A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases

Arup K. Ghose,* Vellarkad N. Viswanadhan,* and John J. Wendoloski

Amgen Inc., One Amgen Center Drive, Thousand Oaks, California 91320

Received August 7, 1998

The discovery of various protein/receptor targets from genomic research is expanding rapidly. Along with the automation of organic synthesis and biochemical screening, this is bringing a major change in the whole field of drug discovery research. In the traditional drug discovery process, the industry tests compounds in the thousands. With automated synthesis, the number of compounds to be tested could be in the millions. This two-dimensional expansion will lead to a major demand for resources, unless the chemical libraries are made wisely. The objective of this work is to provide both quantitative and qualitative characterization of known drugs which will help to generate "drug-like" libraries. In this work we analyzed the Comprehensive Medicinal Chemistry (CMC) database and seven different subsets belonging to different classes of drug molecules. These include some central nervous system active drugs and cardiovascular, cancer, inflammation, and infection disease states. A quantitative characterization based on computed physicochemical property profiles such as log *P*, molar refractivity, molecular weight, and number of atoms as well as a qualitative characterization based on the occurrence of functional groups and important substructures are developed here. For the CMC database, the *qualifying range* (covering more than 80% of the compounds) of the calculated log *P* is between −0.4 and 5.6, with an average value of 2.52. For molecular weight, the qualifying range is between 160 and 480, with an average value of 357. For molar refractivity, the qualifying range is between 40 and 130, with an average value of 97. For the total number of atoms, the qualifying range is between 20 and 70, with an average value of 48. Benzene is by far the most abundant substructure in this drug database, slightly more abundant than all the heterocyclic rings combined. Nonaromatic heterocyclic rings are twice as abundant as the aromatic heterocycles. Tertiary aliphatic amines, alcoholic OH and carboxamides are the most abundant functional groups in the drug database. The effective range of physicochemical properties presented here can be used in the design of drug-like combinatorial libraries as well as in developing a more efficient corporate medicinal chemistry library.

1. Introduction

It is widely believed[1,2] that the pharmaceutical and biotechnology industry will be one of the most active industrial fields in the next century because of the information explosion in the field of genomics. The number of target proteins that can yield important therapeutic agents is expected to increase dramatically in the near future. The pharmaceutical drug discovery research is currently undergoing a tremendous change due to automated parallel organic synthesis[3,4] (combinatorial chemistry) and high-throughput biochemical screening.[4] However, it is necessary to avoid the pitfall of combinatorial explosion because, although the cost of high-throughput screening or automated synthesis per compound may be low, it will become fairly expensive when multiplied by millions. There are several ways to decrease the cost to a manageable level: (i) understanding the target protein structure and designing focused compounds or libraries that fit the protein binding site;[5−11] (ii) designing compounds or libraries around "hits" that are often identified from the initial screening of the existing corporate libraries, using pharmacophore modeling and three-dimensional quantitative structure−activity relationship (3D-QSAR) methods;[12−18] (iii) developing ultrasensitive high-throughput screening (HTS).[4,19] The design process can be further streamlined by focusing on "drug-like" molecules. A convenient starting point to develop a consensus definition(s) of a drug-like molecule is to analyze the databases of known pharmaceutical agents. As a first step, it is necessary to identify biologically and pharmacologically relevant properties which are easily computable from the structure. In this context, it will be instructive to analyze the physicochemical, topological, and electronic properties of all known drugs and compare the properties of different classes of drugs. It will be easy to formulate a consensus definition if the drug molecules are clustered in one or more property spaces. An earlier analysis[20] of known drugs pertained to molecular frameworks and employed shape description methods to prepare a list of common drug shapes. Another analysis[21] of known drugs (the Comprehensive Medicinal Chemistry (CMC) database) and other databases such as the Available Chemicals Directory (ACD) has been devoted to identify the criteria to use in the selection of compounds for screening. This study

10.1021/cc9800071 CCC: $18.00
Published on Web 12/18/1998

Table 1. Classes of Drugs Identified from the CMC Database Based on Keyword Searches and Their Average Physicochemical Properties[a]

database no.	disease type	keyword	no. of drugs	ALOGP	AMR	MW	no. of atoms
1	CMC (clean)		6304	2.30 (2.6)	96.7 (45.3)	357 (174)	48.4 (25.0)
2	inflammation	antiinflammatory	290	3.09 (1.5)	89.2 (31.0)	335 (122)	43.3 (19.0)
3	CNS	antidepressant	208	3.05 (1.5)	85.8 (19.3)	291 (69)	42.0 (9.7)
4	CNS	antipsychotic	105	4.10 (1.5)	108.0 (22.8)	380 (83)	51.7 (12.2)
5	cardiovascular	antihypertensive	269	1.97 (2.1)	97.7 (33.5)	361 (123)	48.3 (18.2)
6	CNS	hypnotic	74	2.20 (1.5)	70.2 (25.3)	277 (99)	33.6 (11.5)
7	cancer	antineoplastic	349	1.59 (2.5)	87.6 (36.5)	332 (129)	44.1 (19.5)
8	infection	antiinfective	39	2.38 (2.7)	89.0 (42.9)	339 (139)	41.7 (25.2)

