Tetrahedron 66 (2010) 1031–1039

Contents lists available at ScienceDirect

Tetrahedron

journal homepage: www.elsevier.com/locate/tet

Toward molecular design for hazard reduction—fundamental relationships between chemical properties and toxicity

Adelina M. Voutchkova[a], Lori A. Ferris[b], Julie B. Zimmerman[b,c], Paul T. Anastas[a,b,*]

[a] *Department of Chemistry, Yale University, New Haven, CT 06511, USA*
[b] *School of Forestry and Environmental Studies, Yale University, New Haven, CT 06511, USA*
[c] *Environmental Engineering Program, Yale University, New Haven, CT 06511, USA*

ARTICLE INFO

Article history:
Received 7 July 2009
Accepted 4 August 2009
Available online 14 November 2009

ABSTRACT

The relationship of in-silico predicted physical/chemical properties and human toxicity is analyzed for a statistically significant sample size of chemical compounds. Results for compounds with known toxicity endpoints, as designated by EPA's Toxic Release Inventory (TRI), are compared to a series of commercial chemicals that are not regulated under TRI. Physical properties for all compounds are predicted using Schrodinger's QikProp, an established tool for predicting adsorption, distribution, metabolism, and excretion (ADME) characteristics. The results of this analysis indicate that the physical/chemical property distributions of TRI chemicals are statistically significantly different from those of bulk commercial chemicals, particularly related to properties associated with bioavailability. Using a partitioning analysis, several key physical/chemical properties and ranges are identified that can be used to readily differentiate TRI chemical characteristics from those of bulk commercial chemicals.

1. Introduction

It is likely that there is no more greatly studied characteristic of molecules than their ability to exhibit biological activity. Several industry sectors, including pharmaceuticals and pesticides, are based on this science. Billions of dollars are spent to evaluate the toxicity of chemicals in the environment and billions more are spent by NIH to discover new chemicals that have therapeutic biological effects.[1] With over a century of scientific focus, the goal of being able to design molecules from first principles with controlled biological activity is still viewed as an immense challenge.[2] With the recent advances in understanding the mechanisms of toxicity,[3] new strides in the field of molecular design can be made.

To this end, an analysis of the relationships of the physical/chemical properties that are most closely linked with established biological activity would be highly informative. Such an analysis is possible due to the availability of empirical and modeling data generated in the study of potential drug targets and environmental toxins. The fields of pharmacology and toxicology are essentially chiral disciplines whose respective tools can both contribute to this investigation. One field focuses on how to maximize a specific targeted therapeutic function, while the other seeks to measure, understand, and even predict adverse biological effects. Our purpose is to combine the insights of these two fields to establish design strategies to achieve wide ranges of chemical function while avoiding biological activity, thereby designing safer chemicals.

Of all of the chemicals that are currently in commerce today, it is estimated that >99% have no *intentional* biological activity.[†] Yet, molecular 'design for function'[4] has not systematically incorporated design for intrinsic safety as a performance criterion.[5] The objective of this study is to begin to inform the generation of heuristic design rules that would increase the probability of reducing or eliminating inherent human health hazard based on an understanding of the relationship between physical/chemical properties, molecular structure, and toxicity.

Despite the mechanistic complexity of biological action, in 1997 Lipinski formulated the 'Rule of Five' of druglikeness[6] of approximately 900 commercial pharmaceuticals. These 'Rules' describe the predominant physical and chemical property ranges of successful pharmaceutical candidates, including molecular weight, log $P_{o/w}$ (octanol/water partition coefficient) and the number of hydrogen bond donors and acceptors. Lipinski's rules have become accepted by the pharmaceutical industry as standards for screening lead active compounds. A significant body of work has been devoted to furthering understanding of how these physiochemical properties are associated with favorable drug toxicokinetics.[7–10] Based on

* Corresponding author. Tel.: +1 203 432 6165.
E-mail address: paul.anastas@yale.edu (P.T. Anastas).

[†] Based on 48M commercially available chemical substances registered in the CAS Database and ~23,000 FDA-approved drugs on the market.

0040-4020/$ – see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.tet.2009.11.002

these studies, the properties Lipinski described have been associated with the potential for bioavailability, ADME (Adsorption, Distribution, Metabolism, and Excretion), and favorable toxicokinetics. Some properties, such as molecular weight, have also been well-correlated with bioavailability and have established cut-offs.[11] Many other properties, such as surface area, are known to have effects on bioavailability but have no definitive limits.

In addition to Lipinski's rules, Quantitative Structure Activity Relationships (QSARs), statistical correlations between biological activity of fragments or whole molecules, and one or more physiochemical properties or descriptors, have also been applied extensively in both pharmacology[12] and toxicology.[13] In toxicology, they have been used to facilitate toxicity and fate predictions by associating physical properties with either specific biochemical interactions that are part of toxicity mechanisms (e.g., enzyme binding),[14] or in some limited cases, with whole-organism non-specific baseline toxicity.[15] This is distinct from our efforts, which do not aim to describe predictive models for toxicity, but rather to attempt to derive property limits associated with reduced toxicity.

To derive a set of property criteria associated with toxic compounds the vast complexity of toxicity mechanisms must be considered. This complexity is partially derived from the complexity of the mode of action of biologically-active compounds in whole organisms. In this case, the focus is on toxic, rather than therapeutic endpoints. However, there are significant similarities in these two objectives: that is, designing for therapeutic effect and minimizing toxic endpoints. For a molecule to act either as a therapeutic or a toxic agent, it must first have a mode of entry into the organism. The main routes of entry in humans are four—gastrointestinal tract, skin, eyes, and lungs. The extent to which the chemical is absorbed into the bloodstream after entering through one of these routes is its bioavailability. This is to say, toxicity, for the most part, is directly dependent on the extent of the compound's bioavailability.[‡] A compound must be bioavailable to be potentially toxic, but often at much lower blood concentrations than pharmaceutical agents.[16]

An understanding of bioavailability is therefore critical when considering how to alter physical/chemical properties so as to reduce or eliminate inherent toxicity. To reduce toxicity potential, in addition to reducing bioavailability, one can also increase excretion/reduce storage, reduce the rate of distribution, inhibit bioactivation pathways/promote detoxification pathways, and reduce rate of toxicodynamic interactions (covalent modifications). These steps are illustrated in Figure 1.

Medicinal chemistry has shown that most biological interactions (such as decreased bioavailability or increased detoxification) can be associated with specific physical and chemical properties of potentially toxic compounds. Yet, to the best of our knowledge, an analysis analogous to Lipinski's that associates physical/chemical properties with their toxicity has been not been reported.

