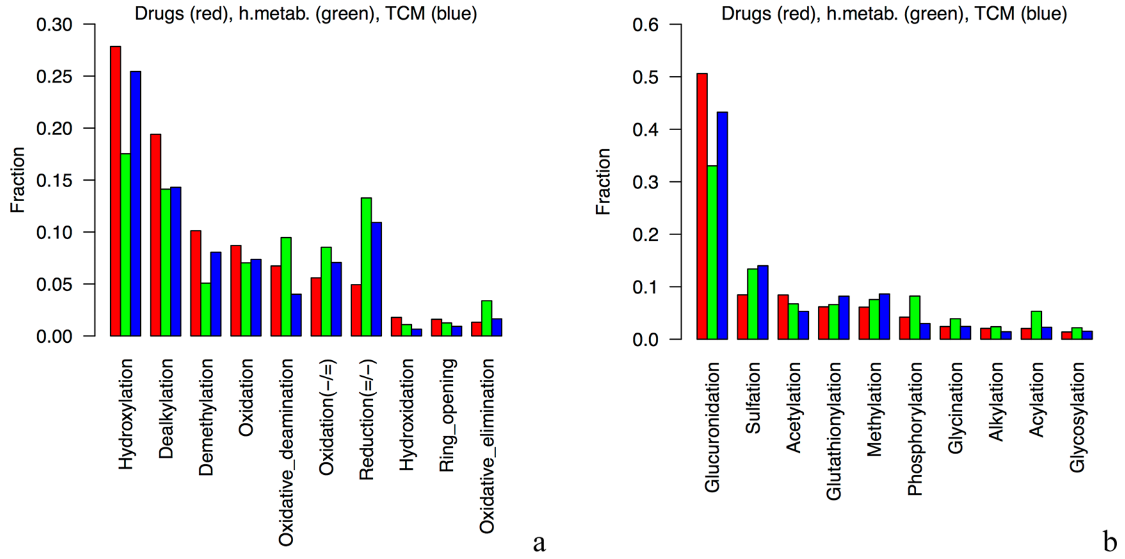


Figure 7. Most common (a) phase I and (b) phase II reactions for drug molecules (red), human metabolites (green), and TCM molecules (blue). Reaction types were determined using MetaPrint2D.[37,53] MetaPrint2D defines the reaction type "oxidation" as a metabolic reaction resulting in the addition of a double-bonded oxygen atom (and possibly also a further single/double-bonded oxygen atom), while the reaction type "oxidation (—/=)" denotes the transformation of a single bond into a double bond, and vice versa for the "reduction (=/—)".

Metabolite Likeness. An increase in metabolite likeness can be observed when comparing approved drugs with their terminal metabolites (Figure 5b)—we would expect this as drugs are often exposed to similar metabolizing enzymes (although this depends on which cell compartments are being considered). This is not the case for human metabolites and TCM molecules, which themselves contain more structural features that are characteristic of metabolites. For these chemical spaces, weak countertrends are found (i.e., their metabolites obtain lower metabolite-likeness scores than the parent compounds; see Supporting Information Figure 7).

Compliance with the Rule of Five. The percentage of molecules complying with any of the criteria of the rule of five drops by 15%, 10%, and 2% for approved drugs, human metabolites, and TCM molecules, respectively.

Molecular Complexity. Changes in the number of rotatable bonds and the number of ring systems are minor. A limited increase in the number of chiral centers per molecule can be observed, which tends to be higher for approved drugs (0.9) than for human metabolites (0.6) and TCM molecules (0.7).

What shifts in Physicochemical Property Space Do Individual Metabolic Reactions Introduce to Molecules (Question 3)? Here, we investigate and quantify the change in physicochemical properties introduced by single metabolic reactions. For this analysis, the chemical space classification of the top-level substrate was assigned to all individual reactions of a metabolic scheme. Reported are the individual shifts observed between a specific substrate (this includes top-level substrates) and its metabolite, averaged on a per-substrate basis. Data sets of biotransformations based on diverse molecular structures were generated using all substrates as reference structures for clustering for maximum diversity (see Materials and Methods).

Metabolic transformations are typically classified into phase I and phase II reactions. Phase I reactions are carried out by a variety of metabolic enzymes that catalyze chemical modifications of small organic molecules, such as oxidation, hydrolysis, reduction, ring closure, ring-opening, etc. Carboxyl, hydroxyl, amino, and sulfhydryl groups are often introduced by phase I reactions, which results in the attachment or modification of highly hydrophilic moieties which generally expedite excretion. Phase II biotransformations (generally conjugation reactions) often follow and the most common components for these are glucuronic acid, glutathione, glycine, acetyl, and methyl.