[a] Mean and standard deviation (in parentheses) are listed for (a) calculated log *P* (ALOGP), (b) calculated molar refractivity (AMR), (c) molecular weight (MW), and (d) number of atoms.

showed that the MDL keys provide at least one way to eliminate compounds least likely to satisfy "drug-likeness". Lipinski et al.[22] studied the USAN (United States Adopted Names) compound list in terms of the computed lipophilicity by Moriguchi method[23] and gave a set of rules ("the rule of 5"). According to "the rule of 5", a poor permeation or absorption is more likely when there are more than 5 H-bond donors, 10 H-bond acceptors, the molecular weight is greater than 500, and the calculated log *P* (Clog *P*) is greater than 5 (or Moriguchi log $P > 4.15$); compound classes that are substrates for biological transporters are exceptions to the rule. Hydrogen bond donors, acceptors, and molecular weight are easily computed for any library, real or virtual, but calculating lipophilicity accurately entails choosing a well-tested, commercially available method that is easy to use and is generally applicable to all classes of organic molecules of medicinal interest. A recent study[24] shows that the ALOGP method[24−26] satisfies these criteria and is therefore likely to be more useful. The main objective of the present work is to profile some pharmacologically relevant physicochemical properties including log *P* (using the ALOGP[24−26] method) and pharmacophorically relevant chemical functionalities of some important classes of drugs along with the whole CMC data set, in order to empirically define a drug-like molecule. This, in turn, will help to design the drug-like combinatorial libraries and to develop guidelines for prioritizing large sets of compounds for biological testing.

2. Materials and Methods

(a) Molecular Databases. There are several commercially available drug-related databases, e.g., Chapman and Hall's Dictionary of Pharmacological Agents[27] which contains over 30 000 compounds of pharmacological interest, the MDL Drug Data Report (MDDR) which has the structures and biological activity data for over 85 000 compounds,[28] and the CMC database (version 97.1) which contains the structures and biological properties of 7183 compounds.[28] We used here the CMC database, since the major source for compounds in this compendium was the drugs identified by either USAN[29] or INN (International Nonproprietary Names)[29] generic names. These names include practically all medicinal agents or compounds intended for clinical study in the advanced countries. In other words, the CMC database is by far the closest database of drug-like molecules. The other databases contain a large fraction of early discovery stage compounds which may not be drug-like. The current CMC version (v. 97.1) contains 7183 structures. However, it has several classes of compounds such as diagnostic imaging agents, solvents, and pharmaceutical aids that are important in the pharmaceutical industry but not necessarily drug-like. The analyzed CMC database was therefore cleaned of these agents (see also ref 20). The search expression for the removed agents was built from ISIS "query builder" and had the syntax

MOL > CLASS > class like "%*keyword*%"

where the *keyword* was one of the following: radiopaque, contrast, disinfectant, spermicide, wetting, flavoring, pharmaceutical aid, surgical aid, dental, surfactant, sunscreen, ultraviolet screen, preservative, aerosol, chelator, insecticide, astrigent, herbicide, solvent, laxative, sweetner, adhesive, dentistry, veterinary, buffer, scabicide. In addition we screened compounds with elements X (a symbolic representation of a resin), Li, Be, and various transition elements as well as a few compounds with radioactive elements. A few drugs with a mixture of more than one compound such as haloquinol were also removed. All this screening dropped the number of "acceptable" compounds to 6454. The substructure search was done with this data set. However, during the physicochemical property calculation a few compounds were deleted because of complex structures or unavailable parameters. This database had 6304 compounds.

In addition to the whole CMC database, we analyzed several drug classes such as central nervous system (CNS), cardiovascular, cancer, inflammation, and infectious diseases. These types and the specific keywords used in these searches are shown in Table 1.

(b) Molecular Physicochemical Properties. The selection of physicochemical properties in the profiling may need some discussion. Experimentally determined values are not directly useful in the design process since we need the properties before the compounds are made. The best choice may be the experimentally measurable (pharmaceutically and biologically relevant) properties that can be computed reliably. Calculated log *P*, molar refractivity, number of hydrogen bond donors and acceptors, molecular size (number of atoms and molecular weight), and molar refractivity are some examples of the properties that satisfy the criterion. Mathematical properties such as topological indices and functional characteristics such as substructure counts may also be interesting to study. The quantum chemical properties such as highest occupied molecular orbital and lowest unoccupied molecular orbital[30] often are conformation dependent and are

difficult to study for large data sets. In the current study we included molecular weight, number of atoms, calculated log *P* using the ALOGP[24,31] method, and molar refractivity using the AMR[25,31,32] method. The latter property is related to the volume of the molecule and its molecular weight.

