Journal of Computer-Aided Molecular Design, **14:** 251–264, 2000.
KLUWER/ESCOM

Property distribution of drug-related chemical databases*

Tudor I. Oprea
Medicinal Chemistry, Astra Hässle AB, S-431 83 Mölndal, Sweden (E-mail: tudor.oprea@astrazeneca.com)

Received 9 March 1999; Accepted 15 September 1999

Key words: combinatorial chemistry, computer chemistry, database filtering, drug-likeness, drug research, hydrogen bonds, property distribution, 'rule of 5' test

Summary

The process of compound selection and prioritization is crucial for both combinatorial chemistry (CBC) and high throughput screening (HTS). Compound libraries have to be screened for unwanted chemical structures, as well as for unwanted chemical properties. Property extrema can be eliminated by using property filters, in accordance with their actual distribution. Property distribution was examined in the following compound databases: MACCS-II Drug Data Report (MDDR), Current Patents Fast-alert, Comprehensive Medicinal Chemistry, Physician Desk Reference, New Chemical Entities, and the Available Chemical Directory (ACD). The ACDF and MDDRF subsets were created by removing reactive functionalities from the ACD and MDDR databases, respectively. The ACDF subset was further filtered by keeping only molecules with a 'drug-like' score [Ajay et al., J. Med. Chem., 41 (1998) 3314; Sadowski and Kubinyi, J. Med. Chem., 41 (1998) 3325] below 0.8. The following properties were examined: molecular weight (MW), the calculated octanol/water partition coefficient (CLOGP), the number of rotatable (RTB) and rigid bonds (RGB), the number of rings (RNG), and the number of hydrogen bond donors (HDO) and acceptors (HAC). Of these, MW and CLOGP follow a Gaussian distribution, whereas all other descriptors have an asymmetric (truncated Gaussian) distribution. Four out of five compounds in ACDF and MDDRF pass the 'rule of 5' test, a probability scheme that estimates oral absorption proposed by Lipinski et al. [Adv. Drug Deliv. Rev., 23 (1997) 3]. Because property distributions of HDO, HAC, MW and CLOGP (used in the 'rule of 5' test) do not differ significantly between these datasets, the 'rule of 5' does not distinguish 'drugs' from 'nondrugs'. Therefore, Pareto analyses were performed to examine skewed distributions in all compound collections. Seventy percent of the 'drug-like' compounds were found between the following limits: $0 \leq \text{HDO} \leq 2$, $2 \leq \text{HAC} \leq 9$, $2 \leq \text{RTB} \leq 8$, and $1 \leq \text{RNG} \leq 4$, respectively. The number of launched drugs in MDDR having $0 \leq \text{HDO} \leq 2$ is 4.8 times higher than the number of drugs having $3 \leq \text{HDO} \leq 5$. Skewed distributions can be exploited to focus on the 'drug-like space': 62.68% of ACDF ('nondrug-like') compounds have $0 \leq \text{RNG} \leq 2$, and $\text{RGB} \leq 17$, while 28.88% of ACDF compounds have $3 \leq \text{RNG} \leq 13$, and $18 \leq \text{RGB} \leq 56$. By contrast, 61.22% of MDDRF compounds have $\text{RNG} \geq 3$, and $\text{RGB} \geq 18$, and only 24.73% of MDDRF compounds have $0 \leq \text{RNG} \leq 2$ rings, and $\text{RGB} \leq 17$. The probability of identifying 'drug-like' structures increases with molecular complexity.

Introduction

Combinatorial chemistry (CBC) and high throughput screening (HTS) are two rapidly evolving technologies that aim at synthesis (CBC) and evaluation (HTS) of libraries of compounds. A compound library is a collection (10^2–10^6) of virtual and/or existing compounds, that can be either synthesized in-house or purchased from external sources, if not already available. Once compounds are available, they are typically stored in a 96-well format (or higher), then the library is screened in the HTS program against the desired target(s). As the range of synthetic possibilities, as well as the number of commercially available compounds are increasing every day, the process of compound selection and/or prioritization has become crucial [1].

*Dedicated to Prof. Dr. Garland R. Marshall on the occasion of his 60th birthday.

Selection involves the evaluation of molecular diversity and a number of tools are available for this process [2].

Two additional steps are performed at the library level, prior to compound selection via diversity analysis and/or experimental design: (1) with respect to the chemical structures, it is recommended to eliminate those classes of compounds [3] that would generate false positives in the HTS assays; and (2) with respect to the chemical properties, it is preferable to exclude those with undesired properties [4]. Any library of virtual or existing compounds contains structures that may be deemed unsuitable for screening against a particular target. These can be rapidly eliminated by applying property filters, in a manner that is convenient to the end-user.

To assist chemists who need to handle large compound libraries, we have examined a series of commercially available compound databases, as follows: MDDR [5] (MACCS-II Drug Data Report), FALERT [6] (Current Patents *fast-alert*), CMC [7] (Comprehensive Medicinal Chemistry), PDR [8] (Physician Desk Reference), NCE [9] (new chemical entities) and ACD [10] (the Available Chemical Directory). Each of these databases was chosen for a different reason: MDDR contains structures that were synthesized, screened in vitro and are intended for medical use (but have yet to pass clinical trials); FALERT represents a non-redundant set of MDDR, in the sense that each patent is represented by one (or a few) key structure(s), whereas MDDR includes more structures per patent; CMC is a subset of the structures included in the *Comprehensive Medicinal Chemistry* [11] textbook, chosen for their availability of *measured* Log P values; PDR is a set of orally available drugs (currently available on the market); NCE is a set of structures that have passed phase I and II of clinical trials (some of these are drugs, but some are not); and ACD, which is a set of buyable starting materials that are used worldwide for drug synthesis. Some compounds included in the ACD database are drugs themselves. Therefore, we have removed 'drug-like' structures from ACD, and filtered reactive species in both MDDR and ACD (as described in the Experimental section), to produce the ACDF and MDDRF subsets. For the sake of clarity, in this study we refer to MDDR (MDDRF) and FALERT compounds as 'pharmaceutically active', to PDR and NCE compounds as 'drugs', and to ACD and ACDF compounds as 'nondrugs'.

To maintain the focus of drug discovery towards effective and orally absorbable compounds, there has been an increasing interest in absorption, distribution, metabolism and excretion (ADME). Therefore, we studied the same properties that were recently examined by Lipinski et al. in their attempt to address ADME issues [12]. Lipinski and co-workers suggested, after studying a subset of 2245 drugs from the World Drug Index, that poor absorption or permeation are more likely when there are more than 5 H-bond (hydrogen bond) donors (HDO – expressed as the sum of O-Hs and N-Hs), when the molecular weight (MW) is over 500, when the calculated octanol/water partition coefficient (CLOGP) is over 5, and when there are more than 10 H-bond acceptors (HAC – expressed as the sum of Ns and Os). The 'rule of 5' probability scheme is currently used to assist chemists to avoid compounds with potential permeability problems. Any pairwise combination of the following conditions: MW $>$ 500, CLOGP $>$ 5, HDO $>$ 5, and HAC $>$ 10, may result in compounds with poor permeability.

We chose the same parameters as Lipinski and coworkers: MW, CLOGP (using Leo's method available from Daylight) [13], HDO and HAC. In addition, we included the number of rings (RNG), the number of non-terminal rotatable bonds (RTB), and the number of rigid bonds (RGB). These three additional properties were chosen to reflect the molecular complexity of a library, as they relate to the flexibility (RTB) and the rigidity (RGB, RNG) of a molecule.