Phase I and phase II reactions differ substantially in the physicochemical shifts they introduce to molecules. In order to obtain a better resolution on the physicochemical property shifts, we have divided our data sets according to these two reaction categories. For the sake of clarity, we hereafter report on the statistical analysis of data based on the clustered data sets. The results summarized in this section are provided in Supporting Information Tables 5 and 6.

Propensity of Metabolic Reaction Types and Metabolic Enzymes. Phase I reactions are dominated by hydroxylation reactions (28%, 18%, and 25% for approved drugs, human metabolites, and TCM molecules, respectively), followed by dealkylation (19%, 14%, and 14%) and demethylation reactions (10%, 5%, and 8%; Figure 7a). The largest fraction of phase II conjugations is glucuronidation (51%, 33%, and 43% for approved drugs, human metabolites, and TCM molecules), which is followed by sulfation (8%, 13%, and 14%) and acetylation (8%, 7%, and 5%); see Figure 7b. The propensities for the reaction types are comparable among the different chemical spaces investigated.

About one-third of the reactions present in these data sets do have their catalyzing enzyme annotated. The large variety of metabolic enzymes and ambiguities in the enzyme name renders the data too sparse to allow a detailed analysis of the quantitative contributions of metabolic enzymes to physicochemical property shifts (at least from this data set). Cytochrome P450 is clearly the primary metabolizing superfamily of enzymes, with a propensity of 62%, 30%, and 50% for approved drugs, human metabolites, and TCM molecules, respectively. The proportion for approved drugs metabolized by cytochrome P450s is comparable to earlier reports, which estimate that more than 70% of all orally administered drugs are metabolized by cytochrome P450 3A4 and 2D6.[46] It should be noted that these numbers do not reflect reaction rates, which