(c) Computational Steps for Physicochemical Profiling and Comparison of Drug Classes. The various computational steps in this analysis are summarized below: (i) once the clean CMC and other drug class lists were created, structures were exported from ISIS as SD files; (ii) the SD files were then converted into Galaxy databases for the calculation of relevant properties, taking care to correct the representation of functionalities such as nitro or *N*-oxides which are represented differently in ISIS[28] and Galaxy;[24] (ii) most of the simple hydrogen halide salts were converted to the corresponding neutral system for a better computation of the physicochemical properties; (iv) analysis of the physicochemical properties (range and percentile distribution) was performed using the "Database Analysis" module of the Galaxy software. We determined two different ranges for the physicochemical properties: the *qualifying range* which covers approximately 80% of the drugs studied and the *preferred range* which covers approximately 50% of the drugs. Having two ranges may be useful as the distribution, being γ type,[33] may need a considerably larger range to cover 80% of known drugs, whereas the search/design for new drugs may be more efficient if we simply consider compounds having a considerably shorter property range (preferred range!) which has a high density of occupation within the qualifying range. On the basis of a careful analysis of the property histograms showing the ranges occupied by different percentages of drugs in each drug class and in the CMC database, the following definition of the preferred range appeared satisfactory: the preferred range is the smallest range within the qualifying range occupied by approximately 50% of the drugs. It is thus necessary to determine both the interval and location of these ranges. The interval of the range was determined by starting from (mean − standard deviation) to (mean + standard deviation). The range was expanded symmetrically on either side of the mean if it did not cover approximately 80% (or 50%) of the drug and contracted if it was considerably higher. Once the interval of the range was determined, it was shifted on either side until the most densely populated area was obtained.

(d) Analysis of the Chemical Functionalities and Important Substructures. We determined the frequency of occurrence of common organic functional groups and aromatic ring systems and a few interesting structural moieties in the CMC database and in a few different drug classes. The substructure search was done using the ISIS software.[28] The search queries sometimes were a simple structure in a substructure search; sometimes they were a complex query with the presence or absence of several substructures. These organic functional groups, ring systems, and other structural moieties are shown in Figure 1.

3. Results and Discussion

The objectives of this work are (i) to develop a consensus definition of a drug-like molecule, physicochemically and structurally; (ii) to compare property distributions among different drug classes and the complete database and analyze the differences; (iii) to develop a practical strategy for designing combinatorial libraries or a corporate medicinal chemistry compound library.

Physicochemical Properties. Molecular lipophilicity and molar refractivity of drug molecules are long known to be important features strongly influencing receptor binding, cellular uptake, and bioavailability. As fragmental constants, they are used to represent the hydrophobic and dispersive (van der Waals) interactions,[32] respectively. These properties are used in QSAR[34,35] and 3D-QSAR[12,13,36,37] studies. Thus the range and distribution of these properties can be used to fingerprint or characterize a library or drug class. It must, however, be cautioned that this characterization pertains to the overall property of the molecule and not to the distribution of the property within the molecule. Nevertheless, it may be regarded as a useful filter, and hence it may be used to develop a consensus definition of drug-like character. The mean values and the corresponding standard deviations of these properties are shown in Table 1. The frequency distributions of these properties are shown in Figures 2−5.

Table 1 shows the average values calculated for log *P* (ALOGP[24−26]) and molar refractivity (AMR[25,31,32]) using the atomic constant approach. The average ALOGP[24−26] value of the CMC database is 2.3 with a standard deviation of 2.6. The qualifying and preferred ranges for the whole CMC database as well as the seven drug classes are shown in Table 2. The qualifying range for the CMC database is −0.4 to 5.6 (see Figure 2 and Table 2). The corresponding preferred range is 1.3 to 4.1. It may be interesting to analyze the compounds which are well beyond this range, for example, the very hydrophilic drugs whose ALOGP[24−26] values were less than −5.0 are shown in Table 3. These compounds are mostly polyhydroxy polyamine antibacterial compounds, unblocked (zwitterionic) peptides, and quaternary ammonium salts. Unlike the antibacterial compounds, the peptides and quaternary salts have several hydrophobic cores in their structures. The antibacterial compounds are definitely a special class of biologically active compounds, which are very different from the regular drugs. These findings clearly show that unless one is interested in a special class of drugs such as antibacterial, the chance of success is at least 1 order of magnitude higher if we keep log *P* of the compound within −0.4 to 5.6. The analysis of the drugs whose ALOGP[24−26] values were greater than 7.0 (Table 4) did not show any predominant class, although quite a few steroids were in this list. Most of these compounds had a relatively high molecular weight (>500). Some of these compounds had a very hydrophobic hydrolyzable group. It is possible that some of these compounds were prodrugs,[38] and the hydrophobic group helped to cross the cell membrane, blood brain barrier, or to enhance its chemical stability, etc. Some of these outliers may resemble some naturally occurring compounds of the body and may have an active transport mechanism over passive transport.[39] Both of these lists showed some compounds that did not have any "drug class" in the CMC database. These compounds were deleted from the analysis.

