Pharmacophore Features Distributions in Different Classes of Compounds

Fabio Zuccotto*

Inpharmatica Ltd., Molecular Design, 60 Charlotte Street, London, W1T 2NU, United Kingdom

Received April 11, 2003

A pharmacophore analysis approach was used to investigate and compare different classes of compounds relevant to the drug discovery process (specifically, drug molecules, compounds in high throughput screening libraries, combinatorial chemistry building blocks and nondrug molecules). The distributions for a set of pharmacophore features including hydrogen bond acceptors, hydrogen bond donors, negatively ionizable centers, positively ionizable centers and hydrophobic points, were generated and examined. Significant differences were observed between the pharmacophore profiles obtained for the drug molecules and those obtained for the high-throughput screening compounds, which appear to be closely related to the nondrug pharmacophore distribution. It is suggested that the analysis of pharmacophore profiles could be used as an additional tool for the property-based optimization of compound selection and library design processes, thus improving the odds of success in lead discovery projects.

INTRODUCTION

At present, drug discovery is concentrated on a relatively small number of targets.[1] However, it is estimated that the completion of the human genome sequence will dramatically increase the number of new drug targets available to pharmaceutical research.[2] The increasing number of targets and the high attrition rate of drug candidates normally associated with the drug discovery process[3] are clear signals that, to effectively sustain its financial and R&D success, the pharmaceutical industry needs more and better lead molecules (i.e. small molecules that show target binding affinity, have a favorable absorption, distribution, metabolism and excretion profile, and which are chemically tractable to allow their optimization in the context of a medicinal chemistry project).

The recent development of technologies such as Combinatorial Chemistry (CC) and High Throughput Screening (HTS) has significantly increased the capacity for the preparation and screening of a large number of compounds in a fast and cost-efficient way. Compound libraries of 10^2−10^6 molecules are routinely prepared and screened in the industry. However, despite this enormous effort, only a small fraction of the chemical space, containing an estimated 10^{60}−10^{100} molecules, is available for screening.[4] Synthetic accessibility, reagents, time and costs, are among factors that limit the portion of chemical space that can be investigated.

In Silico methods allow the evaluation of a larger portion of molecules and can have a significant impact in drug discovery. The contribution of virtual screening (VS) to the lead generation process, for instance, has been discussed by Walters and colleagues.[4] It is clear that in the design of combinatorial libraries, or during the selection of compounds for HTS, there is a need to restrict the molecular space to investigate, ideally enriching it with molecules that are diverse and good drug candidates (i.e. drug-like compounds).

The concept of drug-likeness has been extensively explored, and many computational efforts have been directed at determining the basic requirements for a molecule that may become a drug.[5−7] In one of the best known studies, Lipinski[8] and co-workers discussed the oral-absorption of a molecule in terms of molecular weight, calculated octanol/water partition coefficient, and number of hydrogen bond donors and acceptors. Although the study was not originally aimed to classifying drug-like molecules, it highlighted, in the so-called 'rule of five', a simple and successful relationship between a set of molecular descriptors and oral availability that is intrinsically related to the notion of drug-likeness.

Recently, several attempts have also been made to develop computational methods to discriminate between drugs and nondrugs using either neural networks,[9] genetic algorithms,[10] property filters[11] or scoring methods.[12,13] At the same time, a large number of physicochemical and structural features have been used in the attempt to characterize drug-like molecules. In addition to Lipinski's 'rule of five', several other one or two-dimensional descriptors have been used including the following: number of rotatable bonds, number of rigid bonds, number of aromatic rings, shape and branching descriptors, and fingerprint keys.[9]

In an effort to reduce the number of compounds failing in a late stage of development, there is a clear trend to incorporate the notion of drug-likeness as early as possible in the drug discovery process and to place more emphasis on what is now referred to as property-based drug design.[14]

The concept of pharmacophore is key in medicinal chemistry and is part of the common background of each medicinal chemist. The binding of small molecules to a macromolecule, for instance, is often described as a pharmacophore recognition event, and different combinations of pharmacophore points representing potential interactions are used to query databases in lead generation and lead optimization programs.

To date, none of the available studies on property distribution has investigated this important set of descriptors in a systematic and exhaustive manner.

* Corresponding author phone: +44-207-0744681; fax: +44-207-0744600; e-mail: f.zuccotto@inpharmatica.co.uk.

10.1021/ci034068k CCC: $25.00 © 2003 American Chemical Society
Published on Web 07/31/2003

In the 'rule of five', two of the chosen parameters are pharmacophore features: hydrogen bond donors and hydrogen bond acceptors expressed as sum of NH and OH bonds and as sum of N and O atoms, respectively. The pharmacophore definitions used by Lipinski and co-workers are a fine example of simplicity and effectiveness and are easy to implement in a computational tool, but they can only provide a limited and approximated view of the functional space associated with a molecule.

The aim of the present study was to investigate and compare different classes of compounds, ranging from drug-like to nondrug-like, in terms of an extensive set of molecular functional elements defining five different pharmacophore features. Clearly, the molecular functional space is intrinsically related to the molecular structure, but because it does not represent a single specific (sub)structural component, it is probably better suited to capturing the important elements of ligand binding in their essence rather than their specificity. Another advantage in using pharmacophore features to describe the molecular space is that they are not calculated molecular properties and hence not subject to calculation errors.

Initially, a set of (sub)structural descriptors covering drug relevant functional groups was used to analyze a set of drug structures, and then a suitable combination of the descriptor distributions in the drug space was used to derive a drug-like pharmacophore profile. Similarly, a nondrug-like pharmacophore profile was calculated by analyzing molecules not associated with biological activity.

In the lead identification process, high-throughput screening of compound collections and combinatorial libraries has become the norm in the pharmaceutical industry. In this context, chemical diversity and sample availability are obviously issues of significant importance, and an extraordinary effort has been made throughout the industry to enhance corporate databases of compounds, both in terms of number of samples and molecular diversity. Compound acquisition programs have fueled the growth of a parallel business focused on the production and supply of chemical entities for screening. Several large databases of commercially available compounds are now playing an important role in the supply of compounds for screening. It was our interest, therefore, to also analyze compounds in such databases and derive an 'HTS' pharmacophore profile to assess their drug-likeness.

The possibility to identify and define portions of pharmacophore space that include a higher or lower proportion of drug-like molecules would allow us to introduce landmarks in the complex chemical space, providing a valuable guide in the identification of new active compounds. The analysis of the pharmacophore distribution in collections of molecules should provide a valuable tool in the optimization of the diversity and biological relevance of the compounds in novel combinatorial libraries for HTS or virtual libraries for VS, increasing the effectiveness of the lead discovery process. In the context of a particular medicinal chemistry project, target-specific pharmacophore profiles could also be derived and used in the design of focused libraries for lead optimization screening. To help in the interpretation of the results, a 'lead-like' profile was also generated from a set of 62 molecules identified as lead compounds,[15] and the pharmacophore profile for the cycloxygenase enzyme (COX), a common target in the treatment of pain and inflammation, was derived as an example.

Table 1. Database Studied and Percentage of Analyzed Compounds

	MDDR	CMC	ACD	HTS1	HTS2	BbCc
initial	128269	8474	403050	187802	55619	13318
analyzed	111738	6631	298626	187026	54776	11385
analyzed %	87.13	78.25	74.09	99.59	98.48	85.49

METHODS

This study of pharmacophore profiles was performed using structures extracted from a set of widely available commercial chemical databases including the MACCS-II Drug Data Report (MDDR)[16] and the Comprehensive Medicinal Chemistry (CMC)[17] as drug-like databases, the Available Chemical Directory (ACD)[18] as nondrug-like database, two databases of commercial compounds for high-throughput screenings, and one database of building blocks used in combinatorial chemistry (Table 1).