Two schemes that distinguish between 'drug-like' and 'nondrug-like' compounds were recently published [14]. These schemes are capable of distinguishing compounds associated to a 'nondrug' database (e.g., ACD), as opposed to a 'drug' database (e.g., MDDR or CMC). Although quite successful, these approaches that separate 'drugs' from 'nondrugs' cannot be interrogated to determine the underlying basis of their classification in a chemical sense. We describe how, based on differences in property distribution between MDDRF and ACDF, it is possible to draw a boundary between two volumes of the chemical space, one populated mostly by 'nondrugs', the other by 'drugs'.

Experimental section

Database contents

All databases were filtered to remove salts, metals and isotopes. The FALERT database was filtered to

include only compounds with 'structure contains a carbon atom' – this limited the number of structures to 16226, out of over 25000 entries. Of these, CLOGP was calculated for 11954 structures. The MDDR subset contained 82903 structures, of which 69581 were calculated with the CLOGP program. For comparison, we examined a subset of 984 structures from CMC for which measured Log P values were available. Only MW and CLOGP distributions are reported for CMC. PDR, a compilation of orally available drugs, included 413 structures examined, of which 373 were calculated with CLOGP. NCE contains 420 new chemical entities, of which 400 were estimated using CLOGP.

The ACD database was filtered to contain only C, H, N, O, S, P and X as elements, similar to the drug-like databases. ACD was also filtered for MW, which was limited to values between 86 and 600. The resulting ACD subset contained 194511 structures, of which 149305 were calculated with the CLOGP program.

MDDR and ACD were filtered to remove compounds with reactive functionalities. Acyl-halides, sulfonyl-halides, Michael acceptors, etc. were removed from these databases using the SMARTS language in the Daylight Toolkit [15]. A complete list of the removed classes of compounds is given in Figure 1. These filtered databases are referred to as MDDRF and ACDF, respectively. Furthermore, 'drug-like' scores [16] were calculated for all compounds in the ACDF subset, and only those with scores below 0.8 were included in ACDF. Reactive species removal resulted in 155926 compounds in ACDF (80.12% of the ACD subset). Another 5616 structures (2.88%), deemed as 'drug-like' in this set, were further removed to yield a total of 150310 compounds (77.24%) in ACDF. In MDDR, 21.94% of the structures were found to contain reactive substructures [17]. No significant differences were found when comparing distribution differences in ACDF vs. ACD, and MDDR vs. MDDRF, respectively, unless otherwise mentioned.

Descriptors

All property estimates were performed using a subset of the Selma module of SaSA [18]. Selma uses the Daylight Toolkit [15] to evaluate CLOGP, RNG, RTB, RGB, HDO and HAC. *The number of rings* (RNG) is evaluated using the SSSR (smallest set of smallest rings) algorithm [19], as implemented in the Daylight Toolkit [15].

The number of rotatable bonds (RTB) is formulated in Equation 1:

$$\mathrm{RTB} = \mathrm{N}_{nt} + \sum_i (\mathrm{n}_i - 4 - \mathrm{RGB}_i - \mathrm{ShB}_i) \qquad (1)$$

where N_{nt} is the number of non-terminal freely rotatable bonds (but single bonds observed in groups like, e.g., sulfonamides (N–S) or esters (C–O), are excluded); n_i is the number of single bonds in any non-aromatic ring *i* with 6 or more bonds; RGB_i is the number of rigid bonds in ring *i*; ShB_i is the number of bonds shared by ring *i* with any other ring. The number of rigid bonds (RGB) is defined as the difference between the total number of bonds and the total number of rotatable bonds (including terminal single bonds).

The numbers of hydrogen-bond donors and acceptors (HDO, HAC) are based on a look-up table of known fragments that are involved in hydrogen bonding, that includes only nitrogen and oxygen. Other donors, such as thiols, or acceptors, such as halides, are ignored. HDO counts all N–H and O–H fragments. Exceptions are all acids (which are considered deprotonated). Amide and amide-like (e.g., urea, sulfonamide) nitrogens, as well as tertiary amines, are not considered as H-bond acceptors. Since no $\mathrm{pK_a}$ estimator is included in this scheme, protonation states are not considered (e.g., amines are not protonated). However, HDO and HAC are counted separately, meaning that an O–H group can be both a donor and an acceptor.

Property distributions were analyzed using the *Histogram* program, written by Stefan Ervik and Tudor Oprea. This routine performs a histogram-type of analysis on large numbers of data, in a manner that can be user-defined (with respect to, e.g., the bin size and the cut-off). *Histogram* counts the number of data points between the current bin number and the adjoining higher bin, if any. A value is counted in a particular bin if it is equal to, or less than the bin value down to the last bin. For integer variables (e.g., RTB, HDO), results are displayed for the corresponding values (bin counts for that particular bin range). Values below the first bin value are reported in the first bin, while values above the last bin value are the difference from the highest percentage reported and 100%. All missing values (e.g., for CLOGP) are reported in the final percentage count, unless otherwise stated. Missing values are occasionally reported in the 'skipped' column. Additionally, we define a *peak* as the point where the distribution reaches *exactly* 50%, whereas the *mid-50%* values are the regions between the 25% and 75%