It is seen that CNS drugs differed considerably in their

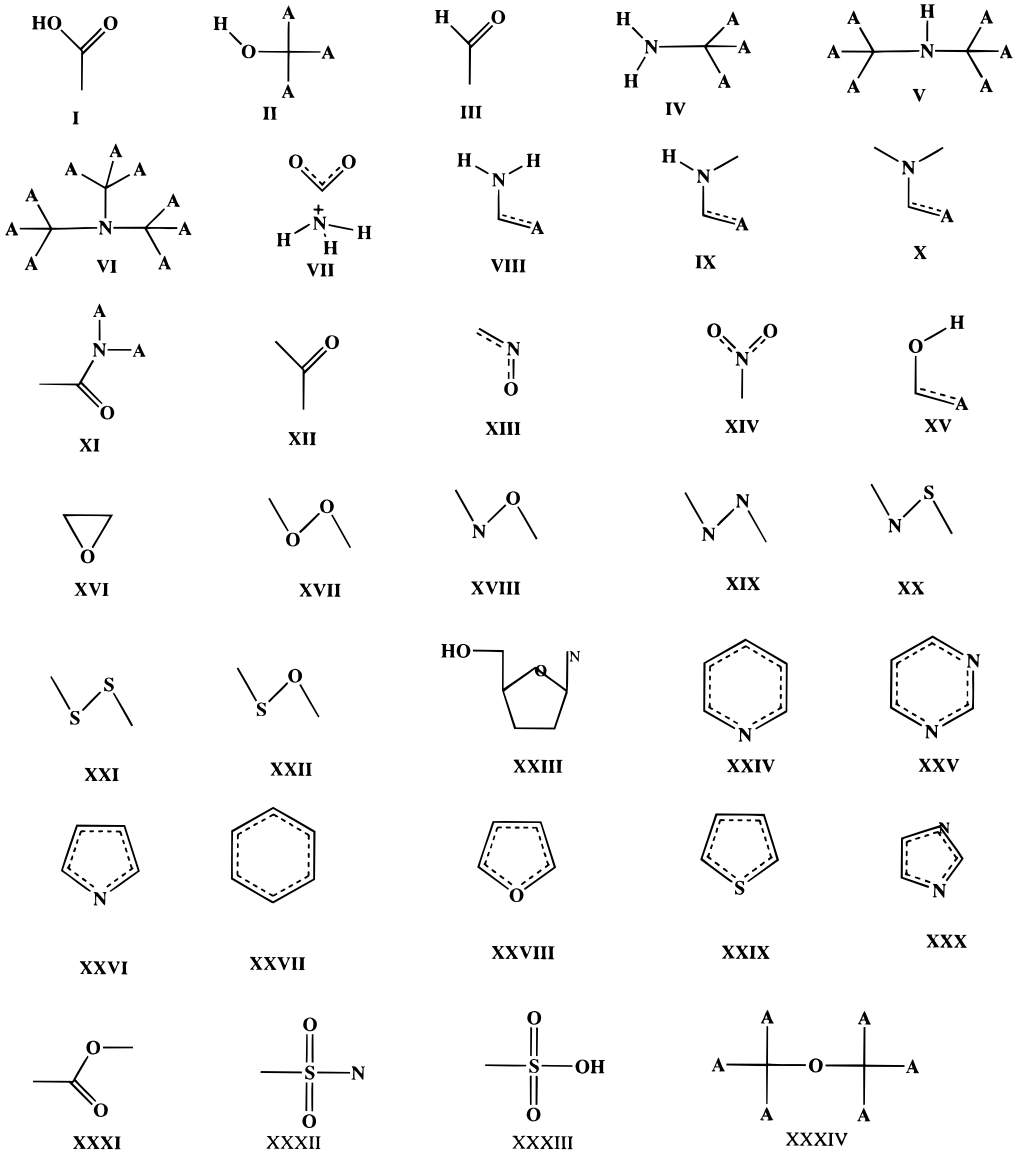

Figure 1. Representation of substructures searched for the comparative analysis of drug databases of Table 1. In these structures, the atoms are carbon unless otherwise indicated. "A" represents any element. The bonds with dashed lines stand for aromatic or delocalized bonds.