For each database, a filter cascade, similar to the one used by Muegge and co-workers,[11] was applied to remove the following: molecules with atoms other than C, N, O, S, H, P, Si, Cl, Br, F, I; counterions; solvent molecules; and compounds with molecular weight outside the 150−750 Dalton range. Duplicate entries and entries with errors were also removed. Chemically reactive compounds were retained.[19]

The pharmacophore profile for a drug-like molecule was derived from a subset of the MDDR, here referred to as MDDR-Launched (MDDR-L), containing compounds classified as launched, phase 1−3, preclinical, clinical, preregistered, and registered. Other compounds in the MDDR classified as biological-testing were included in a different subset, here referred to as MDDR biological test (MDDR-BT), whereas the compounds classified as withdrawn were ignored. The molecules in the MDDR databases are assigned to activity classes. To further characterize the drug-like element of the databases, compounds with no therapeutic class assigned and compounds belonging to undesired classes[20] were removed. Of the final 111 738 compounds in the filtered MDDR, 103 487 were assigned to the MDDR-BT subset and the remaining 8251 to the MDDR-L subset which was assumed to represent the drug-like chemical space.[9] Ideally, MDDR-L should include a wide variety of active, structurally diverse compounds and not be biased toward any particular therapeutic area. An inspection of the MDDR-L subset showed that a wide range of biological activities is represented. To assess the structural diversity of MDDR-L, the nearest neighbor similarity was calculated using the Tanimoto[21] coefficients (Unity[22] 2D fingerprints). It was found that for at least 66% of the compounds there was no other similar structure in the database (cumulative percentage of 66.55 for a nearest neighbor Tanimoto value greater than 0.85).[23] In a similar way, compounds that fell into undesired classes[20] were removed from the CMC database, and a second drug-like pharmacophore profile was derived.

The pharmacophore profile for a nondrug-like molecule was derived from the ACD database. To further characterize the nondrug-like components of the database, compounds

Table 2. Main Structural Features Used To Define Pharmacophore Points

hydrogen bond donors	hydrogen bond acceptors	positively ionizable centers	negatively ionizable centers	hydrophobic site
nitrogen donors (aliphatic)	carboxylic	amidine	carboxylic acid	six/five-member rings
nitrogen donors(aromatic)	carbonyl	protonable amine	tetrazole	*tert*-butyl
hydroxyl	S/P oxygens	guanidinium	acid sulfona(i)mides	hydrophobic moiety
thiols	hydroxyl/phenol	2/4-amino pyridine	hydroxamic acid	halogens
	ether	charged nitrogen atoms	S/P acids	
	ester			
	nitro			
	nitrogen acceptors			

also found in MDDR-L or CMC were removed. The remaining set was assumed to represent the nondrug-like chemical space. In practice, it is possible that other molecules in ACD are biologically active compounds, but the relevance of these was considered to be negligible.

Two databases of compounds for HTS and one database of building blocks for combinatorial library generation from well-established suppliers were investigated: the Gold collection from Asinex,[24] the Maybridge HTS collection, and the Maybridge CC collection[25] (here referred to as HTS1, HTS2, and BbCc). The results obtained for the BbCc database are clearly not directly comparable with the other data. Although they are used to prepare compounds for lead identification/optimization, the molecules used as building blocks for the combinatorial library are reactive intermediates and are usually poorly functionalized and have low molecular weights.

To derive the COX pharmacophore profile, a dedicated database was generated including 1865 compounds with reported COX activity.[26] Compounds in this set were not filtered.

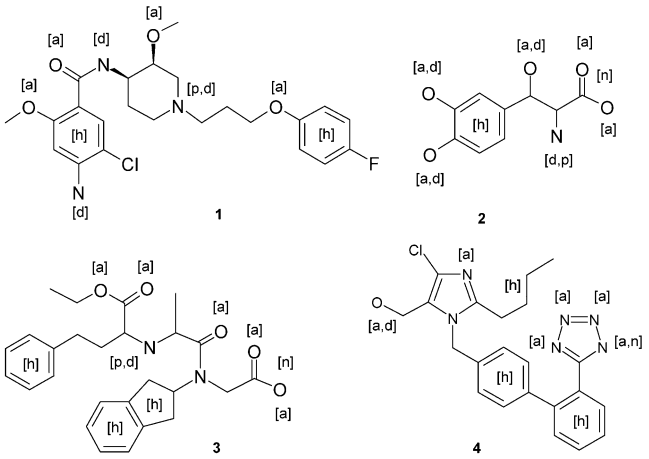
The pharmacophore profiles of the studied databases were obtained by evaluating the distribution of a set of functional features (pharmacophore points) in the databases. The pharmacophore descriptors included in this study were as follows: hydrogen bond acceptor (HbA); hydrogen bond donor (HbD); negatively ionizable center (NI); positively ionizable center (PI); and hydrophobic sites (Hy).

It is common in this type of study to classify groups that might display basic or acid behavior at physiological pH as hydrogen bond donors and acceptors. Davis and Teague[27] discussed the different contribution of charged and neutral ligand−protein interactions to the overall binding energy reporting that, typically, a neutral−neutral hydrogen bond is responsible for a 2−15-fold increase in ligand affinity, whereas a charged hydrogen bond contributes up to 3000-fold increase. The presence of ionized groups in a compound clearly also has serious implications in the cross membrane processes, reducing the rate by which the compound can diffuse across the cell membrane. Therefore, the use of a set of pharmacophore points that include and differentiate groups that can be charged during the binding should better capture the functional profile of the compounds and lead to an improved description of drug-likeness.

The Tripos' line notation for chemical structures SLN (Sybyl Line Notation)[28] was used to define a set of functional groups or substructural motifs responsible for the pharmacophore features as described in Table 2. These SLN definitions[29] were then implemented in the molecular modeling package Sybyl (version 6.8),[30,31] and SPL (Sybyl Programming Language)[28] scripts were used to identify the type and number of pharmacophore points in each molecule of the databases.

It is well accepted that the binding of a drug is mediated by ion−ion interactions, hydrogen bonding, dipole−dipole interactions, lipophilicity, and shape complementarity. However, the contribution of each of these interactions is still poorly understood although it is clearly a system-dependent event, influenced by, among other things, the different structural features of the molecule, the pH of the solvent, and the type of residues involved in the interaction. For instance, an acidic group can be involved in a charge−charge interaction or act as a hydrogen bond acceptor depending on its ionization state and the residue(s) it interacts with. Similarly, a hydroxyl group will behave as a hydrogen bond donor or acceptor in different circumstances.

Even with the aid of structural information on the target(s) and sophisticated drug design tools, the determination of the correct orientation and conformation of the bound ligand is still a lengthy and sometime unreliable process. It is not easy then to determine which particular functional groups (pharmacophores) of the molecule are involved in the ligand−protein interaction, with hydrophobic interactions being particularly challenging to determine.[27] Even if this type of information were easily available, again it would be completely target and method-dependent. Bearing this in mind, our effort was focused on mapping all possible points of interaction (pharmacophore points) present in each structure in the databases. It was assumed that a single atom or functionality can provide the molecule with more than one pharmacophore point. For example, in a given compound, a carboxylic acid group could take part in a charge−charge interaction or be involved with up to two hydrogen bond donors present within a protein binding site. Its contribution to the total count of pharmacophore points for the molecule would therefore be 1 NI and 2 HbAs. In our work, no pK_a evaluation was attempted; each NI and PI center was considered as potentially ionizable regardless of the presence of other such groups in the molecule. Overall, the most general possible SLN definitions were implemented using exclusion definitions to focus on the desired substructure. In particular, the NI definitions included carboxylic and hydroxamic acids, tetrazoles, sulfonamides, and acid sulfonamides. Acids of sulfur and phosphorus in different oxidation states were also considered. PI was defined as amidine, guanidinium, protonable amine, and 2/4-amino pyridine type of compounds. Nitrogen atoms with a formal positive charge were also included. Several substructures containing different oxygen atoms were classified as HbA: carbonyl and hydroxyl groups, esters, carboxylic acids, ethers, and nitro groups. Only a limited number of substructures containing a nitrogen atom were included in this group,