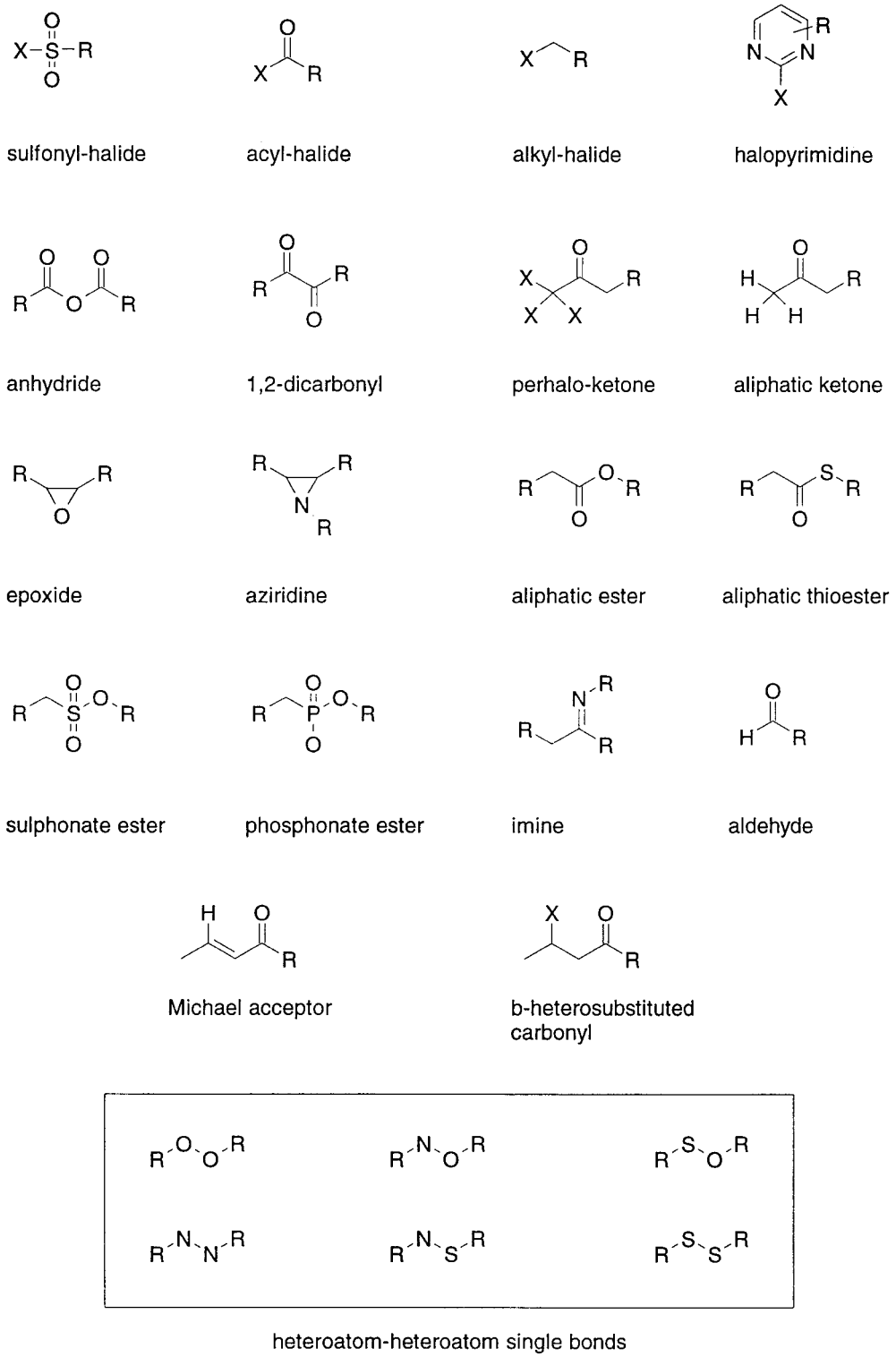

Figure 1. Reactive functional groups removed from the ACDF and MDDRF subsets. Modified from Reference 3.

of the distribution range, precisely. We report these distribution moments instead of average values, since averages give different information than the median, and neither can handle missing values appropriately.

The 80/20 principle

The 80/20 (or Pareto) principle, discovered by Vilfredo Pareto in 1897, can be formulated as follows: A minority of causes usually leads to a majority of results [20]. This is valid in business (20% of the products account for 80% of the sales value), in software (80% of computer time is spent executing 20% of the operating code), on the road (20% of the motorists cause 80% of the accidents), and in society (20% of those who marry comprise 80% of the divorce statistics) [20]. The essence of the 80/20 principle is the pattern of imbalance, that can take any range of values, from 60/40 to 99/1 or more. The 80/20 analysis splits objects into two categories: high-, and low-impact. It highlights the 50% unimportant part (50% of the input has sometimes 5% or less contribution) that – when ignored – can decrease the complexity of the model and improve our understanding of the data. The principle of 'the vital few and the trivial many' [20] is clearly illustrated in HTS results: only a few (good) hits emerge from a library screen, whereas a vast majority of the screened compounds did not achieve the intended activity[1]. We therefore performed Pareto analyses of the property distributions presented above, focusing on those properties that have asymmetrical distribution (e.g., RNG, RGB, RTB), and have used the 80/20 principle to distinguish between the two sets of compounds.

Results

Molecular weight

The molecular weight distribution in the MDDR, ACD and CMC databases is shown in Figure 2. All databases follow a similar, Gaussian, distribution: PDR follows ACD, and both are similar to CMC, whereas MDDR and FALERT have lower percentage values; NCE is in-between PDR and MDDR. In fact, pharmaceutically active compounds (MDDR, FALERT) display an increased molecular weight, with peak around 400, and mid-50% between 320–500. Drugs show a

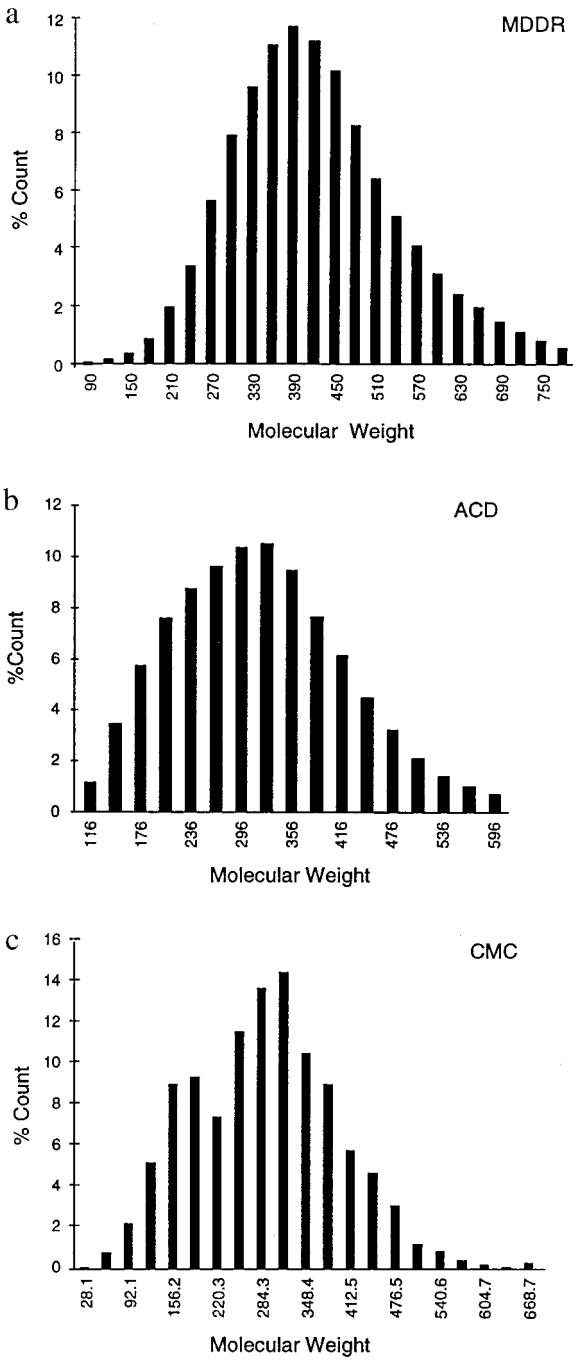

Figure 2. Molecular weight (MW) distribution in MDDR (82903 compounds), ACD (194511 compounds) and CMC (984 compounds with measured LogP).

[1]This situation may not apply to genetic-algorithm and/or other target-directed libraries.

small tendency towards increased MW, the peak going from 307 to 355, and the mid-50% moving from 238–376 to 285–427, when one compares PDR to NCE. ACD peaks around 300, and mid-50% between 224–368 – with 80% of the compounds below 385. CMC peaks around 270, with mid-50% between 198 and 348. Eighty percent of CMC structures have MW below 372. In summary, FALERT compounds have increased MW values, quite similar to MDDR, whereas ACD and CMC have lower MW structures. PDR and NCE are intermediate in terms of MW distribution.

Partition coefficient

The distribution of the calculated partition coefficient (CLOGP) in the MDDR and ACD databases is shown in Figure 3. For comparison, we also show the distribution of the experimental LogP values in the CMC database in Figure 3. We chose not to compare the calculated and experimental LogP values for the CMC database, since this type of comparison has been published elsewhere [21]. All databases follow a Gaussian distribution: CMC, PDR and NCE appear to have high-percentage values, and appear to be different from the low-percentage group, ACD, MDDR and FALERT. These results are influenced by the number of compounds with missing values: 16.1% in MDDR (13.41% in MDDRF), 26.5% in FALERT, 8.1% in NCE, 9.7% in PDR and 23.2% in ACD (15.91% in ACDF). These compounds were not taken into account in Figure 3, and were also excluded in Table 1, where the percentage of compounds included in the −1 to 5, and the 0 to 3 intervals are given. NCE, CMC and PDR (high percentage) are still distinct from FALERT, MDDR, MDDRF, ACD and ACDF (low percentage) in the 0 to 3 interval. Pharmaceutically active compounds (MDDR, FALERT) peak around 3.1, with mid-50% range between 1.3 and 4.75. Drugs do not show changes regarding CLOGP distribution: The peak remains around 2.35, whereas the mid-50% remains between 0.8–3.7 (both PDR and NCE). ACD peaks around 3.0, with the mid-50% between 1.64 and 4.45. CMC peaks at 1.78, with mid-50% between 0.34 and 2.85 (90% below 4.0). Note that compounds with 'missing' CLOGP values were excluded from these quartile ranges. To summarize, NCE, CMC and PDR have more compounds within the 0 to 3 LogP range, by comparison to FALERT, MDDR (MDDRF) and ACD (ACDF).