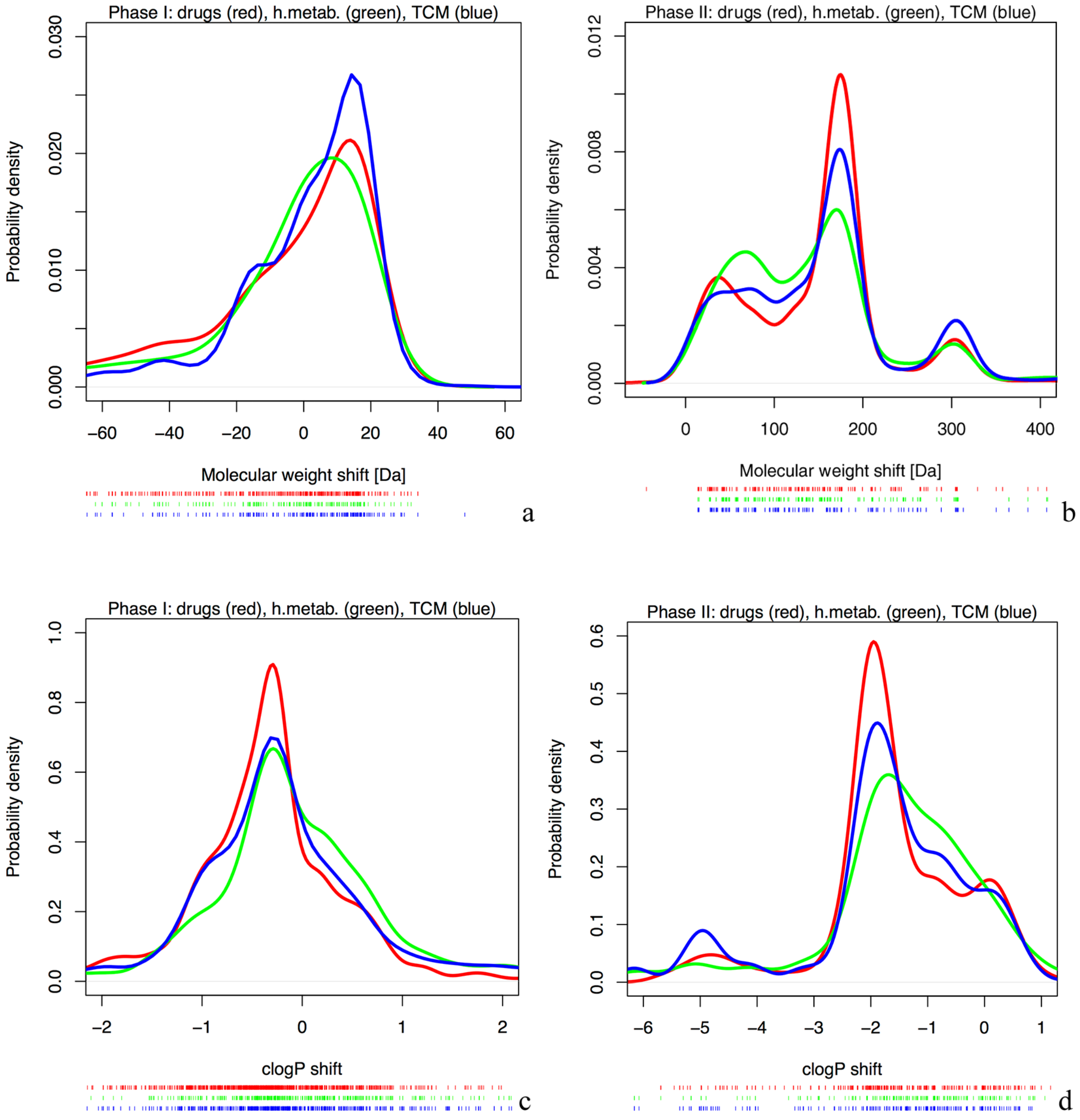


Figure 8. Shifts in MW and clogP for phase I and phase II reactions. Approved drugs (red), human metabolites (green), and TCM molecules (blue). (a) Phase I reactions have a tendency to produce metabolites of lower MW, while (b) phase II transformations lead to substantially heavier molecules. For phase II, three peaks can be identified, for the conjugation with glucuronic acid (main peak at ∼176 Da), methyl (peak at ∼14 Da), and glutathione (peak at ∼305 Da). (c) The peak of the clogP shift distribution at about −0.3 log units is a result of a variety of reactions such as aromatic hydroxylation, N-oxidation, demethylation, hydrogenation reactions and others. (d) Reductions of clogP are most significant for phase II reactions. The main peak at about −1.7 log units is mainly from glucuronidation reactions. A peak observable for approved drugs and human metabolites at ∼0.3 log units is from (O-)methylation reactions, and the second peak at about −4.8 log units (TCM molecules) is dominated by glutathionylation.

indicate an even more dominating role of cytochrome P450 enzymes in xenobiotic metabolism.[47,48] Interestingly, the percentage of molecules metabolized by this enzyme family is much lower for human metabolites, which points out their key role in xenobiotic metabolism, where, presumably due to first pass metabolism in the liver, xenobiotics are exposed to high concentrations of principally cytochrome p450s. Many endogenous compounds are metabolized in cellular compartments with different enzyme concentrations.

Basic Molecular Properties and Elementary Composition. Phase I and phase II reactions have a distinct impact on the MW of substrates. The distributions of shifts in MW caused by phase I reactions resemble skewed normal distributions (Figure 8a). Values are spread over a range of approximately −25 to 25 Da, peaking at a shift of ∼16 Da caused by hydroxylation. Overall, phase I reactions result in a loss in MW of 21−26 Da. Phase II reactions by definition result in an increase of molecular weight (up to ∼350 Da), with average gains of about 150 Da. See Figure 8b for an analysis of the major contributors and Supporting Information Figure 8a−c for scatter plots of the very distinct shifts of MW introduced by phase I and phase II reactions.

The negative shift in clogP introduced by phase I reactions is less than half a log unit (Figure 8c). Reduction of clogP is most apparent for phase II reactions, where reactions lead to a decrease by 1.5, 1.0, and 1.4 log units for approved drugs, human metabolites, TCM molecules, respectively (Figure 8d). For scatter plots of the shifts of clogP introduced by both metabolic phases, see Supporting Information Figure 8d−f.

Interestingly, not all metabolic reactions increase the predicted aqueous solubility of molecules. A minority of 4−9% of all phase I reactions and 8−13% of all phase II reactions result in more lipophilic metabolites. These biotransformations could be of interest, e.g., for the design of skin care cosmetics and pharmaceuticals that exhibit a prolonged pharmacological effect. It makes sense in evolutionary terms that metabolism in the skin is geared more toward increasing hydrophobicity. If logP goes up due to biotransformation in the skin, the more lipophilic metabolites likely stay in the skin, associated with lipids, and are less likely to partition/penetrate deeper into the (more aqueous) body. Once attached to lipids, they slowly move up to the stratum corneum, as all skin cells do, and are finally shed via desquamation. Hence, increasing the logP is an excretion mechanism in skin. Such biotransformations could hence be taken into account when designing new cosmetics and pharmaceuticals for skin targets.