Compound	[a]	[d]	[p]	[n]	[h]
1	4	3	1	0	2
2	5	4	1	1	1
3	5	1	1	1	3
4	6	1	0	1	3

Figure 1. Examples of pharmacophore mapping on four compounds from the MDDR-L set, with [a] denoting acceptor groups, [d] donor groups, [n] negatively ionizable center, [p] positively ionizable center, and [h] hydrophobic site, respectively.

namely pyridine-like and imino nitrogens. Tertiary amines, amide and amide-like (e.g. sulfonamides, urea) nitrogens were not considered as HbA. All nitrogens carrying a hydrogen atom (including those that could be protonated) and the hydroxyl groups were classed as HbD. Despite their weak propensity to give H-bonding,[32] thiol groups are normally treated as HbD by the software used for pharmacophore screening, hence they were also included. If the definition of the NI, PI, HbD and HbA pharmacophore points was made following specific criteria, the definition of the Hy points was slightly more arbitrary. It was decided to consider hydrophobic points all the five- and six-membered rings (tetrazole excluded),[33] *tert*-butyl groups, nonaromatic halogen, and any carbon atom that was at least two carbon atoms distant from any heteroatom or any other carbon atom already classified as hydrophobic points.

The definitions used are in no way exhaustive of all possible HbAs, HbDs, PIs, NIs, and Hys present in the chemical space, but the author believes that the set used well describes the most likely ones to be present in the molecules that are typically screened in the pharmaceutical industry. For a set of four compounds, Figure 1 exemplifies the pharmacophore point mapping. The SLN used to define the pharmacophore points used in this study are available as Supporting Information.

For each database, the total number of occurrences and their distribution were calculated for the five pharmacophore features described above. The distributions were represented as a set of bins with bin size set to unity. Each bin contained the total number of molecules in the database with a number of the corresponding feature equal to that particular bin range. The last bin also contained all the molecules with a number of features greater than its bin size. For the pharmacophores HbA, HbD, and Hy, sets of 10 bins were used, while for Ni and PI, sets of 5 were employed. For example, for HbA, the first bin represented the number of molecules with no hydrogen bond acceptor moieties, the second bin represented the number of molecules with exactly one hydrogen bond acceptor, and so on (the last bin contained the number of molecules with 9 or more acceptors). A total of 670 182 molecules were processed, and more than 5.2 million pharmacophore points were classified.

Table 3. Values of χ^2 Calculated To Compare the Distribution of Features in the MDDR-L Set and the Other Databases[d]

	HbA[a]	HbD[a]	PI[b]	NI[b]	Hydro[a]
MDDR-BT	31.05	**23.41** (1)[c]	32.43	**9.26** (3)[c]	240.10
CMC	330.53	118.74	**14.12** (2)[c]	130.46	622.01
ACD	2838.39	4303.91	6798.88	1364.02	1156.99
HTS1	1314.60	6891.35	5407.27	6566.35	647.40
HTS2	3955.00	4950.51	6178.07	3642.18	1613.89
BbCc	3177.80	4546.54	3456.03	1203.71	2038.16

[a] $\upsilon = 9$, χ^2 critical value ($p \leq 0.001$) 27.88. [b] $\upsilon = 5$, χ^2 critical value ($p \leq 0.001$) 20.51. [c] (1) $p = 0.0053$, (2) $p = 0.1486$, (3) $p = 0.099$. [d] In bold nonsignificant values.

The significance of the differences between the feature distributions in the MDDR-L set and the distribution in the other databases was evaluated using the χ^2 test (Table 3).

The Pareto principle[34,35] was applied to the feature distributions to identify the most relevant subset of data points. In particular, for each pharmacophore, we were interested in the smallest number of features that could represent at least 70% of the data.

To aid the analysis and visualization of the results, in some instances the five-dimensional space defined by the pharmacophore descriptors HbA, HbD, NI, PI, and Hy was projected in two dimensions using multidimensional scaling (MDS).[36] The distributions of the singular structural moieties incorporated in the pharmacophore element were also investigated.

RESULTS

Overall, the distributions of the pharmacophore features clearly show a considerable difference in the composition of the examined databases (Table 4). In agreement with the results of other studies[5,15] it is shown that the drug-like compounds are likely to be more complex chemical entities containing a larger number of pharmacophore features compared to the molecules in the nondrug-like space. The average number of pharmacophore elements decreases from 10.15 in MDDR-L to 7.06 in ACD (Table 4). As shown in Table 5, at least 70% of the compounds in the MDDR-L database presented 6−13 pharmacophore points, compared to 4−9 in ACD. HTS and BbCc compound collections have a number of features similar to nondrug-like compounds (average value of 8.04 for HTS1, 6.86 for HTS2, and 5.59 for BbCc). The χ^2 test results in Table 3 clearly highlight the differences between the distributions of the pharmacophore features in the nondrug-like and HTS molecular space and the corresponding distributions in the drug-like space. Interestingly, HbD and PI provide the greatest discrimination for ACD, HbD, and NI for HTS1 and PI and HbD for HTS2. In all cases the smallest difference in distribution is observed for the hydrophopic features. Not surprisingly, less pronounced differences are observed between MDDR-L and other sets of compounds close to the drug-like space, such as MDDR-BT and CMC, but in these cases the hydrophobic

Table 4. Average Distribution of Pharmacophore Points in the Studied Databases, the Set of Lead Molecules, and the COX Set[a]

database	HbA	HbD	PI	NI	Hy	total
MDDR-L	4.18	1.93	0.60	0.30	3.14	10.15
MDDR-BT	4.13 (−1.0%)	1.86 (−3.80%)	0.57 (−5.2%)	0.29 (−2.2%)	3.34 (+6.2%)	10.19 (+0.4%)
CMC	3.67 (−12.1%)	1.76 (−8.8%)	0.59 (−1.7%)	0.21 (−28.6%)	2.59 (−17.6%)	8.82 (−13.1%)
ACD	3.10 (−25.8%)	1.08 (−44.2%)	0.18 (−70.3%)	0.14 (−54.4%)	2.71 (−13.7%)	7.22 (−28.9%)
HTS1	3.56 (−14.8%)	1.19 (−38.4%)	0.20 (−66.1%)	0.05 (−82.4%)	3.30 (−3.4%)	8.04 (−20.8%)
HTS2	2.90 (−30.7%)	1.04 (−46.1%)	0.16 (−73.9%)	0.06 (−79.4%)	2.71 (−13.9%)	6.86 (−32.4%)
BbCc	2.40 (−42.5%)	0.69 (−64.2%)	0.13 (−77.8%)	0.08 (−73.8%)	2.29 (−27.2%)	5.59 (−44.9%)
Lead set	2.63 (−37.1%)	1.87 (−3.1%)	0.58 (−3.3%)	0.23 (−23.3%)	2.08 (−33.8%)	7.39 (−27.2%)
COX set	3.00	0.98	0.12	0.10	3.32	7.52

[a] In parentheses, when present, the percentage difference from the MDDR-L set that is used as a benchmark.