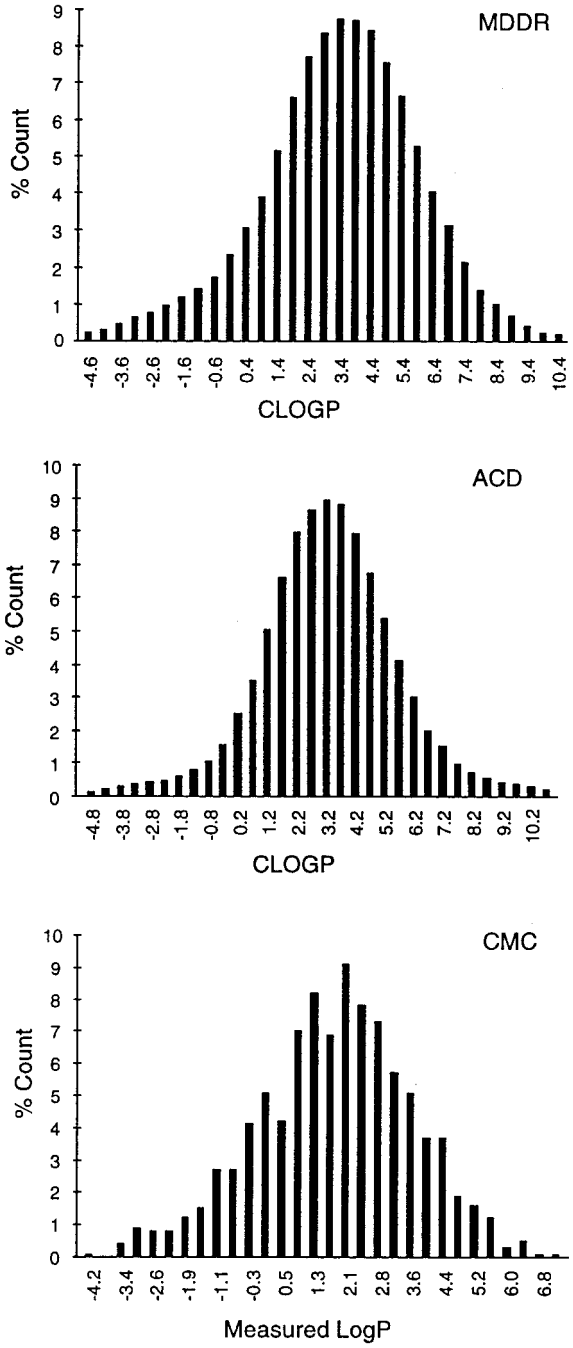

Figure 3. Distribution of the calculated LogP (CLOGP) in MDDR (69580 compounds) and ACD (149308 compounds) and distribution of the measured LogP in CMC (984 compounds).

Table 1. Some intervals for the partition coefficient across different databases (percentage values apply only to compounds with available values)

Database	−1 to 5 interval	0 to 3 interval	Compounds[a]
MDDR	75.5%	41.6%	69581
MDDRF	73.64%	38.93%	59333
FALERT	72.1%	42.9%	11954
NCE	75.75%	71%	400
PDR	82.3%	52.4%	372
ACD	78.1%	40.82%	149305
ACDF	79.26%	44.33%	126127
CMC[b]	87.3%	57.5%	984

[a]Compounds with available partition coefficient values only.
[b]Measured Log P values only.

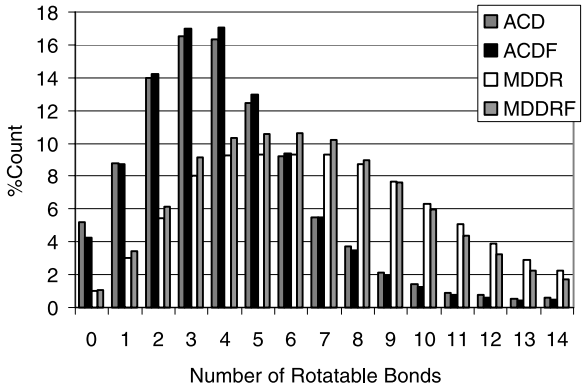

Figure 4. Distribution of the rotatable bonds (RTB) in MDDR (82903 compounds), MDDRF (68523 compounds), ACD (194511 compounds) and ACDF (150310 compounds).

Number of rotatable bonds

The distribution of the number of rotatable bonds in MDDR, MDDRF, ACD and ACDF is shown in Figure 4. All the examined databases follow an asymmetrical Gaussian distribution, that is truncated towards low values. This distribution is observed in other properties (RNG, HDO, HAC), although it is more skewed in some (e.g., HDO). For the RTB distribution, ACD is followed closely by PDR, while FALERT and MDDR have lower percentage values. NCE is between PDR and MDDR. Pharmaceutically active compounds (MDDR, FALERT) peak around 6–7 RTB, with mid-50% between 4 and 11. Drugs tend to increase RTB, the mid-50% going from 2–7 (peak at 4) in PDR, to 3–9 (peak at 6) in NCE. ACD peaks at 4, with mid-50% between 2 and 6. Over 60% of the ACD/ACDF compounds have RTB ≤ 4, whereas over 70% of MDDR/MDRF have RTB ≥ 5. No significant

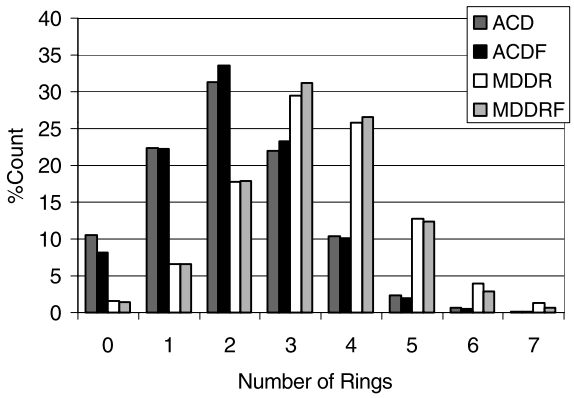

Figure 5. Ring (RNG) distribution in MDDR (82903 compounds), MDDRF (68523 compounds), ACD (194511 compounds) and ACDF (150310 compounds).

differences in RTB distribution can be observed by comparing ACD vs. ACDF, and MDDR vs. MDDRF, respectively (Figure 4).

Number of rings

The distribution of rings in MDDR, MDDRF, ACD and ACDF is shown in Figure 5. It is a truncated Gaussian distribution that shows how ACD (ACDF) compounds have, on average, 1 ring less when compared to biologically related compounds. These compounds (MDDR, FALERT, PDR and NCE) peak at 3 rings, with mid-50% between 2 and 4 rings. Over 40% of MDDR (MDDRF) and FALERT have 4 rings or more. ACD and ACDF peak at 2 rings, with mid-50% between 1 and 3 rings. Sixty-four percent of the ACD and ACDF compounds have RNG ≤ 2, whereas 74% of the MDDR and MDDRF subsets have RNG ≥ 3. PDR and NCE have RNG distributions similar to those of FALERT and MDDR.