In the current communication, a potential base set of physical/chemical property value ranges that statistically correlate with adverse biological activity is reported. Physiochemical properties are computationally predicted for three groups of chemicals—one with established human toxicity, one consisting of commercial drugs, and a third that encompasses a broad range of commercially available chemicals. The property distributions of these three datasets are analyzed. For certain properties the toxic group of compounds shows fairly narrow value distributions, providing insight to the feasibility of a physical/chemical property-based understanding of human toxicity. A comparison of physiochemical properties of the toxic and drug groups also yields insights that in some cases these distributions are distinct. This analysis is the first step to establishing value ranges of properties associated with reduced toxicity. Such property 'rules' can be directly used to inform the designers of molecular structures to minimize or reduce toxicity and increase the safety of synthetic chemicals generally.

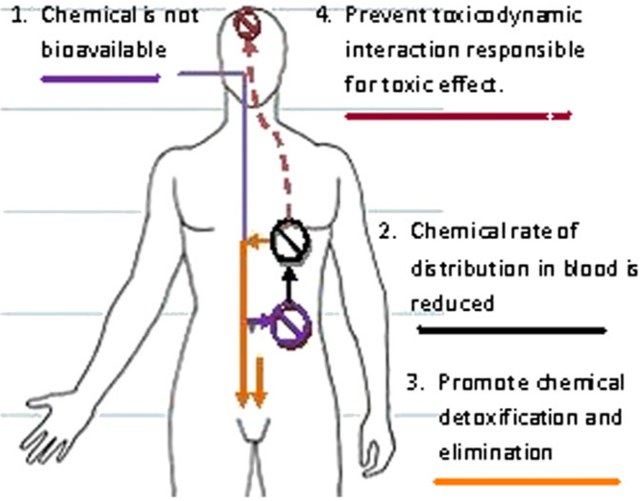
1. Chemical is not bioavailable
4. Prevent toxicodynamic interaction responsible for toxic effect.
2. Chemical rate of distribution in blood is reduced
3. Promote chemical detoxification and elimination

Figure 1.

2. Experimental

2.1. General

This work analyzes computationally predicted physiochemical properties of three groups of chemicals: (1) toxic compounds, with established human toxicity, (2) commercially available active pharmaceutical ingredients (APIs) of drugs, and (3) a broad group of commercially available chemicals whose toxicity is in most cases unknown. Physiochemical properties are predicted computationally using a software package well-established for development of APIs, Schrodinger's QikProp.[17] Subsequently, the biological activity of these compounds based on their physical/chemical characteristics was predicted. The property distributions of the three datasets are analyzed using SAS Institute JMP software version 7.0.1.

2.2. Property predictions

Property predictions were carried out using Schrodinger's QikProp version 2.4, a well-established program utilized in the field of drug discovery to provide predictions for significant physical descriptors and pharmaceutically-relevant biological properties of neutral organic molecules via semi-empirical methods. Multi-structure 2D SD files were generated for the 625 TRI compounds, 546 pharmaceuticals, as well as the 13 M compounds from the virtual screening database ZINC, and converted to 3D structures using the freeware molecular coordinate converter program Babel.[18] Where necessary, molecular structures were optimized prior to property predictions with an AM1 basis set using Gaussian 03 W software.

2.3. Chemicals of known toxic effect

The 625 chemicals listed in the U. S. Environmental Protection Agency's (EPA) Toxic Registry Inventory (TRI) were chosen as a toxic dataset. The Pollution Prevention Act (PPA) of 1990 mandates collection of data on toxic chemicals that are treated, recycled, and combusted for energy recovery. Together, these laws require facilities in certain industries, which manufacture, process, or use toxic

[‡] Exceptions include chemicals that have topical effects, such as ones that are corrosive, or ones that cause skin sensitization.

chemicals above specified amounts, to report annually on disposal or other releases and other waste management activities related to these chemicals.[19] Of the 625, QikProp could not analyze 57 compounds, which are organometallic compounds or inorganic salts. The total number of TRI compounds analyzed in this study was therefore 568.

2.4. Chemicals of known therapeutic effect

A group of 546 oral commercially available drugs were chosen to represent chemicals with known therapeutic effect. The Active Pharmaceutical Ingredients (APIs) of these drugs represent a variety of therapeutic effects, and are either proprietary or generic.

2.5. Commercially available chemicals with no experimental toxicity data

As a large bank of chemical compounds and properties the ZINC database was utilized. The ZINC database, provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF), contains about 13 M commercially available molecules. Medicinal chemists often utilize it as reference bank for screening active pharmaceuticals. As a result, about 8 M of its compounds are considered drug-like in that they meet Lipinski's rules.

From ZINC Database of compounds (a) organometallic compounds and inorganic salts, and (b) biological polymers with poorly defined chemical structure were eliminated due to the inability of QikProp to predict the properties of these classes of compounds. Properties of all the remaining compounds in the database were predicted using QikProp. 133,000 compounds were randomly selected from ZINC for this analysis because inclusion of additional chemicals did not yield significant differences in the means or distributions of the properties reported.

2.6. Physical/chemical properties

The physical/chemical properties evaluated include molecular weight; number of freely rotating bonds; number of reactive functional groups; partition coefficients for octanol/water ($P_{o/w}$), water/gas ($P_{w/g}$), and octanol/gas ($P_{o/g}$); aqueous solubility; Solvent Accessible Surface Area (SASA) and its hydrophobic components (FOSA); polar surface area (PSA); globularity; molecular volume; number of H-bond donors and acceptors, polarizability; dipole; electron affinity (EA); ionization potential (IP); number of rotatable bonds; and number of atoms in rings.

2.7. Biological properties

QikProp was used to predict some properties that are considered most useful in predicting biological activity in the field of drug discovery. These include apparent cellular permeability, in nm/s, of two cell lines—Madin-Darby canine kidney (MDCK) cells (Affymax scale) and Caco cells. MDCK cells are considered to be a good model for the blood–brain barrier,[20] while Caco intestine cells model absorption across the gut–blood barrier. It should be noted that QikProp predictions on MDCK cells are for orally-delivered drugs and behavior for both MDCK and Caco predictions only take into account non-active transport. Chemicals with cellular permeability values for both cell lines exceeding 500 nm/s are considered to be well-absorbed through the respective barriers.[21]

The potential blocking ability of K^+ channels, encoded by the human ether-a-go-go related gene (*HERG*), is often a potential red flag for toxicity.[22] QikProp was able to predict the log IC_{50} (half maximal inhibitory concentration) values for blockage of mammalian HERG K^+ channels.

Skin permeability is a similarly important biological property that was predicted by QikProp, as it governs dermal absorption. Finally blood–brain partition coefficient was calculated.

2.8. Statistical methods

The ANOVA (ANalysis Of VAriance between groups) method is a commonly utilized statistical test that determines whether means of several groups are all equal. ANOVA was used to compare the predicted mean values for the physical/chemical and biological properties discussed previously for the three groups of chemical compounds. To provide a measure of the range of the properties, the 2.5% and 97.5% quantile values were used.