Table 5. Minimum Range Describing at Least 70% of the Data Is Reported for Each Pharmacophore Feature and Database[a]

	HbA	HbD	PI	NI	Hydro	total
MDDR-L	2−6 (75.0%)	1−3 (73.1%)	0−1 (91.1%)	0 (75.7%)	2−4 (73.2%)	6−13 (74.9%)
MDDR-BT	2−6 (75.1%)	0−2 (74.4%)	0−1 (91.1%)	0 (76.3%)	2−4 (73.2%)	6−12 (70.6%)
CMC	1−5 (77.0%)	1−2 (79.3%)	0−1 (91.9%)	0 (83.1%)	1−3 (76.0%)	5−11 (74.4%)
ACD	1−4 (74.2%)	0−1 (72.1%)	0 (84.4%)	0 (88.1%)	1−3 (74.0%)	4−9 (75.1%)
HTS1	2−5 (77.7%)	0−2 (91.83%)	0 (82.3%)	0 (95.2%	2−4 (82.1%)	6−10 (73.3%)
HTS2	1−4 (81.0%)	0−1 (71.7%)	0 (86.2%)	0 (94.3%)	2−4 (81.0%)	5−9 (76.1%)
BbCc	1−4 (82.1%)	0−1 (83.6%)	0 (88.2%)	0 (92.7)	1−3 (85.2%)	3−7 (77.9%)

[a] In parentheses the actual percentage of data covered.

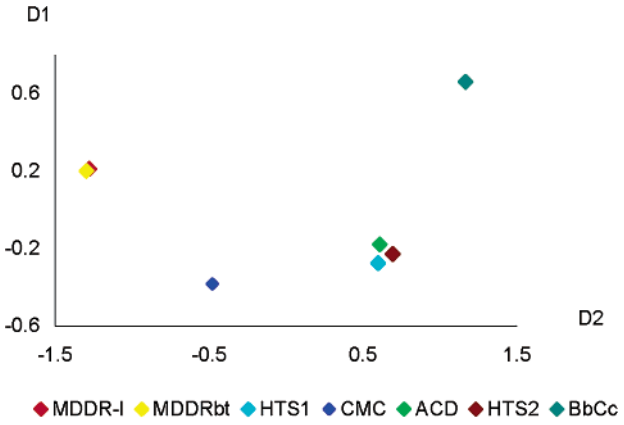


Figure 2. Position of the examined databases in the pharmacophore space described by the average values of the descriptors HbA, HbD, NI, PI, and Hy. Multidimensional scaling was performed to reduce the dimension of the space from 5 to 2. The new dimension D1 and D2 in the lower dimensional pharmacophore space do not retain the same chemical meaning.

feature gives the greatest discrimination. The position of the databases in the pharmacophore space described by the average of the chosen descriptors HbA, HbD, NI, PI, and Hy is illustrated in Figure 2. The result shows two distinct points and two distinct clusters. One that includes both subsets of the MDDR database, and another where the ACD and the HTS compounds are closely grouped together. The CMC lies in a region in the middle of the two, whereas BbCc is the most isolated in the pharmacophore space.

If the total number of pharmacophore features can describe the complexity of a molecule from a functional point of view, the distribution of the single pharmacophore descriptor is probably more helpful in identifying clear differences between the drug and nondrug-like space. For each pharmacophore, the differences between their average occurrences in the MDDR-L molecules (assumed to be representative of the drug-like space) and their average number in the other databases are presented in Figure 3.

Hydrogen Bond Acceptors (HbA). As expected, hydrogen bond accepting groups are the most likely features to be present in a molecule, although the average number ranges quite significantly across the studied databases. The average number of HbA in the launched molecules in MDDR is 4.18, more than 25% higher than the number of HbA present in the molecules of the ACD. In all databases, HbA had an asymmetrical Gaussian distribution with maximum between 2 and 4 (Figure 4). In databases of drug-like compounds the maximum of the curve was characterized by low values, between 15 and 20%, compared to the values of 22−27% for the molecules in nondrug-like databases. A large portion (>70%) of the compounds in MDDR-L and MDDR-BT had HbA ranging from 2 to 6, whereas HTS2 and BbCc shared the same range of $1 \leq \text{HbA} \leq 4$ observed for the nondrug-like ACD. With a number of hydrogen bond acceptor ranges of 2−5 and 1−5, respectively, HTS1 and CMC were placed between the two extremes.

As shown in Table 6, the most common HbA structural feature in all databases was the carbonyl group. The other HbA groups that are the most likely to be related to the drug-like space are groups containing nitrogen atoms as acceptors, followed by ethers, hydroxyl, and finally carboxylic acids. Interestingly, the distribution of the HbA features account for the major difference between the MDDR-L and MDDR-BT. In fact, the average number of ether groups was more than 60% lower in MDDR-BT, whereas the number of esters was 170% higher.

The two HbA structural moieties that differed the most between drug and nondrug-like spaces were the nitro and hydroxyl groups. The number of nitro groups in the MDDR-L accounted for only 1.8% of the total number of HbA and was less than a third of the number of the nitro groups present in ACD (6.4%) suggesting that, in a drug molecule, its major role is probably the fine-tuning of the pharmacokinetic properties rather than H-bonding. The contribution of the hydroxyl groups to the HbA portion of MDDR-L is about 15% and is almost double the one in ACD.

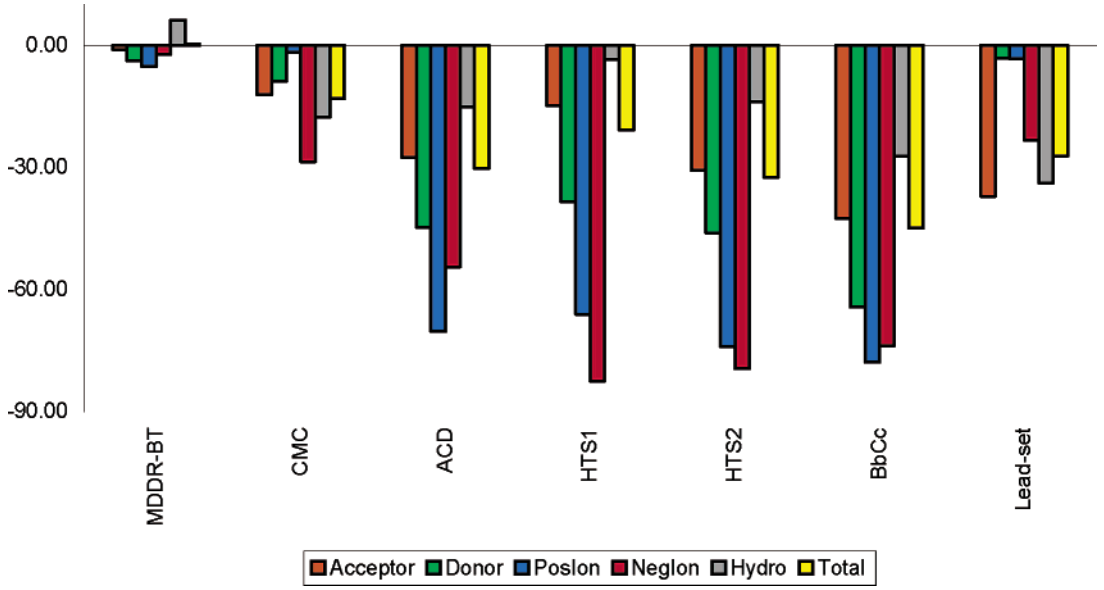


Figure 3. Percentage difference between the average of the pharmacophore points in MDDR-L (benchmark) and the other studied databases.

The distribution of the HbA groups observed for the HTS libraries resembled the one in ACD with even larger differences in the relative number of nitro and hydroxyl groups. As discussed in one of the following sections, another striking difference between the HbA distribution in the HTS libraries and MDDR-L was the number of carboxylic oxygens.

Hydrogen Bond Donors. The distributions obtained for the HbD varied across the databases. In ACD and BbCc, the number of HbD decreased across the range, whereas in the other database it peaked at 1 (Figure 4). The average number of structural moieties capable of donating a hydrogen bond decreased from 1.93 for a molecule in MDDR-L to 1.08 in ACD. Despite presenting a distribution similar to the drug-like space (peaking at 1) and with average values of 1.19 and 1.04 HbD respectively, HTS1 and HTS2 were clearly more closely related to the nondrug-like space.