The most frequently observed phase I metabolic reactions resulting in metabolites with a >1 log unit increase in clogP are dealkylation (33−46% of all reactions) and oxidative deamination (about 13% of all reactions); see Supporting Information Figure 9. In this context, we have analyzed the heterogenic shifts of MW and clogP introduced by common phase I reactions to approved drugs (Figure 9). In particular dealkylation reactions show a widespread effect on both properties.

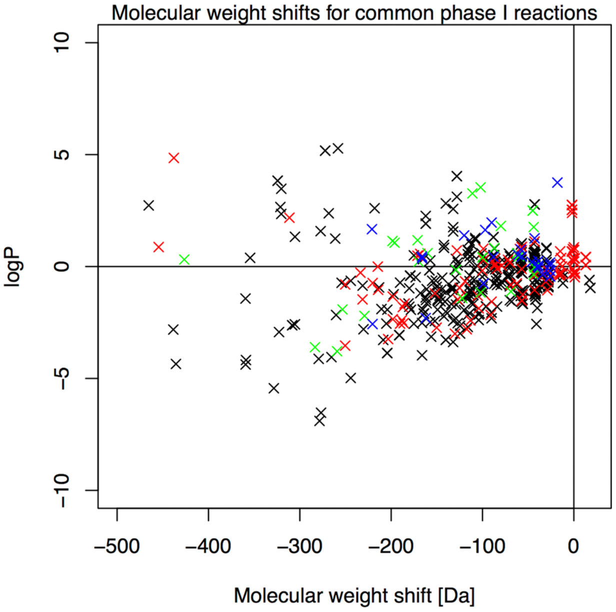


Figure 9. Shifts in MW and clogP for common phase I reactions: dealkylation (black), oxidative deamination (red), oxidative elimination (green), and elimination (blue).

Cytochrome P450s dominate these transformations, producing more lipophilic metabolites. Some members of the cytochrome P450 enzyme family are known to be major contributors to skin metabolism, such as CYP1B1, CYP2B6, CYP1A1, and CYP1A2, besides the primary drug-metabolizing cytochromes, CYP2D6 and CYP3A4.[49,50] A more detailed analysis of which other enzymes catalyze such reactions is confounded by the lack of data.

On average, the number of nitrogen and oxygen atoms is largely preserved for phase I reactions, while there is an average of about 0.5 nitrogen atoms and 4 oxygen atoms observed for phase II reactions.

Chemical Features. Shifts in the number of hydrogen bond donor and acceptor features closely resemble the shifts observed for the number of oxygen and nitrogen atoms described above. Phase I is largely balanced; phase II adds about four hydrogen bond acceptors and three hydrogen bond donors, respectively.

We can observe a tendency for a gain in acidic atoms by phase I biotransformations. This is a clear trend for phase II reactions, where on average 1.5−1.8 acidic atoms are added, resulting in a decrease of the average formal charge by about 0.7. The change in the number of aromatic bonds is negligible for phase I and II reactions.

Molecular Complexity. Phase I metabolism does not substantially alter the number of rotatable bonds, in particular when relating the shifts to the number of atoms per molecule. Phase II reactions increase the absolute number of rotatable bonds by 3.1−3.8, rings by 0.4−0.6, and chiral centers by 2.2−2.9.

What are the Physicochemical Properties of Metabolites Found in the Bile, Faeces, and Urine (Question 4)?

Here, we report on the physicochemical property space populated by metabolites found in bile, faeces, and urine, as annotated in the AMDB. Metabolites (including intermediates and terminal metabolites) were taken from metabolic schemes that are assigned to one or multiple chemical spaces. Diverse subsets were generated from cluster center molecules identified by clustering the metabolites for maximum diversity (see Materials and Methods). The most important properties are discussed below. For more data on a variety of physicochemical descriptors see Supporting Information Table 7.

Molecules found in the bile are on average 13−24% and 19−26% heavier than those present in the faeces or urine, respectively. The probability density diagrams of the three investigated chemical spaces indicate a trend for MW bile > faeces > urine (Figure 10a−c). For approved drugs the maxima are closely related, located just above 300 Da for bile and urine and, slightly higher, at about 315 Da for faeces. The distinctive feature of the three probability density distributions is the accumulation of metabolites with MW > 500 Da in the bile, while such large metabolites are only rarely found in the urine (6−8%). This is in good agreement with earlier estimates.[51,52] Nearly all of such large molecules found in the urine include a charged group, most of them being glucuronic acid conjugates and about one-third of them being zwitterionic. Other than that, only a few instances of peptidic conjugates and macrocyclic metabolites were found.