Number of rigid bonds

The distribution of the number of rigid bonds in ACD, ACDF, MDDR and MDDRF is shown in Figure 6. Pharmaceutically active compounds (MDDR, FALERT) peak at 20 rigid bonds, with mid-50% values between 14 and 26. Drugs tend to increase RGB, the mid-50% going from 10–20 (peak at 15) in PDR, to 12–22 (peak at 17) in NCE. ACD and ACDF peak at 14 rigid bonds, with mid-50% values between 8 and 18. Over 70% of ACD (ACDF) compounds have RGB ≤ 17, whereas over 60% of MDDR (MDRF) have RTB ≥ 18. The distribution of RGB parallels that of RNG, but shows that PDR is closer to ACD, whereas

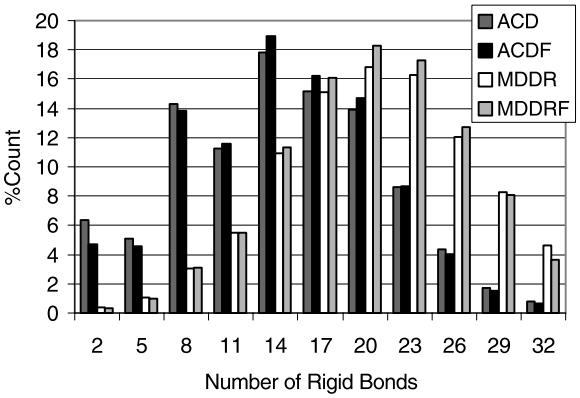

Figure 6. Distribution of rigid bonds (RGB) in MDDR (82903 compounds), MDDRF (68523 compounds), ACD (194511 compounds) and ACDF (150310 compounds).

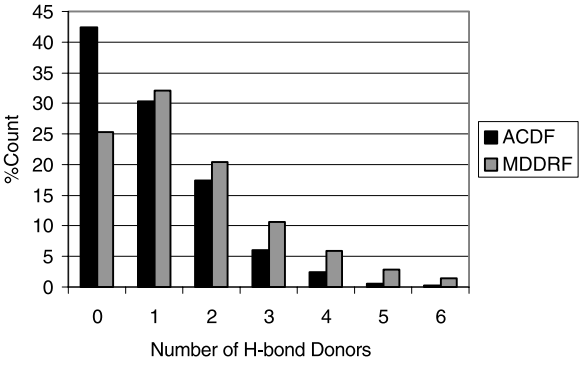

Figure 7. Hydrogen bond donor (HDO) distribution in MDDRF (68523 compounds) and ACDF (150310 compounds).

NCE approaches MDDR and FALERT. The RGB parameter is more sensitive to subtle structural changes, whereas RNG is more crude in estimating compound rigidity.

To summarize the RTB, RNG and RGB findings, one can observe that the ACD and ACDF compounds show less molecular complexity (lower number of rotatable and rigid bonds, as well as rings), when compared to pharmaceutical compounds. This is not surprising, since ACD is a collection of chemical reagents, whereas MDDR, MDDRF, FALERT, PDR and NCE contain products intended for medical use. Novel compounds (FALERT, MDDR) tend to have increased complexity (higher number of rings, rigid and rotatable bonds), compared to drugs (PDR and NCE). This could be explained by the increasing pressure to find proprietary structures, which in turn demands increased molecular complexity.

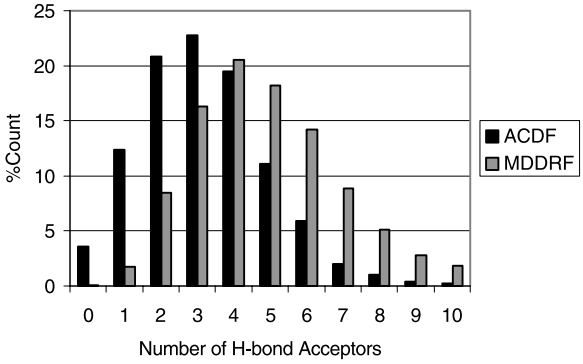

Figure 8. Distribution of hydrogen bond acceptors (HAC) in MDDRF (68523 compounds) and ACDF (150310 compounds).

Number of hydrogen bond donors

The distribution of the number of H-bond donors in MDDRF and ACDF is given in Figure 7. HDO distribution across all the examined datasets was somewhat surprising, if one considers that over 60% of all compounds examined in this study have either 0 or 1 H-bond donors[2]. In particular, 43% of ACD (84445 compounds) have 0 donors, while 28% of ACD (54415 compounds) have 1 donor. Of the pharmaceutical compounds, around 25% have HDO $= 0$, and around 25% have HDO $= 1$. FALERT and NCE compounds peak at 2 HDO, with mid-50% values between 1 and 3, whereas PDR drugs show a decrease in the number of donors: peak at 1, and mid-50% between 0 and 3. MDDR has an intermediate pattern: peak at 1, and mid-50% between 1 and 3. Over 75% of the pharmaceuticals have 1 H-bond donor, or more. ACD and ACDF peak at 1, with mid-50% between 0 and 2. There are 4.8 times more launched drugs in MDDR with $0 \leq$ HDO ≤ 2, when compared to the number of drugs with $3 \leq$ HDO ≤ 5 (see Table 2). A similar, albeit less remarkable, skewness is observed in NCE and PDR: there are 3.11 (in NCE), and 2.56 (in PDR) times more compounds having between 0 and 2 H-bond donors, compared to the number of compounds having between 3 and 5 donors, respectively (see Table 2).

Number of hydrogen bond acceptors

Hydrogen bond acceptors have an asymmetrical Gaussian distribution that resembles the RNG, not the

[2]There are 108320 structures with HDO $= 0$, and 82979 structures with HDO $= 1$, in the combined set of databases (ACD, FALERT, MDDR, NCE and PDR). This number is not corrected for redundant structures.

Table 2. Hydrogen bond donor (HDO) distribution in MDDR, with emphasis on the launched drugs. NCE and PDR distributions are also shown for comparison

HDO	Compounds	Launched drugs	NCE	PDR
0	20298	319	104	108
1	24497	292	106	104
2	16263	210	86	77
3	9044	82	46	65
4	5563	62	37	41
5	3089	27	12	7

HDO distribution (see Figure 8). For the HAC distribution, NCE and MDDR have similar percentage values, that are consistently higher than those from FALERT, whereas ACD and ACDF have the highest cumulative percentage values. PDR distribution resembles ACD and ACDF, not NCE. Over 55% of the pharmaceutically active compounds have 5 or more acceptors. MDDR and FALERT peak at 5 acceptors, with mid-50% between 4 and 8 acceptors. Approved drugs show a slight increase in the number of acceptors: the mid-50% moves from 2–5 (PDR) to 3–7 (NCE). Seventy percent of the drugs have HAC $\geq$ 3 (PDR), and HAC $\geq$ 4 (NCE), respectively. ACD and ACDF peak at 3 acceptors (mid-50% between 2–4), with over 75% of the compounds having HAC $\leq$ 4.

To summarize the H-bond donor and acceptor distributions, one can observe that ACD and ACDF compounds have less donors and acceptors, compared to pharmaceuticals. MDDR and NCE compounds have more similar distributions, whereas FALERT remains at lower percentage values. This is in part caused by the fact that 20% of the FALERT compounds have MW over 600, and contain peptide and peptide-like structures. PDR structures have an intermediate distribution, often comparable to ACD. There are at least 3 times more drugs having between 0 and 2 H-bond donors, compared to those having between 3 and 5 donors.