Similar average HbD values were obtained for all drug-like databases (1.93 HbD in MDDR-L, 1.86 in MDDR-BT, and 1.76 in CMC). The difference in the number of compounds that did not have a single hydrogen bond donor was remarkably high across the databases. About 15% of all the compounds in the MDDR and 17% of the compounds in CMC had no HbD compared to 37% in ACD.

In all databases, the HbD landscape was clearly dominated by nitrogen donors. In this study the nitrogen donors were divided into three categories: aliphatic NHs, aromatic NHs, and protonable amines (tertiary amines that can become donors following protonation). Their combined value ranged from 63% in CMC to 72% in MDDR-L to top at 90% in HTS2. This variation was also reflected in the distribution of the hydroxyl groups which had its maximum in CMC (37%) and its minimum in HTS2 (9%). With values below the 0.5% threshold in both MDDR-L and CMC, the contribution of thiols to the functionalization of a drug molecule was clearly negligible.

Negative Centers. The distribution of the groups that can adopt a negative charge followed a similar trend in all databases: the vast majority of compounds did not have such groups and the number of molecules containing NI centers decreased very fast across the range with a negligible number of compounds having more than two. Despite this similarity, the absolute values of the distributions were significantly different. In the MDDR-L database, 75% of the compounds had zero NI with an average of 0.3 negatively ionizable groups per molecule whereas in ACD almost 90% of the compounds had zero NI with an average value of 0.14, more than 50% smaller than the one in MDDR-L. It was interesting to observe that this pharmacophore descriptor provided the greatest discrimination between the MDDR-L and CMC databases both representing drug-like chemical space (the number of hydrophobic sites was the other major difference between the two databases). In CMC the average NI value was 0.21, almost 30% less than in MDDR-L. Surprisingly the HTS collections had a very limited amount of negatively ionizable groups; about 95% of the compounds did not have a potential negative charge, a fraction even smaller than in ACD. The NI average value in HTS1 and HTS2 was 0.05 and 0.06, respectively, and it was almost exclusively associated with carboxylic acids. In the nondrug-like space the contribution of carboxylic acids to NI was also considerably high at 87%, whereas in the drug-like space it was possible to observe a greater diversity. It was interesting to note that hydroxamic acids and tetrazoles provided more than 5% of the overall NI groups in both MDDR-L and MDDR-BT, a fraction 10 times larger than in ACD.

Positive Centers. The positively ionizable centers were the pharmacophore feature that provided the greatest discrimination between drug-like and nondrug-like compounds (an average of 0.60 PI per molecule in MDDR-L and 0.18 in ACD) and, at the same time, showed the smallest difference in the drug-like class of compounds (0.57 PI in MDDR-L and 0.59 in CMC). It was possible to note two different patterns of distribution of the PI feature in each database and distribution following the same trend, also showed a remarkable similarity to one another. For MDDR-L, MDDR-BT, and CMC, the number of compounds containing PI pharmacophores tended to decrease in almost a linear fashion across the range with about 50−55% of the compounds not having a positively ionizable group. The other

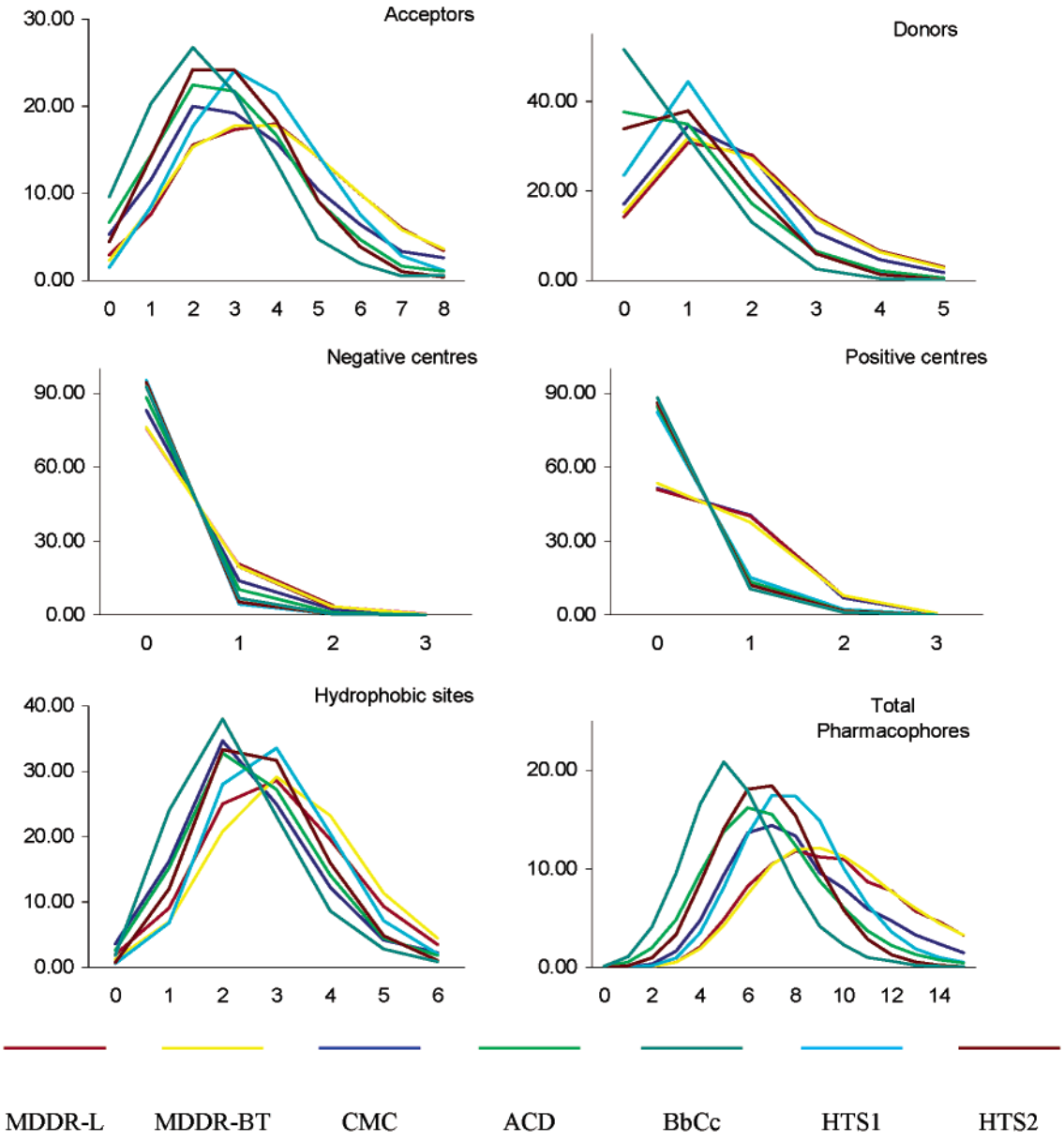


Figure 4. Distribution of the pharmacophore descriptors in the MDDR-L, MDDR-BT, CMC, ACD, HTS1, HTS2, and BbCc databases. The graphs show the percentage of molecules containing a particular number of a defined pharmacophore feature in each database.

databases were characterized by a larger number of compounds with no PI groups (more that 80% overall with 82% in HTS1 and 88% in BbCc). In the drug-like space, about 85% of the positively ionizable centers were represented by compounds containing protonable amines and 2/4 amino pyridine moieties. Amidine and guanidinium groups account for an additional 10%. A similar distribution was observed in the nondrug-like space, although there was a larger number of 2/4 amino pyridine-like groups and charged nitrogens.