With respect to clogP, highly hydrophobic metabolites are more likely to be present in the bile or faeces than in the urine, which is consistent with the physiological requirements of the excretion systems. The differences are however only noticeable for high clogP ranges (Figure 10d−f). Also MW/clogP scatter plots do not indicate a significant accumulation within particular property ranges (Supporting Information Figure 10). There is a tendency for the excretion of charged molecules (in particular acidic molecules) via the bile (Supporting Information Table 7).

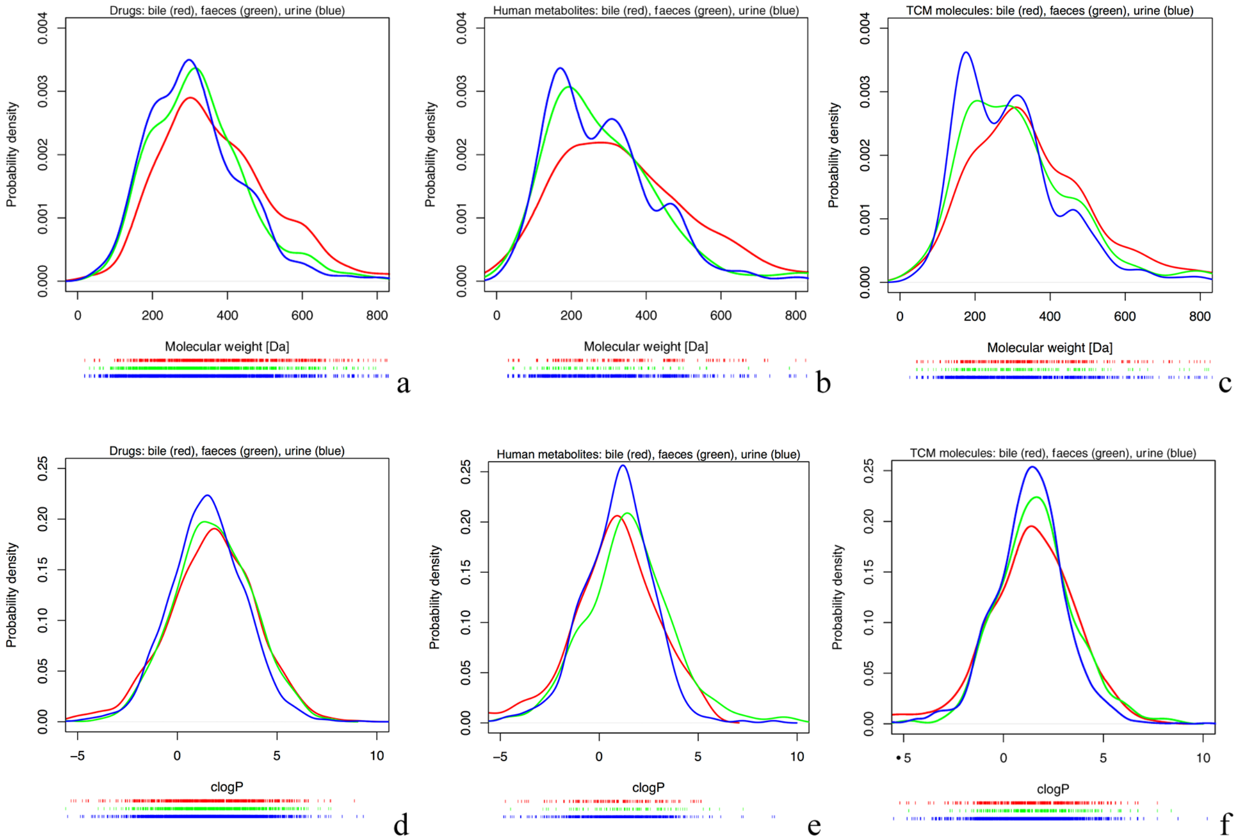


Figure 10. MW and clogP distribution of metabolites found in the bile (red), faeces (green), and urine (blue): (a, d) approved drugs, (b, e) human metabolites, and (c, f) TCM molecules. For small molecules, the individual probability densities are largely alike. Large metabolites are predominantly found in the bile, in particular when including a (negatively) charged group. Metabolites present in the urine mostly have a MW < 500 Da. If larger, they usually comprise one or more charged groups and are often zwitterionic (a−c). A trimodal distribution is observed for the MW of human metabolites and TCM molecules in the urine (b, c). The three peaks correspond to scaffolds such as purines and phenols (MW ∼ 170 Da), steroids (MW ∼ 330 Da), and glucuronidated steroids (MW ∼ 470 Da). Highly hydrophobic metabolites are more likely to be found in the bile and faeces (d−f).

■ CONCLUSIONS

We have investigated experimentally determined metabolic data on approved drugs, human metabolites, and molecules related to traditional Chinese medicine (TCM) to extend knowledge on four aspects of ADME decisive for pharmacokinetics and pharmacodynamics.