Overview of the property distribution

A graphical summary of the property distribution across five databases is presented in Figure 9. For each database, the 50% statistical moment ('peak') is presented. For example, ACD and PDR have identical peak values in 2 of the examined properties (RTB and HDO), and have quite similar peak values in 2 other properties (RGB and HAC). Chemical reagents (ACD)

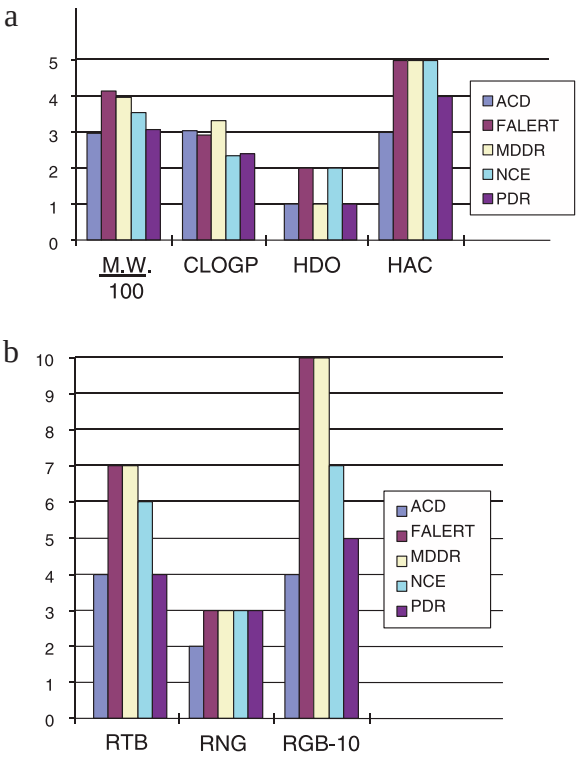

Figure 9. Peak-point (exactly 50% of the distribution) values for the seven analysed properties. To reach a similar scale, M.W. was divided by 100, and RGB by 10.

have reached, today, a similar degree of complexity as the time-tested, orally available drugs (PDR). On the other hand, PDR and MDDR also have 2 identical peaks in the examined properties (RNG and HDO), and similar peaks in HAC. However, when comparing MDDR and ACD, one finds more differences than similarities, as discussed further in this paper. ACDF and MDDRF were not included in Figure 9, since the 50% values are identical to those in ACD and MDDR, respectively.

Discussion

Advantages and limitations of the chosen parameters

The parameters included in this study were chosen after balancing their caveats and advantages. MW is clearly related to size and shape, when metals are not considered [22], but is not a valid estimate for molecular size when halogens (fluorine and iodine, in particular) are included. Other computational options were the molecular volume and the calculated molar refrac-

tivity (CMR). The calculation of the molecular volume necessitates the generation of 3D coordinates with a choice of conformation, whereas CMR is significantly associated with polarizability. MW was preferred due to its simplicity, unequivocal interpretation, and speed of calculation.

Furthermore, CLOGP as an estimate of membrane transferability has also been questioned [23], since membrane transport is a highly complex phenomenon that can be active or passive, and depends on the pH conditions of the microenvironment. This parameter can be corrected according to the partition of various ionized microspecies (namely, LogD), that in turn involves the calculation of pK_a values for each compound in its various tautomeric forms. However, the computational tools that estimate pK_a values for such a diverse set of chemicals are expected to be less reliable. Therefore, we chose CLOGP as rapid, yet reliable estimate of hydrophobicity. In a similar vein, the lack of reliable pK_a estimators reduces the accuracy of the HDO and HAC parameters. For example, the ionization state of barbiturates and coumarols is not accurately estimated in this report. This study was aimed at obtaining a global, albeit less reliable, picture of a chemical space spanning over 300000 compounds, rather than performing an accurate analysis for a significantly lower number of compounds.

The influence of reactive and 'drug-like' compounds

Despite the successful discrimination between 'drugs' and 'nondrugs' [14, 16], an analysis of this type raises legitimate concerns regarding its meaning and its interpretation. One such concern is the presence of chemically reactive species in these databases. Alkylchlorides of the type R-CH_2-Cl, for example, are present in 2773 ACD ACD structures, but also in 598 MDDR structures and in 96 CMC compounds (mostly antineoplastic). Because such reactive compounds may cloud the analysis, these were removed (see also Figure 1). The effect of such structural filtering on the examined property distributions was not found to be significant. Statistical moments for the distribution of MDDR vs. MDDRF, and ACD vs. ACDF always differ by less than 4%. This is not surprising, since both MDDR and ACD are in the order of 10^5 compounds, of which approximately 20% were removed. One can thus assume similar property distributions in each subset. In conclusion, removal of the reactive species did not have a significant influence on property distribution.

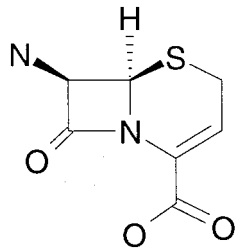

Figure 10. Structure of the cephem nucleus, common to most cephalosporins.

Another source of concern is that ACD contains drugs and 'drug-like' compounds. For example, 28 compounds in ACD contain the cephem nucleus (see Figure 10). Of these, 19 entries are different formulations of cephalosporins, at least 9 of which are approved drugs. However, it is difficult to track down *precisely* which structures in ACD can be of medical use. The amino acid arginine, to take another example, is used as detoxicant in ammonia intoxications, yet it is not often associated with therapeutic purposes. Due to the difficulty of this task, a 'drug-like' scoring scheme [16] was used to remove 5616 compounds from the ACDF subset, to further reduce the influence of 'drug-like' structures on this analysis. We therefore assume that ACDF is composed of 'nondrugs' that do not contain reactive functional groups prone to decomposition in plasma, at pH 7.4.

The content of two other databases, MDDR and FALERT, is often put under scrutiny since these compound collections include structures intended for pharmaceutical use that did not, in an overwhelming percentage, undergo the process of drug development, but are rather disclosed for patent reasons. These databases, however, are likely to reveal trends, since they are likely to include drugs of the future, and should be regarded as relevant to general pharmaceutical use, be it veterinarian or human.

The Pareto principle, the 'rule of 5' test and the 'drug-like' space

CLOGP and MW appear to have a Gaussian distribution in all the compound collections examined so far. They are useful in assisting us to set lower and upper limits, but they tend to drift towards higher values as molecular complexity increases (especially MW, but also CLOGP for the usual hydrocarbon-based chemistry). Both MW and CLOGP lack discriminant power when it comes to differentiating among the databases examined in this study. However, the number of rings,

Table 3. Results from 'rule of 5' test applied to the ACDF and MDDRF datasets. The 'skipped' column includes compounds for which CLOGP calculations failed

	Pass	Fail	Skipped	Total
ACDF	123513	2789	23912	150310
	82.17%	1.86%	15.91%	
MDDRF	54523	4807	9191	68521
	79.57%	7.02%	13.41%	

the number of rigid bonds and the number of rotatable bonds, and less so the number of hydrogen acceptors show an 80/20 relationship that differs among the databases (suggesting an improved discriminant power). For example, 70% of the drug-like compounds have between 0 and 2 H-bond donors, between 2 and 9 H-bond acceptors (except PDR), between 2 and 8 rotatable bonds and between 1 and 4 rings. From another perspective, the number of drugs having between 0 and 2 H-bond donors is at least three times higher than the number of drugs having between 3 and 5 H-bond donors.