Hydrophobic Sites. After the hydrogen bond acceptor, the hydrophobic sites were the second most common pharmacophore feature in small molecules. The average number of Hy ranged from 3.34 in MDDR-BT to 2.29 in BbCc. The number of hydrophobic sites in the biologically tested compounds in MDDR was slightly higher than the one in the launched compounds; this was the only instance where, for a given pharmacophore, the maximum value of the average occurrence did not correspond to the MDDR-L database. It was also the pharmacophore feature that was most similar in the drug and nondrug-like space with 3.14 hydrophobic site per molecule in MDDR-L and 2.71 in ACD. Hy always displayed a similar asymmetrical Gaussian distribution with maximum values at 2−3. Ranges of $1 \leq Hy \leq 3$ (CMC, ACD, and BbCc) or $2 \leq Hy \leq 4$ (MDDR-L, MDDR-BT, HTS1, and HTS2) described more than 70% of the data. Across all databases, five- and six-membered rings were the most abundant hydrophobic moieties. Combined, they accounted for more than 75% of the total hydrophobic groups with very similar values for both drug- and nondrug-like compounds (77.7% in MDDR-L compared to 79.3% in ACD). It was interesting to observe that the compounds in the HTS library were particularly rich in rings (90.2% in HTS1 and in HTS2) with the larger increase observed for the five-membered rings (21.8% and 40.0% more in the HTS1 and HTS2 respectively). The nondrug-like compounds differentiate from the drug-like ones primarily in the larger number of halogen atoms (6.5% in ACD compared to 3.0% in MDDR-L) and a smaller number of hydrophobic centers

Table 6. Distribution of Different Functionalities in the Examined Databases Classified by Pharmacophore Feature[a]

structural feature	MDDR-L	MDDR-BT	CMC	ACD	HTS1	HTS2	BbCc
			NI				
carboxylic acid	74.50	73.97 (−0.5)	74.72 (0.5)	88.24 (18.6)	93.28 (25.4)	92.34 (24.1)	92.16 (23.9)
tetrazole	5.03	5.81 (15.5)	2.21 (−56.1)	0.51 (−89.8)	0.65 (−87.1)	1.73 (−65.6)	1.20 (−76.2)
acid sulfona(i)mide	3.19	3.12 (−2.1)	4.27 (34.1)	1.05 (−67.1)	2.69 (−15.6)	4.24 (32.9)	0.11 (−96.6)
S/P acids	11.83	10.87 (−8.0)	17.09 (44.7)	9.66(−18.2)	2.93 (−75.2)	1.19 (−89.9)	5.12 (−56.6)
hydroxamic acid	5.44	6.23 (14.6)	1.71 (−68.5)	0.54 (−90.0)	0.45 (−91.7)	0.51 (−90.7)	1.42 (73.9)
			PI				
amidine	5.22	5.75 (10.4)	3.69 (−29.1)	6.00 (15.3)	3.13 (−39.9)	9.78 (87.9)	5.64 (8.7)
protonable amine	74.08	72.54 (−1.8)	78.67 (6.5)	60.15 (−18.6)	73.43 (−0.6)	51.85 (−29.8)	42.40 (−42.4)
guanidinium	3.53	3.58 (1.5)	3.02 (−14.2)	3.06 (−13.3)	1.56 (−55.9)	3.17 (−10.0)	2.92 (−17.0)
2/4-amino pyridine	13.16	14.54 (8.6)	7.38 (−44.9)	20.03 (49.6)	16.46 (22.9)	35.56 (150.6)	20.81 (54.3)
charged N	4.00	3.60 (−9.8)	7.32 (81.1)	10.75 (169.4)	5.43 (35.9)	1.63 (−59.2)	28.14 (607.2)
			HbD				
nitrogen donor (aliphatic)	55.13	57.85 (5.0)	43.08 (−21.8)	67.80 (23.0)	73.63 (33.6)	77.60 (40.8)	67.95 (23.0)
oxygen donor	28.02	24.23 (−13.2)	36.76 (31.7)	19.45 (−30.3)	12.22 (−56.2)	8.87 (−68.2)	18.87 (−32.5)
nitrogen donor (aromatic)	2.06	2.72 (32.1)	1.43 (−30.7)	5.14 (149.4)	3.66 (77.5)	5.88 (185.2)	5.09 (155.5)
thiol	0.36	0.45 (23.8)	0.14 (−62.4)	0.94 (158.1)	0.27 (−24.4)	0.78 (115.9)	2.11 (480.4)
protonable amine	14.44	14.74 (2.5)	18.59 (−21.8)	6.67 (−53.6)	10.22 (33.6)	6.87 (40.8)	5.97 (23.0)
			HbA				
carboxylic oxygens	10.60	10.38 (−2.1)	8.62 (−18.7)	7.82 (−26.9)	2.73 (−74.3)	3.90 (−63.2)	5.95 (−43.9)
carbonyl	27.74	29.40 (6.0)	27.98 (0.9)	29.92 (7.8)	34.32 (23.7)	29.45 (6.2)	28.02 (1.0)
S/P oxygen	8.53	8.21 (−3.8)	7.21 (−15.5)	7.94 (−6.9)	7.45 (−12.7)	10.28 (20.6)	8.62 (1.0)
hydroxyl	11.96	9.98 (−16.6)	16.93 (41.6)	6.16 (−48.5)	3.31 (−72.4)	2.53 (−78.9)	4.41 (−63.2)
ether	15.50	5.79 (−62.6)	15.90 (2.6)	13.18 (−15.0)	16.28 (5.0)	11.9 (−23.1)	11.1 (−28.1)
ester	5.74	15.54 (170.6)	7.81 (35.9)	8.02 (39.6)	5.50 (−4.3)	6.37 (10.9)	7.30 (27.2)
other O	0.92	0.79 (−14.0)	1.21 (31.7)	0.90 (−2.3)	1.08 (17.1)	0.64 (−30.5)	0.47 (−48.4)
nitro O	1.78	1.37 (−22.9)	1.87 (5.0)	6.39 (259.4)	8.58 (383.2)	6.76 (280.4)	12.15(584.0)
nitrogen acceptors	17.22	18.54 (7.6)	12.47 (−27.6)	19.69 (14.3)	20.76 (20.5)	28.12 (63.3)	21.93 (27.3)
			Hy				
six-membered rings	62.36	62.90 (0.9)	64.48 (3.4)	63.96 (2.6)	71.61 (14.8)	64.47 (3.4)	58.48 (−6.2)
five-membered rings	15.30	16.48 (7.7)	12.88 (−15.8)	15.35 (0.3)	18.64 (21.8)	21.42 (40.0)	15.34 (0.3)
tert-butyl	1.67	1.46 (−12.6)	1.18 (−29.3)	1.72 (3.1)	0.76 (−54.5)	2.03 (21.6)	1.79 (7.2)
hydrophobic centers	17.65	16.55 (−6.2)	17.68 (0.2)	12.47 (−29.4)	6.44 (−63.5)	3.39 (−80.8)	5.72 (−67.6)
halogens	3.02	2.61 (−13.6)	3.78 (25.2)	6.50 (115.4)	2.55 (−15.6)	8.70 (188.1)	18.67 (518.2)

[a] The occurrence of each structural feature is represented as a percentage of a particular pharmacophore feature. In parentheses the percentage difference between the value of the pharmacophore occurrence observed for a given database and the one observed for MDDR-L.

(12.5% in ACD, 17.6% in MDDR-L). The *tert*-butyl group, probably due to its bulkiness, was found to be the least likely hydrophobic feature in all databases.

COX Pharmacophore Profile. The pharmacophore profile obtained for the COX system showed that a typical COX inhibitor had a limited number of pharmacophore features (an average number of 7.52 compared to 10.15 in MDDR-L) and was predominantly described by the pharmacophore features Hy and HbA (an average of 3.32 and 3.00 per molecule, respectively) which, when combined, accounted for 84% of the total pharmacophore points in the molecule (compared to 74% for a molecule in MDDR-L). HbD contributed an additional 13% (19% in MDDR-L) of the features, while compounds in this set had a negligible number of ionizable groups (the average NI and PI was 0.10 and 0.12, respectively, less than 3% combined, compared to about 9% in MDDR-L). The calculated profile is consistent with the prevailing hydrophobic nature of the COX binding sites.