First, we compared the physicochemical property spaces of databases populated by these three distinct chemical spaces. In this context, we found approved drugs to be more closely related to human metabolites than TCM molecules. The latter differ from approved drugs and human metabolites by a ∼150 and ∼130 Da higher average molecular weight (this and all further values provided for clustered data sets), respectively. While only 13% of all approved drugs have a molecular weight >500 Da, 43% of all TCM molecules do not satisfy this Lipinski criterion and molecules with a molecular weight of up to 1200 Da are not uncommon. Also clogP is on average about 1.5−3 log units higher than for approved drugs and human metabolites. TCM molecules also contain fewer nitrogen atoms but are richer in oxygen atoms. They are more complex than approved drugs, which is obviously related to their biosynthetic origin and reflected e.g. by a higher prevalence of chiral centers.

The compliance of approved drugs with the rule of five (nearly 90% violate fewer than two criteria) is higher than for TCM molecules (59%) and human metabolites (76%). All these data are in alignment with the fact that human metabolites and TCM molecules are in general produced in the subcellular compartment in which they exhibit their biological function (in the respective organism, human and plant) and hence are not required to cross any membranes. In contrast to that, the ability to cross membrane barriers is decisive for drug molecules to reach their site of action, which substantially constrains their physicochemical property space. When using metabolites or TCM molecules as structural templates for the development of drugs, it is hence particularly important to optimize candidate molecules for adequate lipophilicity/solubility and molecular weight. A valid optimization strategy is to identify the pharmacophore of a bioactive (TCM) molecule and to systematically reduce the size of the molecule by eliminating moieties not essential for bioactivity and for a favorable ADME and toxicity profile. In this respect, substructure assembly based on fragment-based approaches are highly promising. Overall, this confirms the usefulness of simple ADME descriptors for prioritising candidate molecules for screening in the preclinical phase.

A metabolite-likeness model allowed good identification of human metabolites and to distinguish them from approved drugs. We observed an interesting difference between the DrugBank data set and the Approved Drugs subset: for the latter, fewer molecules obtained high metabolite-likeness scores. Drug molecules with a high metabolite-likeness score were found to be themselves endogenous metabolites or metabolite-mimetics, underpinning the validity of the model.

The second part of this study investigated the shifts in physicochemical property space induced by metabolism from a holistic perspective: How does an organic molecule absorbed by an organism differ from the terminal metabolites derived from this parent molecule? Metabolism as a global entity raises the molecular weight on average by 15−36 Da. The main drivers of this are hydroxylation and conjugation reactions such as glucuronidation, partly neutralized by deglycosylation, dealkylation reactions, and others. The clogP value of approved drugs and TCM molecules is lowered on average by ∼1 log unit while the effects on human metabolites are half or less. This discrepancy is interesting, as it highlights the function of retention mechanisms for nutrients/micronutrients and their metabolites and rapid excretion for xenobiotics and their metabolites. Careful integration of specific structural elements found in endogenous metabolites into xenobiotics may hence expedite metabolic stability without loss of membrane permeability.

Third, we explored the effect of individual metabolic reactions on the physicochemical properties. We therefore classified all biotransformations into phase I and phase II reactions for individual analysis. Dominated by hydroxylation and dealkylation reactions, phase I reactions lower the average molecular weight by 20−25 Da and tend to reduce lipophilicity. The majority of these reactions (about two-thirds in the case of approved drugs) are catalyzed by cytochrome P450s and the much higher proportion of drugs metabolized (at least orally, with first pass metabolism effects) by this enzyme family compared to human metabolites underlines their key role in xenobiotic metabolism.

Phase I biotransformations can function as precursors of subsequent conjugation reactions, which often use chemical functions introduced during phase I metabolism as reaction centers. These phase II reactions lead to a substantial gain in molecular weight (140−150 Da) and a decrease in clogP of about 1.5 log units for approved drugs and TCM molecules on average. This is lower for human metabolites, with a decrease of only one log unit on average, again indicating the evolution of a retention mechanism for endogenous metabolites. But not all biotransformations result in more hydrophilic metabolites. In particular, dealkylations and oxidative deaminations may result in more lipophilic products and these reactions could be used for the design of agents with extended pharmacological effects.