lipophilicity. The antipsychotic drugs are considerably more hydrophobic (mean log $P = 4.10$) than antidepressant (mean log $P = 3.10$) or hypnotic (mean log $P = 2.20$) drugs. The standard deviation of log P for all the three classes of CNS active drugs was 1.5 which is considerably lower than several other classes of drugs such as anticancer, cardiovascular, or antiinfective drugs. This is due to the requirement that they should cross the blood brain barrier. Surprisingly, the antiinflammatory drugs are very similar to the antidepressant drugs in their physicochemical property profile. The anticancer drugs are the least lipophilic compounds with a high standard deviation. This is a consequence of the fact that cancer is a complex disease affecting different parts of the body and tissues and often works with chemical brute force rather than milder physical interactions. The antihypertensive drugs are more hydrophilic than the average CMC compound, and they have a reasonably high standard deviation. Despite the low number of the antiinfective drugs, the standard deviation is fairly high for this drug class. Analysis of the ALOGP[24−26] distribution curve shows that most drug classes have very distinct sharp peaks. These peaks are distributed over the ALOGP[24−26] range of 1.0 to 4.0 (see Figure 2). Because of this, the distribution curve of the whole CMC database is considerably flattened compared to any particular drug class. Although the distribution peaks for different classes are different, their overall distributions overlap considerably as indicated by their overlapping qualifying and preferred ranges (Table 2). This indicates that most drugs should fall within a particular range of log P to satisfy a proper physiological distribution and it may be a tool for fine-tuning the efficacy of a drug, though log P by itself does not determine the drug class.

Molecular weight, the number of atoms, and molar refractivity all are related to molecular size. Although the average value, the standard deviation, and the distributions of all these properties are given in Tables 1 and 2 and Figures 2−5, we will analyze only the molecular weight in a greater detail. The CMC database has an average molecular weight of 357 and a standard deviation of 174. The average molecular weights of the drug classes are close, with the

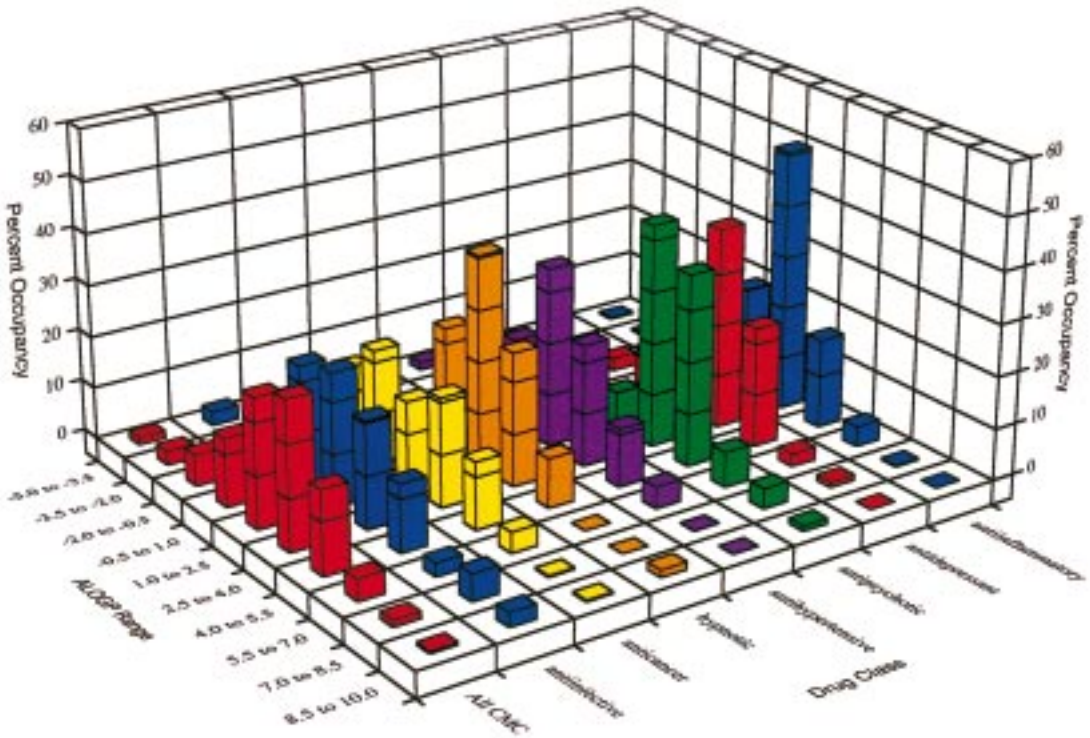

Figure 2. Histogram plots of octanol−water log *P* (ALOGP) distributions for the CMC database and a few other drug classes as described in Table 1.