DISCUSSION

About the Lack of NI and PI in HTS Collections of Compounds. The number of negatively ionizable centers combined with the number of positively ionizable centers which can be expected to be present in the average launched compound in MDDR is 0.90. Similar values were obtained for the average molecule in both the MDDR-BT (0.86) and CMC (0.70). Moving to the nondrug-like space, the average number of charged functional groups dropped dramatically: 0.32 in ACD. Surprisingly, the values obtained for the compounds in the commercially available HTS collections were even more distant from the drug-like space: 0.25 for HTS1 and 0.22 for HTS2. Overall, about 65% of the molecules in MDDR-L contained at least one or more PI or NI groups. Similarly, 60% of CMC compounds had chargeable groups, compared to the much lower proportion of ACD (25%) and HTS2 (19%) compounds

Could this be because HTS collections contain lead-like molecules rather than drug molecules? In a recent study, Opera and co-workers discussed the differences between lead structures and their related developed molecules using different molecular properties relevant to the drug-like chemical space.[15] An analysis of the 62 lead structures presented in their work using our pharmacophore descriptors led to the following 'lead-like profile': PI 0.58, NI 0.23, HbA 2.63, HbD 1.87, Hy 2.08 (Figure 3).

Although this is by no means an exhaustive analysis of the lead-like space, it suggests that from a pharmacophore point of view, the lead molecular space greatly overlaps with the drug space. In particular, it is interesting to note that for PI and HbD, the pharmacophore requirements of the drug molecules are mainly satisfied already in the lead-like space. In fact, the average number of pharmacophore features in the lead space related well to the drug space (in the drug

profile PI 0.60 and HbD 1.93). On the contrary, the lead molecular space seemed to be less rich in negatively ionizable centers, hydrogen bond acceptors, and hydrophobic features (in the drug profile NI, HbA, and Hy were 0.30, 4.18, and 3.14, respectively). This confirms that the number of HbD and PI in a drug molecule seems to be more tightly regulated than HbA, Hy, and NI and is consistent with the view of lead molecules as smaller, less functionalized and hydrophobic entities compared with the corresponding drug molecules.[5,37] It is worth remembering that no attempt was made to determine the ionization state of a molecule by rigorous pK_a calculation. Instead, the number of PI and NI was obtained by simply adding together the occurrences of specific ionizable groups in a molecule. Likely, this led to a slight overestimation in the number of ionizable pharmacophore features. Even so, the average number of ionizable groups in a drug molecule appeared to be higher than the common medicinal chemist's perception of good drug candidates would suggest. The discrepancy between the observed and 'expected' number of ionizable groups in a drug molecule probably reflects the current focus on orally available compounds which tends to obscure the fact that many molecules of pharmacological significance are not administrated orally or rely on protein transport rather than passive diffusion to cross the cell membrane.[38]

About the (Un)success of Combinatorial Chemistry. The urgent need for new lead molecules provided a powerful driving force for the development of HTS technology. In parallel, the pharmaceutical industry made significant investments in Combinatorial Chemistry to provide more and more compounds for screening. The initial goal to generate large and diverse libraries of compounds was successfully achieved, and the employment of large HTS programs became a common feature in the lead identification process. Unfortunately, this did not yield any real improvement in the generation of new lead series.[39] As suggested by Feher and Schmidt, there is increasing evidence that the low hit rate observed during HTS may in part reflect significant deficiencies in the type of structures that are actually screened and, ultimately, in the compounds that are generated using combinatorial approaches.[40] The results of this study undoubtedly show that the pharmacophore distributions in the compounds of both the HTS collections strongly resembled those obtained for the nondrug-like compounds (Table 4 and Figure 3). The distribution of the different structural features is also nondrug-like (Table 5). The limited chemistry available to CC might account for the observation that HTS compounds are less functionalized than drug molecules. Problematic and possibly costly purifications could also be a basis for the lack of charged groups in the HTS sets discussed in the previous section. For instance, from a chemistry point of view, the preparation and purification of an ester is much easier and less elaborate than the preparation of the corresponding acid.

Despite the work that has already been done in this direction, major emphasis should be placed in generating diverse libraries of compounds that also display drug-like physicochemical properties. Pharmacophore profiles similar to the one presented in this study, in conjunction with other property distribution analyses,[5] should provide valuable tools in the optimization of the diversity and drug-likeness of the compounds to prepare and screen. Libraries focused on a particular biological target could be improved in a similar way.

About the Differences between MDDR-L and CMC. In the recent literature, the MDDR and CMC databases (together with the World Drug Index) have interchangeably been used as benchmarks to define the drug-like space.[11−13] In other cases they have been used in the validation of predictive computational tools, one as training set and the other as control set.[9]

It was interesting to note that the pharmacophore profile obtained for CMC was considerably different from the ones obtained for MDDR-L and MDDR-BT, having a partial nondrug-like character. There might be several reasons for this. Both MDDR and CMC databases are collections of compounds that include not only drug molecules but also pharmacological agents and compounds with a proven biological activity that were not developed into drugs. In MDDR, the progress through the different development phases is recorded for each compound, so it was possible to focus on the component of MDDR that includes molecules already launched as drugs or in a late stage of development (MDDR-L) and to separate it from the other compounds (MDDR-BT). However, this division alone is not sufficient to explain the differences of the pharmacophore distributions in each database. In fact, the profile obtained for MDDR-L is remarkably similar to the MDDT-BT profile. A problem associated with large databases such as the ones examined in this work is that they grow over time and not always as if sampled from a static probability distribution, i.e., the probability distribution can vary dramatically. Lipinski, for instance, has shown that there has been a shift to higher molecular weight for the compounds in clinical trial over the past few years.[8] This suggests that the concept of drug, hence the characteristics of the drug molecules, is not static but evolves over time. Several factors are probably contributing to the evolution of the drug molecules, for example they often undergo further development to increase their effectiveness, novel drugs are developed for new targets or for disease without a suitable cure and, in the recent past, more and more emphasis has been placed on orally absorbable compounds. With the decoding of the human genome offering the opportunity to identify new targets and with the numerous diseases still without adequate chemotherapy, the exploration of the drug-like space will inevitably continue in the future and the characteristics of drug molecules are likely to change with it.

Although they are both focused on the drug-like space, the CMC and MDDR databases cover different periods of time. MDDR is focused on the work produced in the last 25 years, whereas the CMC goes back to the beginning of the 20th century. Hence MDDR-L provides a pharmacophore profile based on the most recent perception of the drug-like space, whereas CMC is likely to provide one based on the evolution of the drug-like space, or at least our perception of it, over time. These intrinsic differences between MDDR and CMC should be kept in consideration when analyzing the properties of the compounds in these databases. Due to the certainty about the development phase of its compounds, the MDDR-L was chosen as the representative of the drug-like space in this study.

About the Possibility of Classifying Drug-like and Nondrug-like. The possibility to discriminate between drug-

like and nondrug-like compounds by analyzing the distributions of a limited set of molecular properties would be an invaluable tool in facilitating the lead identification process. Despite much effort, to date it has not been possible to identify any single descriptor or combination of descriptors that capture the difference between drug and nondrug molecules in a simple way. For instance, it has been shown that about 80% of the molecules in the nondrug-like space pass Lipinski's 'rule of five'.[5] This inability to predict drug-likeness is due to the lack of significant differences in the distribution of the descriptors used by Lipinski. Similarly, the drug-like space defined by the pharmacophore features partially overlap with the nondrug like space. Hence, pharmacophore descriptors are not suitable to separate drug-like from nondrug-like.