The 'rule of 5' test [12] was applied to the MDDRF and ACDF datasets. The results are shown in Table 3. Over 80% of ACDF compounds passed the test, while less than 2% failed. These results are similar to those for MDDRF (79.6% passed, 7% failed), if one assumes that compounds in the 'skipped' column distribute in a similar manner in the 'passed' and 'failed' category, respectively. The fact that 80% of ACDF compounds pass the 'rule of 5' test does not imply that these molecules have good permeation properties, since the 'rule of 5' test is only "a very coarse filter that identifies compounds lying in a region of property space where the probability of useful oral activity is very low" [12]. The results in Table 3 show that the 'rule of 5' test cannot be used to discriminate between 'drugs' and 'nondrugs'. This inability is due to the lack of significant differences in the distribution of the four properties examined in this test, i.e. MW, CLOGP, HDO and HAC. Similar results have, in fact, been reported by Sadowski, who attempted to incorporate the 'rule of 5' parameters in his neural network based 'drug/nondrug' discriminating scheme [24].

Pareto analyses, on the other hand, highlight those properties that have uneven distribution between the ACDF and MDDRF subsets. These can be exploited to focus on the 'drug-like space'. Thus, it appears possible to direct the library content and, indeed, synthetic efforts, towards this property space. This can be achieved by focusing on those properties that have a skewed distribution, e.g. RNG and HDO, and not so much on MW and CLOGP.

Property filtering: A simple tool to reduce the number of virtual compounds

Property filters can be used to remove virtual (or existing) structures that have undesired features. These filters are a convenient way to reduce sample size, in particular for datasets that have a large number of compounds. Property filtering was applied to ACDF (to show 'nondrugs') and MDDRF ('drugs'). Results for two such property filters are given in Table 4.

These property filters allow the introduction of (artificial) boundaries between different populations, as follows: 57.55% of the ACDF and 20.07% of the MDDRF compounds have $0 \leq \text{RTB} \leq 5$, and $0 \leq \text{RGB} \leq 18$. 8.17% of the ACDF and 35.81% of the MDDR compounds populate the opposite volume ($6 \leq \text{RTB} \leq 39$, and $19 \leq \text{RGB} \leq 56$, respectively). While this boundary does not effectively encompass a significant part of MDDRF, one can notice that, in absolute terms, 6 times more ACDF compounds populate the lower volume, and that there are twice as many MDDRF compounds in the higher volume.

In another example, 62.68% of the ACDF and 28.88% of the MDDRF compounds have $0 \leq \text{RNG} \leq 2$, and $0 \leq \text{RGB} \leq 17$ (lower volume). In the higher volume ($3 \leq \text{RNG} \leq 13$, and $18 \leq \text{RGB} \leq 56$), there are 24.73% ACDF and 61.23% MDDR compounds, respectively. This boundary encompasses a significant part of ACDF in the lower volume, and a significant part of MDDRF in the higher region. In relative (percentage) terms, there are 2.5 times more ACDF compounds in the lower region, and 2.12 times more MDDRF compounds in the higher region. However, in absolute terms, a 1:1 ratio is observed in the higher region, even though 5.5 times more ACDF compounds populate the lower volume.

By applying various filtering schemes, one can find different ways to discriminate ACDF from MDDRF via property distribution. Thus, one can define an 'inside' volume of MDDRF (e.g., $\text{RNG} \geq 3$, $\text{RGB} \geq 18$ and $\text{RTB} \geq 6$) that does not overlap with an ACDF 'inside' volume (e.g., $\text{RNG} \leq 2$, $\text{RGB} \leq 17$ and $\text{RTB} \leq 5$). The probability of finding an MDDRF ('drug-like') compound is higher in its 'inside volume' (as profiled above), while the probability of finding an ACDF ('nondrug-like') compound is higher in the 'ACDF in-

Table 4. Sample size in ACDF and MDDRF, after applying various property filters with respect to RNG, RGB and RTB. Filter boundaries show the interval allowed for each property. ACDF% and MDDRF% refer to the percentage relative to the total number of compounds

Filter boundaries	ACDF	ACDF%	MDDRF	MDDRF%
$0 \leq$ RTB ≤ 5; $0 \leq$ RGB ≤ 18	86506	57.55	13735	20.07
$6 \leq$ RTB ≤ 39; $19 \leq$ RGB ≤ 56	12278	8.17	24512	35.81
$0 \leq$ RNG ≤ 2; $0 \leq$ RGB ≤ 17	94208	62.68	16930	24.73
$3 \leq$ RNG ≤ 13; $18 \leq$ RGB ≤ 56	43409	28.88	41908	61.23

side volume', as shown above. Thus, property filtering is a simple, yet efficient tool for sample size reduction that can be performed, for instance, prior to compound selection in a virtual library, or prior to purchase in an HTS library acquisition program.

The skewness of property distribution: Additional comments

It is not surprising that the ACDF 'inside volume', as defined above, contains more 'nondrug' compounds, by comparison to MDDRF. As previously discussed, ACDF is derived from a collection of commercially available compounds, whereas MDDRF includes more elaborate structures, intended for medical use. Ninety percent of all bio-active compounds have a maximum of 5 rings, whereas 64% of ACD compounds have 2 rings or less. Rigid compounds often gain target affinity (free energy of binding) by reducing the entropic contribution. This may explain why more compounds with increased rigidity (RNG, RGB) populate the 'MDDRF inside volume'. The fact that a large proportion of the available drugs are either antagonists or inhibitors may also have an effect on this skewed distribution. ACD compounds have less rotors, but also less rings, on the average, indicating a collection of somewhat simpler molecules.

Ninety percent of all compounds have between 0 and 5 H-bond donors. Over 65% of all pharmaceutically active compounds have $0 \leq$ HDO ≤ 2, while almost 50% of them have 0 or 1 donor. Eighty percent of all compounds have $1 \leq$ HAC ≤ 8, with almost 50% of them between 3 and 7. Over 65% of the compounds in PDR and NCE have $2 \leq$ HAC ≤ 6, while almost 60% of ACD compounds have significantly less acceptors, between 0 and 3. Some possible reasons for this distribution are outlined below.

Typical membrane lipids lack hydrogen bond donors, since lipid head groups are esters. These esters are solvated by water. Thus, hydrogen bonds involving drugs as donors and the lipid head groups as acceptors may occur *twice* during trans-membrane passage. This, in turn, hinders trans-membrane diffusion more than the presence of hydrogen bond acceptors in the drug molecule. Hydrogen bond acceptors are water-solvated, like the lipid head groups. This is less likely to impair the trans-membrane diffusion. This may explain why the number of H-bond donors in a drug structure is more closely regulated, and appears to be more important than the number of acceptors in our property distribution analysis.

The apparent penalty for introducing H-bond donors in a drug-like molecule is further substantiated by the limited diversity of H-bond donors. These are either O-H or N-H groups. Hydroxyl functions are, in fact, incorporated in (hydrophobic) drugs via cytochrome P450 enzymes [25] during phase I catabolic biotransformations [26]. This is followed by phase II, in which hydroxyl groups are conjugated by glucuronic, sulfuric, acetic or other acids prior to urinary excretion [26]. Among the N-H functions, amides are also known to create problems for absorption of peptides, which leaves chemists with only (substituted) amines, guanidines, amidines, ureas, carbamates, sulfonamides and sulfamides to choose from.