The statistical technique of partitioning was used to determine which physiochemical properties can distinguish the TRI chemicals form the ZINC group. Partitioning was performed in JMP. The algorithm recursively partitions data according to a relationship between the physiochemical properties (*X* variables) and the TRI or ZINC designation of each compound (*Y* value), creating a tree of partitions. It finds a set of cuts or groupings of *X* values that best predict a *Y* value. It does this by exhaustively searching all possible cuts or groupings. These data partitions are done recursively, forming a tree of decision rules, until the desired fit is reached.

G^2 is an indication of variation in the data. R^2 is a measure of the amount of variation explained by splitting the data into groups. The Receiver Operating Characteristic (ROC) curve is used to describe the goodness of fit of the partitioning. It represents the count of True Positives by False Positives as one accumulates the frequencies across a rank ordering. The True Positive *y*-axis is labeled 'Sensitivity' and the False Positive *x*-axis is labeled '1-Specificity.'

3. Results and discussion

In attempting to understand the relationship between human toxicity, mechanism of action, and physical/chemical properties, a preliminary analysis of the typical property value ranges for toxic chemicals was performed. Given that prediction methods tested against experimental datasets do exist, calculated values were utilized. QikProp's prediction capabilities were validated with the available experimental data for the TRI dataset and for the properties of log $P_{o/w}$, polar surface area, molecular volume, number of rotating bonds, and hydrogen donors/acceptors. The predicted and experimental values were highly correlated with R^2 values of 0.96–0.99. In concurrence with previous studies,[21] optimization of molecular structures with at least a semi-empirical level of theory prior to property predictions was undertaken.

3.1. Predicted physical/chemical properties

The physical/chemical properties analyzed can be grouped into five broad descriptive categories: size; shape; flexibility/rigidity; electronic nature; and solubility both in water and organics. The results of predicted properties for the three sets of chemicals under consideration, toxic, therapeutic, and bulk/commercial, are presented according to these categories.

3.1.1. Molecular size. The link between size and absorption is related to the biological barriers involved in each absorption route (i.e., dermal, gastrointestinal, pulmonary, and ocular).[8] Compounds with molecular weight >400 Da, for example, are considered incapable of being absorbed by the skin,[11] while those with MW >500 Da cannot generally cross the GI tract and enter the bloodstream.[23,24]

Since the TRI chemicals are single molecule entities, molecular size can be approximated by molecular weight and shape. The molecular weight distribution of the TRI chemicals was significantly different from that of the ZINC set (Fig. 2). The 2.5 and 97.5%

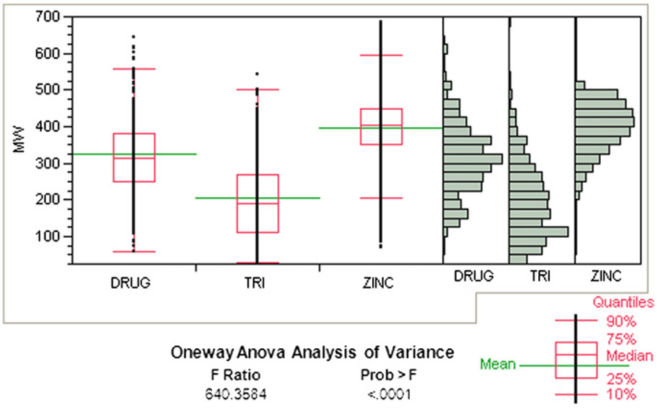

Figure 2.

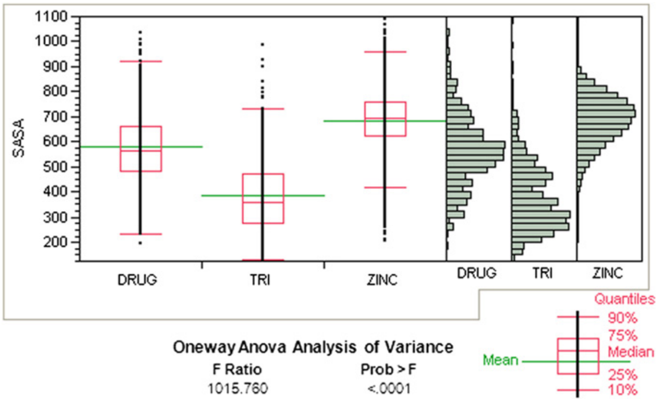

Figure 3.

quantiles observed for the distribution of molecular weights of the TRI compounds fall within 43–461 Da (Table 1). These results indicate that the weights of the large majority (greater than 92%) of the TRI compounds fall well within the molecular weight range for likely absorption. A hypothesis explaining the outliers is that they are compounds that contain hydrolysable linkages, which facilitate their degradation in the stomach to smaller compounds. These metabolites can then pass through the GI tract. The molecular weight distribution of the drug compounds falls between those of the TRI and ZINC, with a mean of 331 Da.

Table 1

Physical and chemical properties of the 568 TRI compounds compared with 546 commercial drug compounds and 133,667 randomly selected compounds from the ZINC Database

Property category	Property	Mean (2.5 and 97.5% Quantiles)			Range or recommended values for drugs[a]
		TRI (toxic) compounds	Drug compounds	ZINC database compounds	
Flexibility	Number of rotatable bonds	2.07 (0, 10.00)	5.48 (0, 15.20)	5.66 (1.00, 11.00)	0–15
	Number of ring atoms	6.92 (0, 22.00)	13.64 (0, 28.00)	16.87(6.00, 27.00)	—
Shape	Globularity	0.91 (0.79, 0.99)	0.84 (0.74, 0.93)	0.80 (0.74, 0.89)	0.75–0.95
Electronics	Dipole (Debye)	4.45 (0, 18.12)	5.23 (0.98, 11.72)	6.42 (1.52, 13.63)	1–12.5
	HB donors	0.62 (0, 3.00)	1.48 (0, 5.00)	1.07 (0, 3.00)	0–6
	HB acceptors	2.68 (0, 9.53)	5.79 (1.54, 14.54)	6.46 (2.53, 11.00)	2–20
	Polarizability (Å^3)	18.32 (2.56, 43.53)	34.33 (13.82, 63.83)	40.84 (23.83, 53.53)	13–70
	IP (eV)	9.44 (5.98, 11.81)	9.01 (7.86, 10.59)	8.85 (7.61, 9.83)	7.9–10.5
	EA (eV)	0.41 (−3.25, 3.46)	0.53 (−0.81, 1.79)	0.83 (−0.22, 1.89)	−0.9 to 1.7
Size	Molecular weight (Da)	206.21 (43.23, 461.17)	331.67 (112.27, 645.21)	395.88 (234.12, 504.12)	130–725
	Solvent Accessible Surface Area (Å^2)	384.12 (176.63, 714.98)	596.23 (349.56, 999.27)	684.64 (461.23, 857.77)	300–1000
	FOSA (Å^2)	120.23 (0, 441.65)	271.45 (0, 650.88)	265.36 (26.25, 531.66)	0–750
	Polar surface area (Å^2)	36.84 (0, 122.26)	77.22 (6.52, 192.56)	84.62 (29.34, 149.76)	7–200
	Volume (Å^3)	617.34 (216.65, 1273.29)	1021.60 (447.03, 1932.34)	1202.76 (768.45, 1530.06)	500–2000
Solubility	log $P_{oct/water}$	2.03 (−2.14, 6.99)	2.47 (−1.96, 6.61)	3.59 (0.56, 6.28)	−2.0 to 6.5
	log $P_{oct/gas}$	10.75 (1.30, 22.70)	16.87 (6.63, 36.08)	19.25 (11.33, 25.97)	8–35
	log $P_{water/gas}$	5.08 (0.16, 12.24)	10.36 (2.79, 25.07)	11.06 (5.17, 17.64)	4–45
	log $P_{water/sol}$	−2.31 (−9.11, 1.79)	−3.50 (−7.39, −0.01)	−5.30 (−8.22, −1.94)	−6.5 to 0.5