CONCLUSIONS

To further characterize the structural properties required to a drug molecule, a set of sub(structural) elements, grouped in five pharmacophore features, were used to analyze different sets of molecules of interest in drug discovery: drug molecules, compounds in high throughput screening libraries, combinatorial chemistry building blocks, and nondrug molecules. It was decided to use pharmacophore features as they are suited to capturing the essence of the molecular functions required in the binding event and because they are a familiar concept to medicinal chemists. The analysis of the pharmacophore feature distribution in the different sets of compounds highlighted several points of interest. It was observed that the pharmacophore composition of drug-like molecules was the most complex and diverse among the classes of compounds investigated and that there are intrinsic differences between two of the databases often used to represent the drug-like space, MDDR and CMC. The distributions of the pharmacophore features that were used in this study, provided pharmacophore profiles that could successfully be used in property-based design to define boundaries identifying regions of the chemical space rich in drug-like or nondrug-like molecules. This could lead to a more effective library design or compound selection for biological screening. It was shown, in fact, that compounds currently available in HTS collections did not share the same pharmacophore space of the drug-like compounds. The drug-like pharmacophore profile was derived from a set of compounds showing a broad range of known activities, and hence it defines a 'general drug-like space' from a probabilistic point of view.[41] Pharmacore profiles derived for specific biological targets, such as the COX pharmacophore profile exemplifies, are likely to be significantly different from one another and from the general drug-like profile reflecting the different functional requirements of each targeted binding site. Target specific pharmacophore profiles can assist in the optimization of focused libraries.

ACKNOWLEDGMENT

I would like to thank Dr. Andrew J. Chalk, Dr. Edith A. W. Chan, Dr. Scott A. Dann, Dr. John P. Overington, and all my colleagues at Inpharmatica Ltd. for valuable discussions.

Supporting Information Available: Pharmacophore definitions and SLN used to define the pharmacophore points used in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960−1964.
(2) Hopkins, L. A.; Groom, R. C. The druggable genome. *Nature Reviews − Drug Discovery* **2002**, *1*, 727−730.
(3) Edwards, R. A.; Zhang, K.; Firth, L. Benchmarking Chemistry Functions within Pharmaceutical Drug Discovery and Preclinical Development. *Drug Discovery World* **2002**, *3* (3), 67−74.
(4) Warters, P. W.; Matthew, S. T.; Murcko, M. A. Virtual Screening − an Overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(5) Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(6) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(7) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(9) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(10) Gillet, V. J.; Willet, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(11) Muegge, I.; Heald, L. S.; Brittelli, D. Simple Selection Criteria for Drug-like Chemical Matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(12) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(13) Xu, J.; Stevenson, J. Drug-like Index: A New Approach To Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.
(14) Van de Waterbeemd, H.; Smith, A. D.; Beamont, K.; Walker, K. D. Property-Based Design: Optimization of Drug Absorption and Pharmacokinetics. *J. Med. Chem.* **2001**, *44*, 1313−1333.
(15) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(16) MACCS-II Drug Data Report (2002.1 version) contains biologically active compounds in the different stages of drug development as presented in the patent literature, journals, meetings and congresses. The database is available from MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA, 94577. Website: www.mdli.com/products/mddr.html.
(17) Comprehensive Medicinal Chemistry (2002.1 version) contains compounds used or studied as medicinal agents in humans and pharmaceutical compounds. It is derived from the Drug Compendium in the Pergamon's Comprehensive Medicinal Chemistry. The database is available from MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA, 94577. Website: http://www.mdli.com/products/cmc.html.
(18) Available Chemical Directory (2002.1 version) contains grade and bulk chemicals. The database is available from MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA, 94577. Website: http://www.mdli.com/products/acd.html.
(19) Oprea (ref 5) concluded that the subset of the reactive compounds in ACD and MDDR have the same property distribution than the remaining set and that their removal from the dataset did not influence the overall property distribution. In this study it was assumed that the pharmacophore distribution is also similar in the each set of compounds.
(20) Compounds that felt into any of the following classes were removed: **MDDR**- anesthetic, pulmonary surfactant, contrast medium, sweetener, dental agent, ultraviolet light absorber, radiosensitizer, radioprotector, radiopharmaceutical for brain imaging, imaging agent, pharmaceutical aid, keratolytic, dermatological aid, anti acne, spermicide, vaccine, photosintesizer, antidote, blood substitute, chemoprotective, wound healing agent, MoAb, lubricant agent. **CMC**- buffer, solvent, anesthetic, disinfectant, aerosol propellant, surfactant, preservative, surgical aid, radiopaque medium, X-ray contrast agent, insecticide, astringent, laxative, sweetener, dental caries prophylactic, ultraviolet screen, flavoring agents, chelating agents, radioprotective agent, sequestering agent, imagining agent, repartitioning agent, alkalizing agent, MRI agent, retinal oxygenizing agent, pharmaceutical aid, diagnostic aid, keratolytic agent, veterinary agent, scabicide, ectoparasiticide, nutrient, radiosensitizer, antiseptic, pigment, depigmentor, dermatological aid, antiacne, acaricide, sacaricide, parasiticide, vaccine, photosensitizer, alcohol denaturant, antidote, prosthetic aid, blood substitute, replenister, emulsifier, emollient, detoxicant, antiperspirant, iodine source, emulsifier.

(21) Willet, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(22) UNITY 4.3, Tripos Inc., 1699 South Hanley Rd., St. Louis, MO 63144, U.S.A.
(23) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behaviour: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(24) Website: http://www.asinex.com/.
(25) Website: http://www.maybridge.com/.
(26) Compounds with Cox activity were identified using an in house literature database covering the last 10 years of the *Journal of Medicinal Chemistry* publication. Compounds displaying either COX-1 or COX-2 activity were included in the set.
(27) Davis, A. M.; Teague, S. J. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew. Chem., Int. Ed.* Engl. **1999**, *38*, 736−749.
(28) **SLN**, a notation developed by Tripos Inc., 1699 South Hanley Rd., St. Louis, MO 63144, U.S.A. was inspired by the SMILE annotation and has several extensions and modifications allowing specification of full substructure queries, including Markush structures. SLN provides access to the fragment recognition routines from within a procedural language such as SYBYL Programming Language (**SPL**).
(29) The SLN definitions used in this work, a modification of the original set provided by Tripos with Sybyl6.8, are available in Appendix 1. It was necessary to modify the default file to introduce the negatively ionizable and positively ionizable groups and to optimize the hydrogen bond donor/acceptor and hydrophobic definition.
(30) Sybyl6.8.1, Tripos Inc., 1699 South Hanley Rd., St. Louis, MO 63144, U.S.A.
(31) The SLN definitions were imported in Sybyl6.8 as sln3dmacro.def file.
(32) Abraham, H. M.; Duce, P. P.; Prior, V. D. Hydrogen Bonding. Part 9. Solute Proton Donor and Proton Acceptor Scales for Use in Drug Design. *J. Chem. Soc., Perkin Trans. 2* **1989**, 1355−1375.
(33) Five- and six-membered rings containing heteroatoms were classified as hydrophobic moieties providing that the heteroatom was not already part of any other functionality.
(34) Principle introduced by Vilfredo Pareto in 1897. In this context the principle can be formulated as follows: a minority of causes usually leads to a majority of results.
(35) Koch, R. *The 80/20 Principle*; Nicholas Brealy Publisher: London, 1997.
(36) Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*; Chapman & Hall/CRC Press: Boca Raton, FL, 2000.
(37) Hann, M. M.; Leach, R. A.; Harper, G. Molecular Complexity and Its mpact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.
(38) Al-Awqati, Q. One hundred years of membrane permeability: does Overton still rule? *Nature Cell Biol.* **1999**, *8*, E201−E202.
(39) Leach, R. A.; Hann, M. M. The *in silico* world of virtual libraries. *Drug Discovery Today* **2000**, *5*, 326−336.
(40) Feher, M.; Schmidt, M. J. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.
(41) The 8251 compounds in the MDDR-L displayed a total of 16 920 biological activities grouped in 525 unique indications.

CI034068K