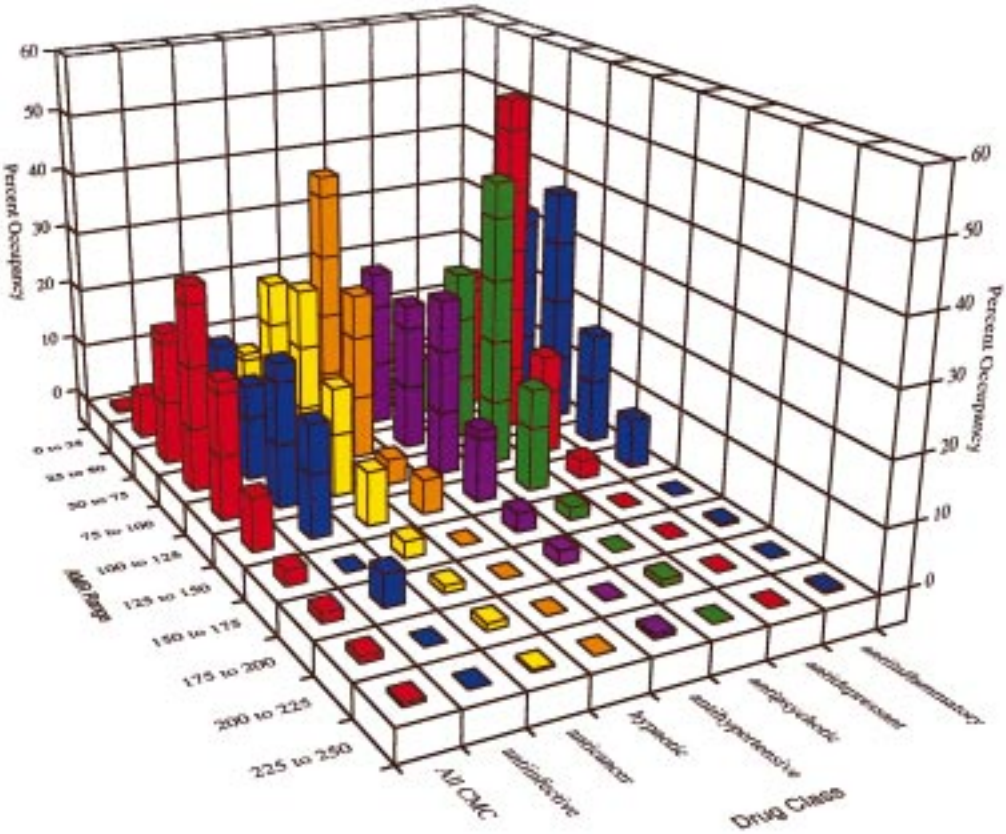

Figure 3. Histogram plots of molar refractivity (AMR) distributions for the CMC database and a few other drug classes as described in Table 1.

exception of the antidepressant and hypnotic drug classes which are somewhat smaller in their average molecular weight. All three classes of CNS active drugs we studied here had considerably smaller standard deviations. We also analyzed the compounds which had a molecular weight significantly outside this range. The compounds in the CMC database which had a molecular weight less than 150 are shown in Table 5. The most important drug classes found