[a] For 95% of known drugs.[21]

Since solvent accessible surface area (SASA) should be correlated to molecular weight, a similar trend to molecular weight is predicted. Although SASA is also known to have an effect on bioavailability,[25] the limits of SASA beyond which a compound becomes significantly less bioavailable are not well defined. The 2.5–97.5% limits of the TRI compounds fall between 176 and 714 Å^2 (Fig. 3). These values are significantly different from those of the commercial chemicals, and the means of all three groups are significantly different (Fig. 2, ANOVA). The drugs are larger than the TRI chemicals on average, with a range between 128 and 664 Å^2.

The non-polar component of SASA (in Å^2) is on average smallest for TRI, followed by the drug and the ZINC groups. However, as a fraction of the SASA its variability is not very high −31%, 45% and 38%, respectively, for the three groups.

Polar surface area (PSA) is considered to be an important property in prediction of oral bioavailability of drugs.[25] In fact, it has been reported that membrane permeability can better be correlated to PSA than molecular weight.[25–29] Veber[9] reported that active pharmaceuticals after oral administration generally showed bioavailability exceeding 20% in rats. Datasets for the artificial membrane permeation rate and for rate of clearance (elimination from body through excretion) in the rat also showed that reduced polar surface area correlates better with increased permeation rate than does lipophilicity (log $P_{o/w}$).[9] From the predicted property data for TRI compounds, there is a predominance of compounds with low PSA, with an upper 97.5% limit at 122 Å^2 (Fig. 4). Since it falls well within the limit of <140 Å^2, the set of TRI compounds appear to be orally bioavailable.

3.1.2. Molecular shape. The predicted molecular descriptor that best describes molecular shape was globularity. Globularity is defined as:

$$Glob = 4\Pi r^2/SASA$$

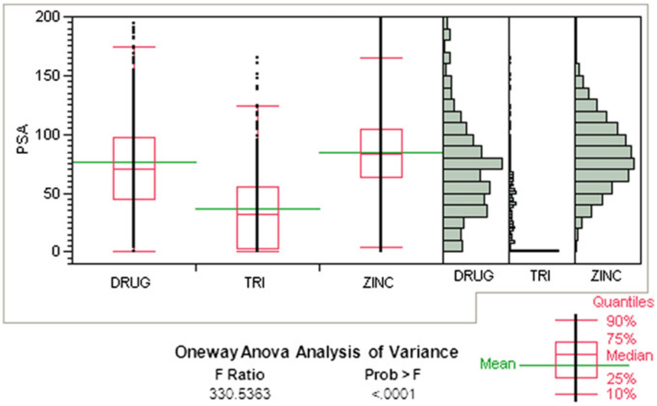

Figure 4.

where r=radius of a sphere with a volume equal to the molecular volume. This implies that highly spherical molecules have globularity values approaching 1.

The 2.5–97.5% limits of the globularity of the TRI chemicals were 0.79–0.99 with a mean of 0.908, significantly higher than the mean of globularity of the randomly sampled ZINC database (Glob=0.80, Fig. 5). Both of these distributions were significantly different from that of the orally-administered drugs, whose range fell between 0.74 and 0.93. For all three datasets, but especially the drugs, these ranges fall within the desirable range for drug candidates of 0.75–0.95.[21] The mechanistic connection between bioavailability and globularity still remains to be experimentally explored, but some feasible possibilities include (a) a direct influence on membrane transport and/or (b) an indirect influence on bioavailability by affecting water and lipid solubility depending on the polarity of the SASA.

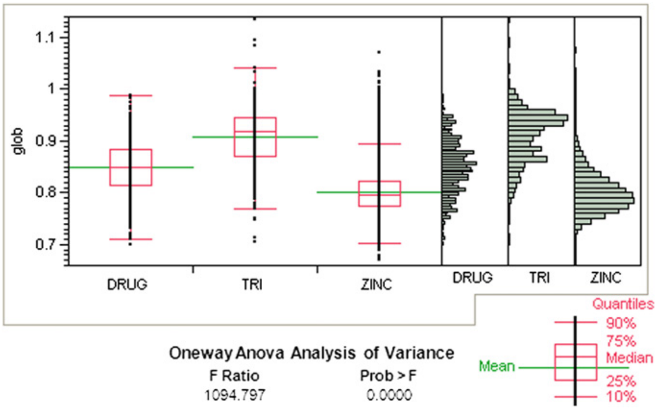

Figure 5.

3.1.3. Flexibility. Understanding the impact of molecular rigidity (or flexibility) on bioavailability is not as straightforward as understanding the impact of molecular size. It is possible that increased rotational degrees of freedom may result in a larger diffusional cross section, adversely influencing membrane permeability.[23] However, flexibility could also decrease crystallinity of the solute, resulting in improved aqueous solubility and enhanced absorption.[30] It has been reported[9] that reduced molecular flexibility, as measured by the number of rotatable bonds, along with other electronic parameters, can be used as an important predictor of good oral bioavailability, independent of molecular weight. Veber[9] reported that active pharmaceutical compounds with less than 10 freely rotatable bonds were most likely to have high bioavailability. An analysis of the distribution of the number of freely rotatable bonds in the TRI chemicals indicates that the 95% limit is nine freely rotatable bonds (Table 1). This indicates that the TRI chemicals either are very rigid (contain rings and double bonds), and/or have low molecular weights (smaller molecules are expected to have fewer freely rotatable bonds) (Fig. 6).

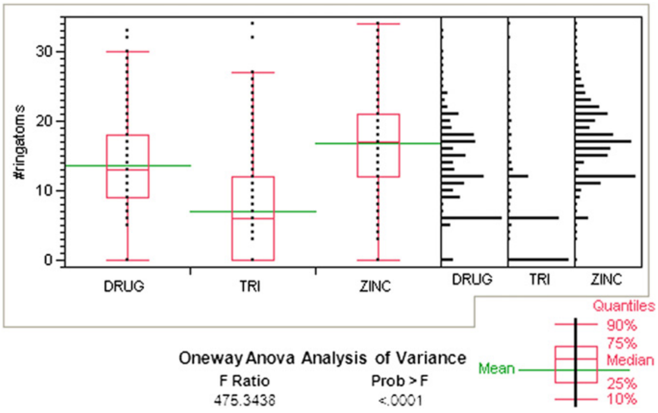

Figure 6.