Conclusions

To expedite the computer-assisted compound selection for libraries of virtual and/or existing compounds, we have examined several commercially available compound collections, with respect to the distribution of several properties: molecular weight, CLOGP (using Leo's method), the number of hydrogen bond donors

and acceptors, the number of rings and the number of rotatable and rigid bonds. Of these parameters, CLOGP and MW have a Gaussian distribution in all databases. However, the number of H-bond donors, the number of rings and, to some extent the number of rotatable and rigid bonds and the number of H-bond acceptors follow an 80/20 relationship. The 'rule of 5' test produced similar results when applied to the ACDF and MDDRF subsets (Table 3), and consequently cannot be used to discriminate between 'drugs' and 'nondrugs'. Therefore, we have focused on RNG, RGB and RTB in an attempt to define simple property boundaries for zooming towards the 'drug-like' property space.

Based on Pareto analyses, we observed that 'drug-like' structures from MDDRF are more likely to be found at the higher end of the property space (RNG $\geq$ 3, RGB $\geq$ 18, RTB $\geq$ 6), whereas 'nondrug-like' structures from ACDF are more probable in the lower end (RNG $\leq$ 2, RGB $\leq$ 17, RTB $\leq$ 5). Such filters can be used to bias library contents prior to purchase (HTS) or synthesis (CBC) of compounds.

These property filters *complement*, and do not replace, the library analysis at the structural (chemical) level or the 'rule of 5' test proposed by Lipinski and co-workers [12]. However, this filtering process is always context-dependent (what is the drug target? what is the purpose of the library?, etc.) and cannot be applied indiscriminantly. One can expect different property filters when looking for, e.g., alpha-adrenergic blockers active in hypertension, compared to chemotherapeutic agents for urinary infections. Property filtering can be combined with chemical structure filtering and with 3D-structure filtering when the receptor structure is available. This will expedite library analysis and rule out putative ligands that will not fit into the binding site. One can, for example, use the steric fit parameter, which is a simple measure to evaluate how much of the ligand is packed into the binding site [27].

Skewed property distributions can be rationalized, at least in part, by the constraints imposed on the 'drug-like' property space. Pharmaceutically active compounds tend to be more rigid due to entropic factors, since increased rigidity is often associated with higher binding affinity. Furthermore, pharmaceutically active compounds and drugs are likely to have low numbers of H-bond donors due to constraints imposed by the membrane transfer process, by the phase II biotransformation, and by the limited diversity of functional moieties that exhibit H-bond donors.

Acknowledgements

Andrew Davis (Astra Charnwood) has been influential in discussing concepts and providing the PDR database. The following people from Astra Hässle are acknowledged: Vladimir Shcherbukhin and Thomas Ohlsson for discussing and implementing filtering schemes; Stefan Ervik for programming assistance; Magnus Björsne, Johan Gottfries, Thomas Kühler, Ingemar Nilsson, Bo Nordén and Bertil Samuelsson for useful discussions and for reading the manuscript.

References

1. a. Kubinyi, H., Perspect. Drug Discov. Design, 9–11 (1998) 225.
   b. Clark, R.D., Ferguson, A.M. and Cramer, R.D., Perspect. Drug Discov. Design, 9–11 (1998) 213.
2. a. Pearlman, R.S. and Smith, K.M., Perspect. Drug Discov. Design, 9–11 (1998) 339.
   b. Van Drie, J.H. and Lajiness, M.S., Drug Discov. Today, 3 (1998) 274.
   c. Walters, W.P., Stahl, M.T. and Murcko, M.A., Drug Discov. Today, 3 (1998) 160.
3. Rishton, G.M., Drug Discov. Today, 2 (1997) 382.
4. Walters, W.P., Stahl, M.T. and Murcko, M.A., in von Ragué Schleyer, P. (Ed.) Encyclopedia of Computational Chemistry, Vol. 2, Wiley, New York, NY, 1998, pp. 1225–1237.
5. Available from MDL Information Systems, http://www.mdli.com/dats/pharmdb.html. The MDDR database is developed in cooperation with Prous Science Publishers, http://www.prous.com/index.html. Our MDDR subset contains 82903 structures.
6. Available from Current Drugs, http://www.current-drugs.com/products/iddb/idpfast.html. FALERT is based on Current Drug's patent coverage of therapeutic compounds.
7. Available from MDL Information Systems, http://www.mdli.com/dats/pharmdb.html.
8. A compilation of orally available drugs from the Physician Desk Reference, 1994.
9. A compilation of new chemical entities introduced between 1980 and 1996. Provided by C.A. Lipinski (Pfizer, Groton, U.S.A.).
10. Available from MDL Information Systems, http://www.mdli.com/dats/pharmdb.html. The ACD database is a compilation of over 250000 commercially available substances from over 500 catalogs worldwide. Our ACD subset contains 194511 structures.
11. Comprehensive Medicinal Chemistry, 6 Volumes, in Hansch, C. (Ed.) Pergamon Press, London, 1990.
12. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., Adv. Drug Deliv. Rev., 23 (1997) 3.
13. CLOGP is included in the PCMODELS 4.6 program, available from Daylight Chemical Information Systems, http://www.daylight.com/products/pcmodels.html.
14. a. Ajay, Walters, W.P. and Murcko, M.A., J. Med. Chem., 41 (1998) 3314.
    b. Sadowski, J. and Kubinyi, H., J. Med. Chem., 41 (1998) 3325.

15. Available from Daylight Chemical Information Systems, http://www.daylight.com.
16. V.V. Shcherbukhin, T.I. Oprea and U. Norinder. Manuscript in preparation.
17. The MDDR file was extracted using the October 1998 version of MDDR (82903 structures). MDDRF was extracted using the July 99 version of MDDR, that contained 103690 entries, of which 78772 were included in MDDRF.
18. Olsson, T. and Shcherbukhin V., Synthesis and Structure Administration (SaSA) (1997) developed at Astra Hässle AB, http://www.astrazeneca.com.
19. Downs, G.M., in von Ragué Schleyer, P. (Ed.) Encyclopedia of Computational Chemistry, Vol. 4, Wiley, New York, NY, 1998, pp. 2509–2515.
20. Koch, R., The 80/20 Principle, Nicholas Brealy Publishers, London, 1997.
21. a. Hansch, C. and Leo, A., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, ACS, Washington, DC, 1995, pp. 125–168.
    b. See the Biobyte webpage for a detailed analysis of the CLOGP accuracy, http://clogp.pomona.edu/medchem/chem/clogp/starlist/index.html.
22. Oprea, T.I. and Waller, C.L., in Lipkowitz, K.B. and Boyd, D.B. (Eds) Reviews in Computational Chemistry, Vol. 11, Wiley-VCH, New York, NY, 1997, pp. 127–182, see the reference 138 in particular.
23. a. Chan, O.H. and Stewart, B.H., Drug Discov. Today, 1 (1996) 461.
    b. Ungell, A.-L., Drug Develop. Ind. Pharmacy, 23 (1997) 879.
    c. See the ACD Labs web site, http://www.acdlabs.com/ for instances where LogP and LogD differ significantly.
24. Sadowski, J., Optimization of the drug-likeness of chemical libraries. International workshop on virtual screening, Schloss Rauischolshausen, March 1999.
25. Oprea, T.I., Hummer, G. and García, A.E., Proc. Natl. Acad. Sci. USA, 94 (1997) 2133 and references therein.
26. Benet, L.Z., Kroetz, D.L. and Sheiner, L.B., in Hardman, J.G., Limbird, L.E., Molinoff, P.B., Ruddon, R.W. and Goodman Gilman, A. (Eds) Goodman & Gilman's Pharmacological Basis of Therapeutics, 9th Edition, McGraw-Hill, New York, NY, 1996, pp. 3–28.
27. Oprea, T.I. and Marshall, G.R., Perspect. Drug Discov. Design, 9–11 (1998) 35.