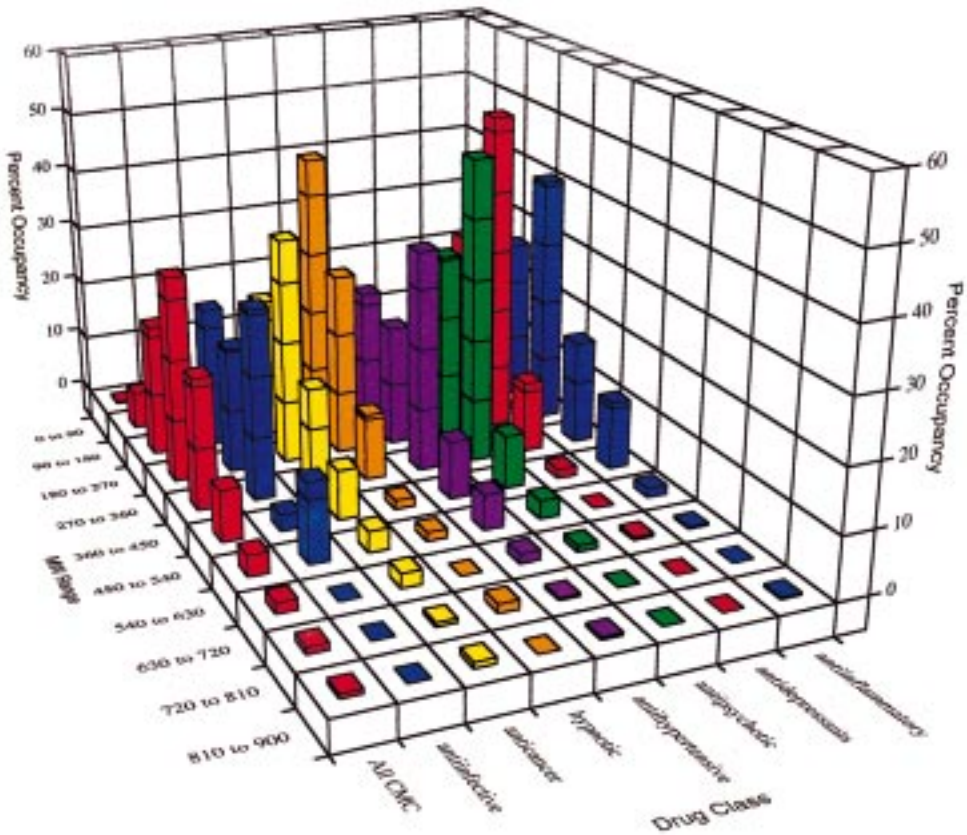

Figure 4. Histogram plots of molecular weight (MW) distributions for the CMC database and a few other drug classes as described in Table 1.

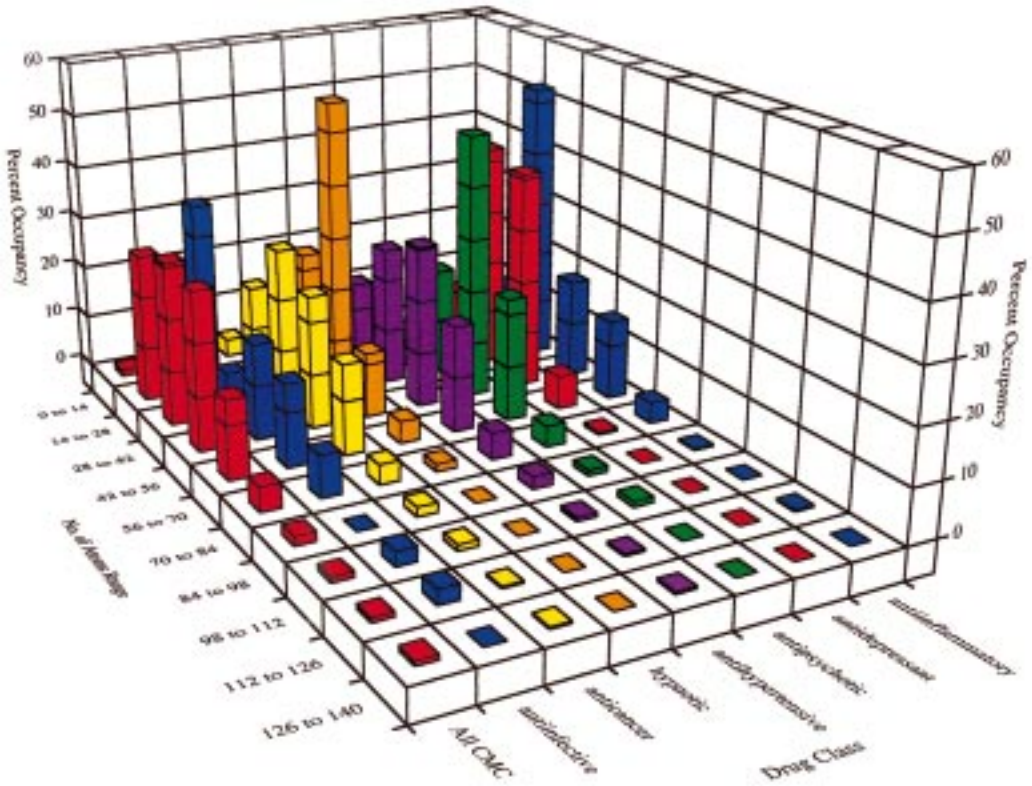

Figure 5. Histogram plots of total number of atoms distributions for the CMC database and a few other drug classes as described in Table 1.

in this list are adrenergic, antineoplastic, anesthetic, anticonvulsant, and nutrient (amino acids). Many of the drug classes such as antineoplastic, anesthetic, and nutrient should be considered differently! In the new drug discovery, we will probably not miss many important drugs if we avoid too small molecules. The unusually large drugs with a