Since rings are generally considered to contribute to molecular rigidity, the number of ring-associated atoms for each compound can be used to compare rigidity. This difference can also be captured in the means of the two distributions (6.92 vs 13.6 atoms in rings/compound). This difference could be attributable to the smaller average molecular size of the TRI compounds compared to the ZINC. Even though most of the rings in TRI compounds are five- or six-membered, an estimated 6% of the compounds do have atoms in three or four member rings. This is significantly different from the distribution of the compounds in the ZINC database, in which only 3% have any atoms in three or four member rings. Such strained rings (three or four member rings), as those found more predominantly in the TRI chemical set, tend to be highly reactive and therefore if they can become bioavailable, have the potential to cause intrinsic damage to cellular macromolecules.

3.1.4. Electronic properties. Considering that most toxicity mechanisms involve some specific toxicodynamic interaction, such as covalent binding to cellular constituents like DNA or proteins, toxic compounds would likely exhibit trends in electronic properties, such as polarizability, dipole moment, and electron affinity. Less clear, however, is precisely how electronic properties also influence bioavailability by specific routes. An indirect effect on bioavailability is expected through an effect of solubility. If one includes hydrogen bond (HB) donors and acceptors in the group of electronic properties, there is the likelihood for indirect effects on water solubility and thus octanol/water partitioning, subsequently influencing bioavailability. Lipinski's rules state that bioavailable compounds have no more than 5 HB donors and 10 acceptors. The TRI dataset indicates a corresponding 97.5% limit of 3 HB donors and 9.5 acceptors (Fig. 7).

Also of note is that the TRI chemicals have significantly fewer nitrogen (N) and oxygen (O) atoms than the average compounds from the ZINC database. A lack of N and O atoms also influences polarity and relates to the smaller observed dipole moments for these compounds. The distribution of dipole moments of the TRI compounds versus those in the ZINC database appears to bias compounds with small dipole moments, with a 97.5% limit of 18.1 D. Although the 90% limit of the ZINC dataset is only 13.6 D, the latter set has a pseudonormal distribution at a maximum of 5 D, while the TRI group has an exponentially decreasing distribution, with a mode of 0–1 D (Fig. 8). For comparison, the highly toxic

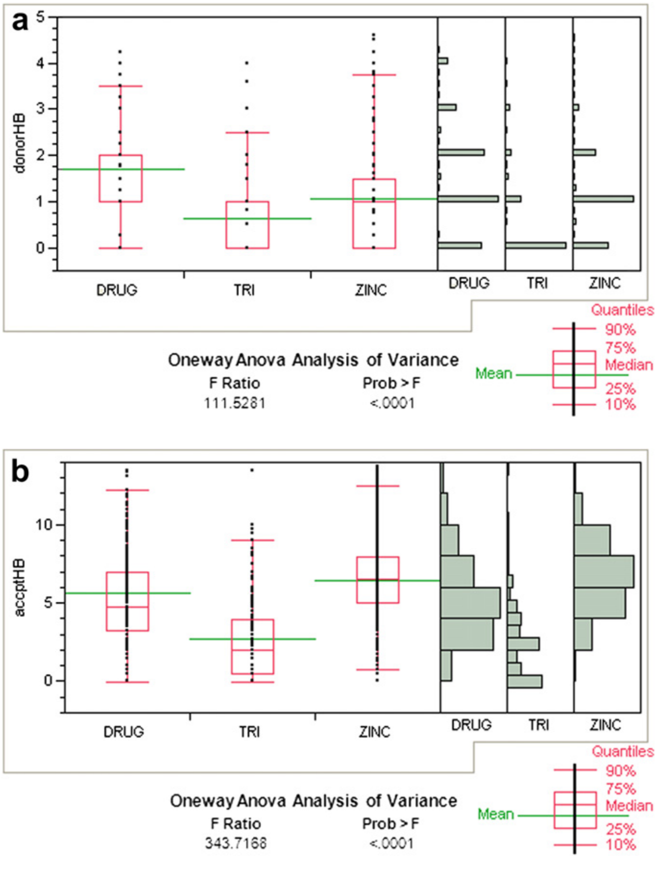

Figure 7.

methyl iodide, which is known to be a powerful methylating agent of biological molecules, has a dipole moment of 2.45 D according to QikProp.

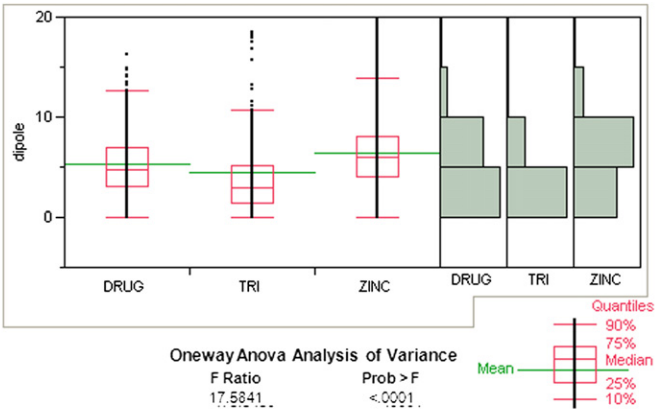

Figure 8.

The calculated polarizability is considered to represent the extent to which an electric field can influence the charge distribution in a molecule. The principal influence of polarizability on bioavailability would be most likely derived from water solubility of crystalline compounds. The difference in distributions of means is significant between the TRI set and the ZINC database, 18.3 and 34.3, respectively. This indicates that this group of toxic compounds do have significantly lower polarizability than the average in the ZINC database. This is linked to the partition between lipid and water solubility.[31] An analysis of our predicted property data from the ZINC database as well as the TRI group indicates an inverse relationship between predicted polarizability and log of water solubility (Fig. 9a) and a positive correlation with solubility in the non-polar hexadcane (Fig. 9b).

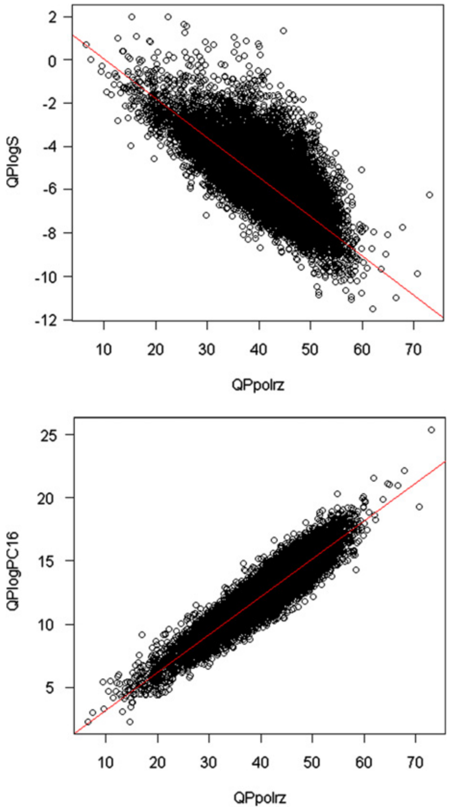

Figure 9.