In the last part of this work, we investigated whether it is possible to distinguish metabolites found in the bile, faeces, and urine with respect to their molecular weight and lipophilicity. Our analysis indicates that differences are quite moderate across low molecular weight and clogP ranges, but some distinctive features can be identified. Large metabolites are predominantly found in the bile and they often contain one or more charged groups. Metabolites found in the urine mostly have a molecular weight of less than 500 Da, and if exceeding this threshold, they are usually charged or even zwitterionic. The absence of a clear preference for compounds to remain within a specific range of physicochemical properties is a strong indicator that transporters play a key role in excretion, in particular for compounds with physicochemical properties that do not meet the physiological parameters of the excretion system. It would be therefore particularly interesting to investigate the data for structural patterns and substructures specific to a particular excretion system. Sparsity of the data is likely to become an issue for these investigations.

Knowledge on the physicochemical property space of molecules with distinct biological functions, their metabolism and excretion has dramatically increased over the past decade. In this work, we quantified the shifts introduced by metabolism to physicochemical properties, from a holistic perspective and by analyzing individual phase I/phase II metabolic reactions. We believe that the observed trends add to the knowledge on the metabolic system and help to extrapolate and predict the likely metabolic fate of small organic molecules.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Ten figures illustrating the data processing scheme and a variety of physicochemical property distributions and shifts. Seven tables reporting on metabolic reaction types, data sets used for statistical analysis, physicochemical properties and descriptors of molecules included in the AMDB, Approved Drugs of DrugBank, HMDB, and TCM Database@Taiwan, physicochemical property shifts observed for each metabolic scheme, each phase I and each phase II reaction, and physicochemical properties of metabolites found in the bile, faeces, and urine. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rcg28@cam.ac.uk. Phone: +44 (1223) 336 432.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Unilever for funding and Accelrys and Schrödinger for software.

■ ABBREVIATIONS

ADME, absorption, distribution, metabolism, excretion; AMDB, Accelrys Metabolite Database; ASA, accessible surface area; HMDB, Human Metabolome Database; MOE, Molecular Operating Environment; MW, molecular weight; p450, cytochrome P450; TCM, traditional Chinese medicine; vdW, van der Waals

■ REFERENCES

(1) Schroeder, K.; Bremm, K. D.; Alépée, N.; Bessems, J. G. M.; Blaauboer, B.; Boehn, S. N.; Burek, C.; Coecke, S.; Gombau, L.; Hewitt, N. J.; Heylings, J.; Huwyler, J.; Jaeger, M.; Jagelavicius, M.; Jarrett, N.; Ketelslegers, H.; Kocina, I.; Koester, J.; Kreysa, J.; Note, R.; Poth, A.; Radtke, M.; Rogiers, V.; Scheel, J.; Schulz, T.; Steinkellner, H.; Toeroek, M.; Whelan, M.; Winkler, P.; Diembeck, W. Report from the EPAA workshop: In vitro ADME in safety testing used by EPAA industry sectors. *Toxicol. In Vitro* **2011**, *25*, 589−604.

(2) Hann, M. M.; Keserü, G. M. Finding the sweet spot: The role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug Discovery* **2012**, *11*, 355−365.

(3) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.* **2012**, *52*, 617−648.

(4) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(5) Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(6) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem., Int. Ed.* **1999**, *38*, 643−647.

(7) Feher, M.; Schmidt, J. M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.

(8) Stahura, F. L.; Godden, J. W.; Xue, L. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.

(9) Dobson, P. D.; Patel, Y.; Kell, D. B. 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discovery Today* **2009**, *14*, 31−40.

(10) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(11) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(12) Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, physical properties, and drug-likeness. *J. Med. Chem.* **2002**, *45*, 3345−3355.

(13) Oprea, T. I. Current trends in lead discovery: Are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.

(14) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.

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

(16) Khanna, V.; Ranganathan, S. Physiochemical property space distribution among human metabolites, drugs and toxins. *BMC Bioinf.* **2009**, *10* (Suppl 1), S10.

(17) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284−289.

(18) Peironcely, J. E.; Reijmers, T.; Coulier, L.; Bender, A.; Hankemeier, T. Understanding and classifying metabolite space and metabolite-likeness. *PloS ONE* **2011**, *6*, e28966.

(19) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Chretien, J. R.; Raevsky, O. A. Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and h-bonding descriptors. *J. Drug. Target* **1998**, *6*, 151−165.

(20) Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51*, 817−834.

(21) Benet, L. Z.; Broccatelli, F.; Oprea, T. I. BDDCS applied to over 900 drugs. *AAPS J.* **2011**, *13*, 519−547.

(22) Khanna, V.; Ranganathan, S. Structural diversity of biologically interesting datasets: A scaffold analysis approach. *J. Cheminf.* **2011**, 3:30.

(23) Nobeli, I.; Ponstingl, H.; Krissinel, E. B.; Thornton, J. M. A structure-based anatomy of the E. coli metabolome. *J. Mol. Biol.* **2003**, *334*, 697−719.