3.1.5. Solubility. Experimental or calculated aqueous and lipid solubilities are essential to understanding a compound's bioavailability considering that aqueous transport and lipid bilayer permeability control bioavailability and distribution of a chemical in the human body. As such, Lipinski's rules indicate that log $P_{oct/water}$ is vital in predicting bioavailability.[6] In addition, in studies aimed at understanding ecotoxicity, log $P_{oct/water}$ has been related directly to bioaccumulation in fish.[32] The aqueous solubility and the solvent/gas partition coefficients, considering both water and octane as solvents ($P_{water/gas}$ and $P_{oct/gas}$, respectively), were evaluated in addition to the log $P_{oct/water}$. These parameters are important in understanding transport and permeability of vapors (such as gases dissolving on alveolar membranes in lungs) (Fig. 10).

The data for $P_{oct/gas}$ indicates decreased lipid solubility among the TRI compounds compared to both drug compounds and the ZINC database (log $P_{o/g}$ mean for TRI=10.7; drug=16.8; ZINC=19.2). The same trend is observed for the water/gas partition (log $P_{w/g}$ mean for TRI=5.08; drug=10.3; ZINC=11.0). This trend agrees with the low polarity and smaller molecular size of the TRI compounds compared to the drug and ZINC datasets.

3.2. Predicted biological interactions/properties

The permeability into cells is typically predicted using several model cell types as described in Section 3. As described previously, the MDCK epithelial cell lines are commonly used as a possible tool for assessing the membrane permeability properties of blood–brain

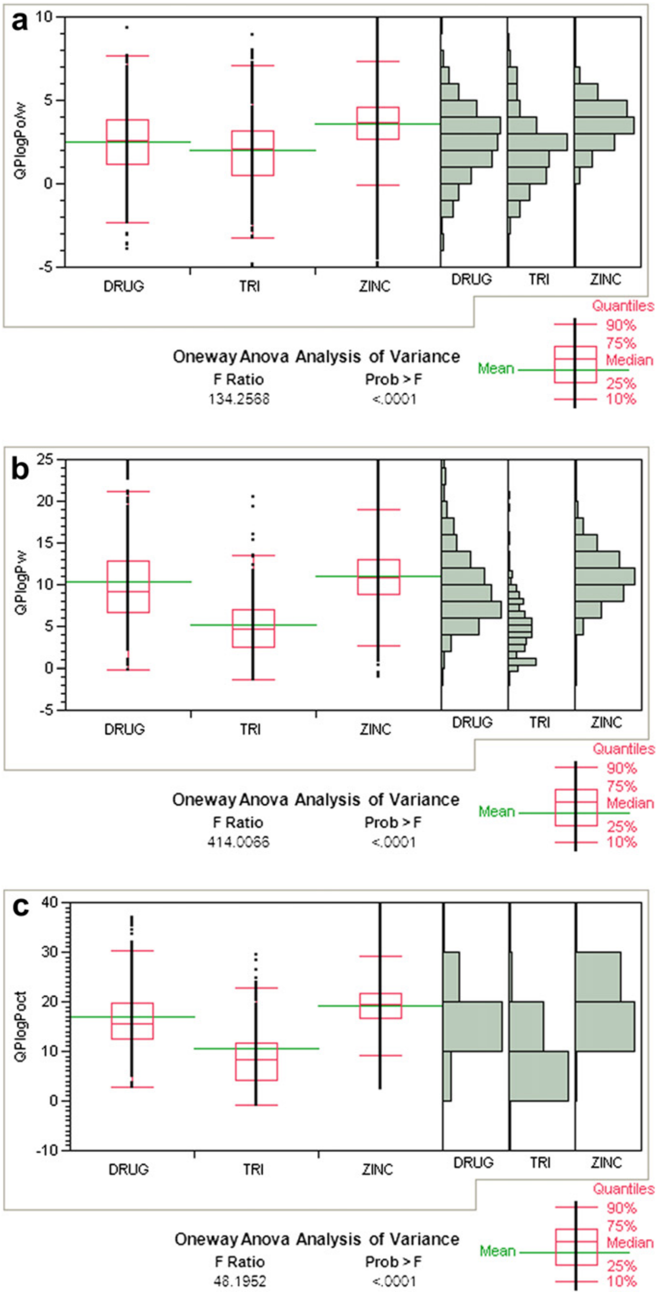

Figure 10.

barrier for early drug discovery. Caco-2 cells, on the other hand, are immortalized lines of heterogeneous human epithelial colorectal adenocarcinoma cells[33] that are widely used with in vitro assays to predict the absorption rate of candidate drug compounds across the intestinal epithelial cell barrier. Notably, for both cell lines considered, the TRI compounds showed significantly increased permeability rates compared to both the drug and ZINC datasets (Fig. 11). This again points to greater bioavailability for the TRI compounds.

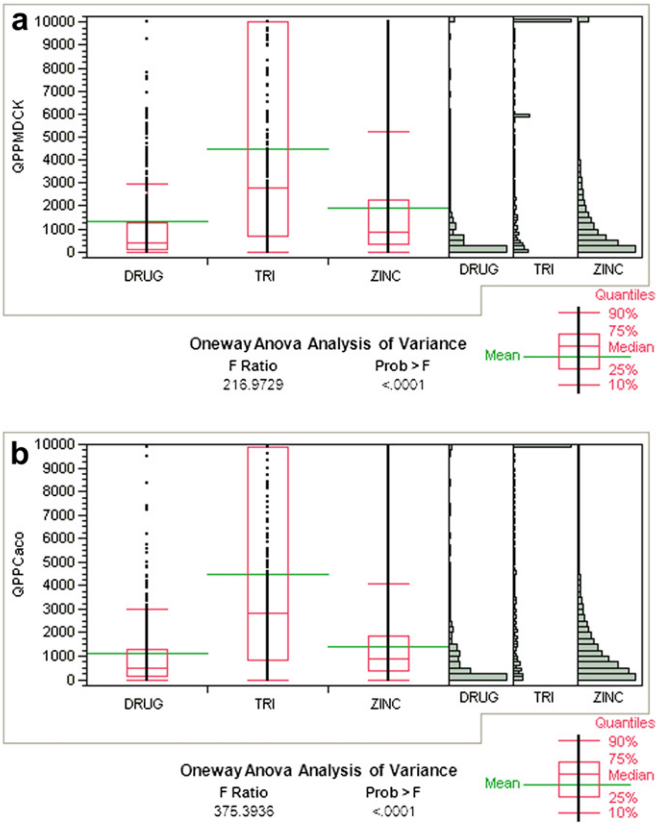

Figure 11.

Finally, human skin permeability, log K_p, was calculated for the three sets of compounds. In this case none of the datasets had means that fell in the range of good skin absorption (−8 to −10), indicating that only 0.5% of the TRI chemicals are significantly bioavailable through dermal absorption. It should be noted, however, that this does not take into account any local effects, such as skin sensitization or corrosive strength of these compounds.

3.3. Partitioning analysis to interpret most relevant physiochemical properties

Partitioning was performed using all the physiochemical and biological properties in Tables 1 and 2, yielding results that identify the groupings of physiochemical values which best distinguish the TRI and ZINC groups. Figure 12 illustrates the tree

Table 2
Biological properties of the 568 TRI compounds compared with 546 commercial drug compounds and 133,667 randomly selected compounds from the ZINC Database

Property category	Property	Mean (2.5 and 97.5% quantiles)			Range or recommended values for drugs[a]
		TRI (toxic) compounds	Drug compounds	ZINC Database compounds	
Organ permeability	log (human skin absorption)	−2.22 (−5.80, 0.13)	−3.39 (−7.58, −0.65)	−2.18 (−4.76, −0.28)	−8 to 10
	log (blood–brain)	−0.16 (−1.67, −0.794)	−0.58 (−3.17, −0.83)	−0.79 (−2.32, 0.39)	−3.0 to 1.2
Cell permeability/ channel blockers	log (HERG IC_{50})	−3.40 (−6.05, −0.866)	−4.69 (−7.88, −1.06)	−5.65 (−7.63, −3.14)	>−5
	Caco (nm/s)	4453.45 (18.02, 9906.45)	1122.30 (1.20, 7354.83)	1381.24 (47.98, 5567.98)	<25 poor, >500 great
	MDCK (nm/s)	4503.34 (25.65, 10,000.00)	1299.70 (1.09, 100,001.00)	1884.34 (31.34, 10,000.89)	<25 poor, >500 great

[a] For 95% of known drugs.[21]

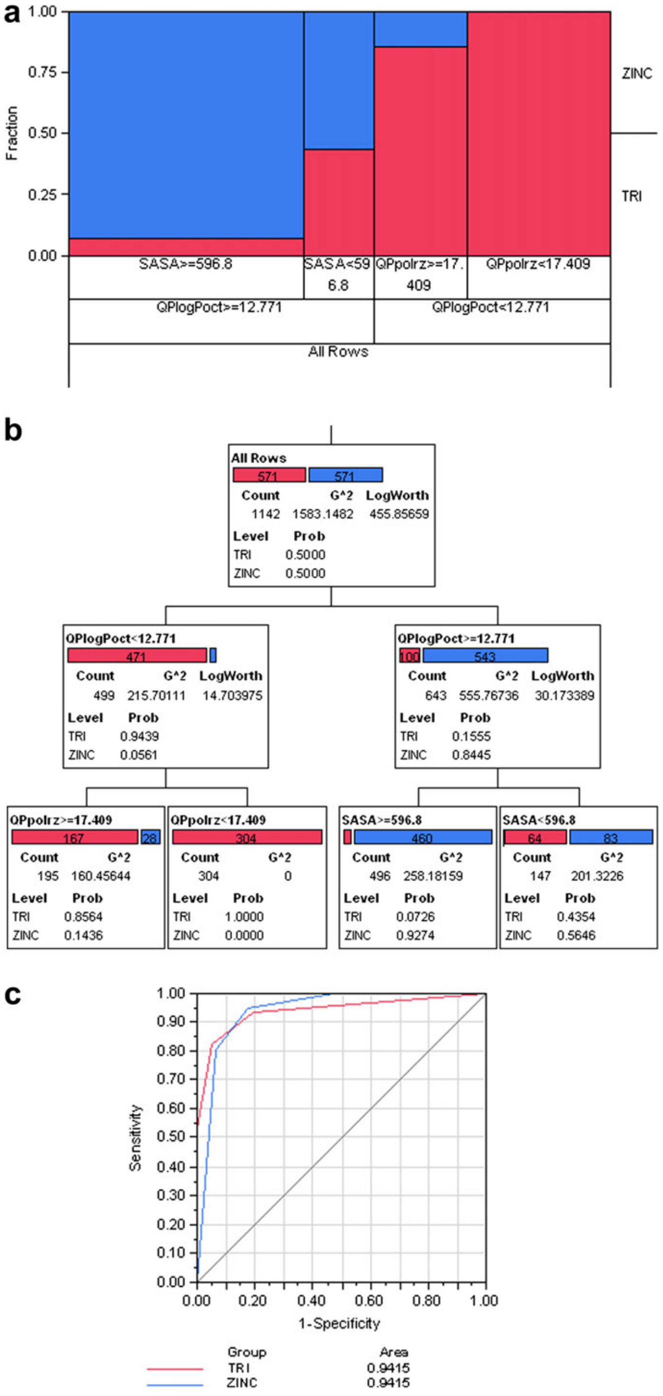

Figure 12.

diagram and partition graph representing the best splitting where TRI is red and ZINC is blue. The first split occurred by using a value of 12.77 for the log $P_{oct/gas}$ predicted parameter, resulting in distinguishing 471 TRI chemicals from the ZINC sample. The next split occurred by using a value of 17.409 for the predicted polarizability, which distinguished 304 TRI compounds and no members of the ZINC set. The real significance of this analysis comes from the fact that if the group of predominantly ZINC compounds is examined (with log P_{oct}>12.77 and SASA>596.8 Å^2) a set of properties that exclude 94% of the toxic TRI compounds has been identified.

This partitioning was repeated with three random samples from the ZINC database, and showed little variation. The goodness of fit was represented by a ROC curve, shown in Figure 13. This analysis implies that the toxic chemicals can be described by as little as three physiochemical properties and specific ranges. In each of the partition analyses the octanol/gas partition coefficient was always selected as the most significant partitioning property, with the dividing value falling between 11.7 and 13.8. Size was the second property, which was usually represented by SASA, but in one case by molecular volume. Finally polarizability was the property used to split the remaining ZINC chemicals from the predominantly TRI grouping.

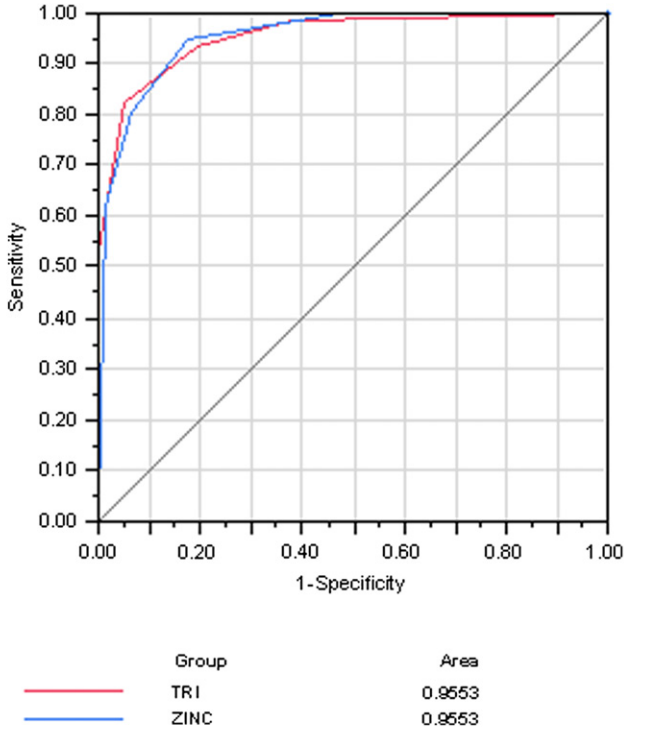

Figure 13.