(24) Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **2003**, *125*, 11853−11865.

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

(26) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68−74.

(27) Eckert, H.; Bajorath, J. Exploring peptide-likeness of active molecules using 2D fingerprint methods. *J. Chem. Inf. Model.* **2007**, *47*, 1366−1378.

(28) Hao, G.; Dong, Q.; Yang, G. A comparative study on the constitutive properties of marketed pesticides. *Mol. Inf.* **2011**, *30*, 614−622.

(29) Medina-Franco, J. L. Interrogating novel areas of chemical space for drug discovery using chemoinformatics. *Drug Dev. Res.* **2012**, *73*, 430−438.

(30) Shen, M.; Tian, S.; Li, Y.; Li, Q.; Xu, X.; Wang, J.; Hou, T. Drug-likeness analysis of traditional Chinese medicines: 1. Property distributions of drug-like compounds, non-drug-like compounds and natural compounds from traditional Chinese medicines. *J. Cheminf.* **2012**, 4:31.

(31) Lopez-Vallejo, F.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. Expanding the medicinally relevant chemical space with compound libraries. *Drug Discovery Today* **2012**, *17*, 718−726.

(32) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. Drugbank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−1041.

(33) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; De Souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603−D610.

(34) Chen, C. Y. C. TCM Database @Taiwan: The world's largest traditional Chinese medicine database for drug screening in silico. *PloS ONE* **2011**, *6*, e15939.

(35) *Accelrys Metabolite Database*, version 2011.2; Accelrys, Inc.: San Diego, CA, 2011.

(36) *MOE*, 2010.10; Chemical Computing Group: Montreal, QC, 2011.

(37) Adams, S. E. Molecular similarity and xenobiotic metabolism. Ph.D. Thesis. University of Cambridge, UK. 2010.

(38) MetaPrint2D. http://www-metaprint2d.ch.cam.ac.uk (accessed Dec 11, 2012).

(39) Waring, M. J. Defining optimum lipophilicity and molecular weight ranges for drug candidates-molecular weight dependent lower logD limits based on permeability. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 2844−2851.

(40) Gleeson, M. P.; Hersey, A.; Montanari, D.; Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug Discovery* **2011**, *10*, 197−208.

(41) Benet, L.; Broccatelli, F.; Oprea, T. BDDCS applied to over 900 drugs. *AAPS J.* **2011**, *13*, 519−547.

(42) Sutherland, J. J.; Raymond, J. W.; Stevens, J. L.; Baker, T. K.; Watson, D. E. Relating molecular properties and in vitro assay results to in vivo drug disposition and toxicity outcomes. *J. Med. Chem.* **2012**, *55*, 6455−6466.

(43) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; Decrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4872−4875.

(44) Bar-Even, A.; Noor, E.; Flamholz, A.; Buescher, J. M.; Milo, R. Hydrophobicity and charge shape cellular metabolite concentrations. *PLoS Comput. Biol.* **2011**, *7*, e1002166.

(45) Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(46) Guengerich, F. P. Human cytochrome P450 enzymes. In *Cytochrome P450: Structure, mechanism, and biochemistry*, 3rd ed.; Ortiz de Montellano, P. R., Ed.; Kluwer Academic/Plenum Publishers: New York, 2005; pp 377- 530.

(47) Wilkinson, G. R. Drug metabolism and variability among patients in drug response. *N. Engl. J. Med.* **2005**, *352*, 2211−2221.

(48) Slaughter, R. L.; Edwards, D. J. Recent advances: The cytochrome P450 enzymes. *Ann. Pharmacother.* **1995**, *29*, 619−624.

(49) Oesch, F.; Fabian, E.; Oesch-Bartlomowicz, B.; Werner, C.; Landsiedel, R. Drug-metabolizing enzymes in the skin of man, rat, and pig. *Drug Metab. Rev.* **2007**, *39*, 659−698.

(50) Svensson, C. K. Biotransformation of drugs in human skin. *Drug Metab. Dispos.* **2009**, *37*, 247−253.

(51) Spalding, D. The importance of the physicochemical properties of drugs to drug metabolism. In *A handbook of bioanalysis and drug metabolism*, Evans, G., Ed.; CRC Press: Boca Raton, 2004; pp 8−31.

(52) Gordon Gibson, G.; Skett, P. *Introduction to drug metabolism*, 3rd ed.; Cengage Learning: Andover, 2001; pp 37−86.

(53) Carlsson, L.; Spjuth, O.; Adams, S.; Glen, R. C.; Boyer, S. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinf.* **2010**, *11*, 362−362.