4. Conclusions

The development of design protocols based on physical/chemical properties will require an understanding of the fundamental relationships between biological activity, specifically toxicity, and these properties. By comparing the dataset of toxic chemicals (Toxic Release Inventory database), therapeutic dataset (pharmaceutical database), and the random dataset (ZINC database), distinctions between these classes of chemicals have been identified. The compounds in the Toxic Release Inventory list are shown to share some relatively narrow property boundaries, all of which are directly or indirectly linked to bioavailability. With this analysis as a basis, the crucial characteristics that make up ‘toxic chemical space’ and ‘safe chemical space’ can be recognized and understood. In addition, the analysis has elucidated which of these properties has the highest degree of distinction between toxic chemicals and the random or therapeutic chemicals set. With this insight, the basis for a set of design rules or guidelines to be considered in the design of safer chemicals begins to emerge.

Future work will include a focus on chemical datasets with full toxicity characterization documenting the lack of any adverse biological activity to provide a verification of these initial analyses. In addition, chemical classes that are known to undergo metabolic transformations and degradations will be explored to understand the impact on the analysis and how these properties affect the fundamentals of potential design rules.

Acknowledgements

For financial support we would like to thank the Johnson Family Foundation. We would also like to thank Dr. Julian Tirado-Rives and Prof. W. Jorgensen for insightful discussions and consultations.

References and notes

1. Fabre, N.; Anglade, I.; Vericat, J. A. *Toxicol. Lett.* **2009**, *186*, 13–17.
2. Horvath, I. T. .; Anastas, P. T. *Chem. Rev.* **2007**, *107*, 2169–2173.
3. Boelsterli, U. A. *Mechanistic Toxicology: The Molecular Basis of How Chemicals Disrupt Biological Targets*; Taylor & Francis: New York, NY, 2003.
4. Horvath, I. T. *Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds*; Elsevier Science: 1992.
5. DeVito, S.; Garrett, R. *Designing Safer Chemicals: Green Chemistry for Pollution Prevention*; American Chemical Society: Washington, DC, 1996; Vol. 640.
6. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
7. Golovenko, N. Y.; Borisyuk, I. Y. *Bioorg. Khim.* **2008**, *54*, 392–407.
8. Hurst, S.; Loi, C. M.; Brodfuehrer, J.; El-Kattan, A. *Expert. Opin. Drug. Metab. Toxicol.* **2007**, *3*, 469–489.
9. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615–2623.
10. Proudfoot, J. R. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1087–1090.
11. Schneider, T.; Vermeulen, R.; Brouwer, D. H.; Cherrie, J. W.; Kromhout, H.; Fogh, C. L. *Occup. Environ. Med.* **1999**, *56*, 765–773.
12. Hansch, C. In *Drug Design*; Ariens, E. J., Ed.; Academic: New York, NY, 1971.
13. Cronin, M. T. D.; Livingstone, D. *Predicting Chemical Toxicity and Fate*; CRC: Boca Raton, FL, 2004.
14. Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. *Chem. Rev.* **2002**, *102*, 783–812.
15. LeBlanc, G. A.; Kaiser, K. L. E. QSAR in environmental toxicology. Proceedings of the workshop on quantitative structure–activity relationships in environmental toxicology held at McMaster University, Hamilton, Ontario, Canada, August 16–18, 1983. **1984**, 235–260.
16. Hu, X. L.; Liu, J. F.; Lu, S. Y.; Jiang, G. B. *Prog. Chem.* **2009**, *21*, 514–523.
17. Jorgensen, W. *QikProp Version 2.3*; Schrodinger, LLC: New York, NY, 2003.
18. Babel, O. S. *File Format Converter v.2.2.2*; 2009; http://openbabel.org/.
19. U.S. E.P.A. *Toxic Release Inventory Program*; U.S. Environmental Protection Agency: Washington D.C., 2009. Last updated June 8th; http://www.epa.gov/tri.
20. Polli, J. W.; Wring, S. A.; Humphreys, J. E.; Huang, L. Y.; Morgan, J. B.; Webster, L. O.; Serabjit-Singh, C. S. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 620–628.
21. Jorgensen, W. *QikProp v. 2.3: User Manual.* Schrodinger, LLC: New York, NY, 2009; p 20014.
22. Anson, B. D.; Weaver, J. G. R.; Ackerman, M. J.; Akinsete, O.; Henry, K.; January, C. T.; Badley, A. D. *Lancet* **2005**, *365*, 682–686.
23. Jackson, M. J. In *Physiology of the Gastrointestinal Tract*; Johnson, L. R., Ed.; Raven: New York, NY, 1987; pp 1597–1621.
24. Zmuidinavicius, D.; Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. *J. Pharm. Sci.* **2003**, *92*, 621–633.
25. Lu, J. J.; Crimin, K.; Goodwin, J. T.; Crivori, P.; Orrenius, C.; Xing, L.; Tandler, P. J.; Vidmar, T. J.; Amore, B. M.; Wilson, A. G.; Stouten, P. F.; Burton, P. S. *J. Med. Chem.* **2004**, *47*, 6104–6107.
26. Dressman, J. B.; Thelen, K.; Jantratid, E. *Clin. Pharmacokinet.* **2008**, *47*, 655–667.
27. Refsgaard, H. H. F.; Jensen, B. F.; Brockhoff, P. B.; Padkjaer, S. B.; Guldbrandt, M.; Christensen, M. S. *J. Med. Chem.* **2005**, *48*, 805–811.
28. Faassen, F.; Kelder, J.; Lenders, J.; Onderwater, R.; Vromans, H. *Pharm. Res.* **2003**, *20*, 177–186.
29. Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. *Pharm. Res.* **1997**, *14*, 568–571.
30. Vippagunta, S. R.; Brittain, H. G.; Grant, D. J. W. *Adv. Drug Delivery Rev.* **2001**, *48*, 3–26.
31. Chenzhong, C.; Zhiliang, L. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1–7.
32. Chiou, C. T.; Freed, V. H.; Schmedding, D. W.; Kohnert, R. L. *Environ. Sci. Technol.* **1977**, *11*, 475–478.
33. Fogh, P.; Ellervik, C.; Saermark, T.; Brynskov, J. *Ann. N.Y. Acad. Sci.* **1999**, *878*, 692–695.