Table 2. Qualifying (Covering 80% of the Total Number of Compounds in That Class) and Preferred (Covering 50% of the Total Number of Compounds) Ranges of the Various Physicochemical Properties in Different Drug Classes

no.	drug class	ALOGP 80%	ALOGP 50%	AMR 80%	AMR 50%	MW 80%	MW 50%	atoms 80%	atoms 50%
1	CMC clean	−0.4; 5.6	1.3; 4.1	40; 130	70; 110	160; 480	230; 390	20; 70	30; 55
2	inflammatory	1.4; 4.5	2.6; 4.2	59; 119	67; 97	212; 447	260; 380	24; 59	28; 40
3	depressant	1.4; 4.9	2.1; 4.0	62; 114	75; 95	210; 380	260; 330	32; 56	37; 48
4	psychotic	2.3; 5.2	3.3; 5.0	85; 131	94; 120	274; 464	322; 422	40; 63	49; 61
5	hypertensive	−0.5; 4.5	1; 3.4	54; 128	68; 116	206; 506	281; 433	28; 66	36; 58
6	hypnotic	0.5; 3.9	1.3; 3.5	43; 97	43; 73	162; 360	212; 306	20; 45	29; 38
7	neoplastic	−1.5; 4.7	0.0; 3.7	43; 128	60; 107	180; 475	258; 388	21; 63	30; 55
8	infective	−0.3; 5.1	0.8; 3.8	44; 144	68; 138	145; 455	192; 392	12; 64	12; 42

Table 3. List of Unusually Hydrophilic Compounds in the CMC Database (ALOGP < −5.0)

generic name	drug class	generic name	drug class	generic name	drug class
acarbose	α-glucosidase (inhibitor)	polymyxin b1	antibacterial	obidoxime chloride	cholinesterase reactivator
anthelmycin	anthelmintic	ristocetin	antibacterial	pralidoxime chloride	cholinesterase reactivator
pentigetide	antiallergic	butirosin	antibacterial	alfadex	complexing agent
paromomycin	antiamebic	homatropine methylbromide	anticholinergic (antispasmodic)	edrophonium chloride	diagnostic aid (myasthenia gravis)
carcainium chloride	antiarrhythmic	tiametonium iodide	anticholinergic (antispasmodic)	pentetic acid	diagnostic aid
capreomycin 1b	antibacterial (tuberculostatic)	edrophonium chloride	antidote (curare)	thiamine	enzyme cofactor
viomycin	antibacterial (tuberculostatic)	sinefungin	antifungal	pentamethonium bromide	ganglionic blocker
streptomycin	antibacterial (tuberculostatic)	azamethonium bromide	antihypertensive	trepirium iodide	ganglionic blocker
enviomycin	antibacterial (tuberculostatic)	pentamethonium bromide	antihypertensive	dicolinium iodide	ganglionic blocker
tobramycin	antibacterial	hexamethonium bromide	antihypertensive	somatostatin	gastro-duodenal ulcers (therapeutic for severe hemorrhage)
propikacin	antibacterial	ademetionine	antiinflammatory	eledoisin	hypotensive
neomycin b [u]	antibacterial	lividomycin	antimicrobial (broad spectrum)	hymotrinan	immunological agent
arbekacin	antibacterial	oxiglutatione	antineoplastic	acemannan	immunomodulator
isepamicin	antibacterial	prospidium chloride	antineoplastic	prolonium iodide	iodine source
dihydrostreptomycin	antibacterial	peplomycin	antineoplastic	eledoisin	lacrimal stimulant
dibekacin	antibacterial	talisomycin	antineoplastic	oxydipentonium chloride	muscle relaxant (skeletal)
ribostamycin	antibacterial	bleomycin a2	antineoplastic	alcuronium chloride	muscle relaxant (skeletal)
amphomycin	antibacterial	capreomycin 1b	antitubercular (antimycotic)	benzoquinonium chloride	muscle relaxant (skeletal)
bluensomycin	antibacterial	viomycin	antitubercular (antimycotic)	succinylcholine chloride	muscle relaxant (skeletal)
streptonicozid	antibacterial	streptomycin	antitubercular (antimycotic)	hexcarbacholine bromide	muscle relaxant (skeletal)
kanamycin	antibacterial	enviomycin	antitubercular (antimycotic)	succinylcholine chloride	neuromuscular blocker
furazolium chloride	antibacterial	tiametonium iodide	antitussive	gallamine triethiodide	neuromuscular blocker
bekanamycin	antibacterial	acemannan	antiviral	suxamethonium bromide	neuromuscular blocker
butikacin	antibacterial	goralatide	bone marrow therapeutic	suxethonium chloride	neuromuscular blocker
colistin	antibacterial	deltibant	bradykinin antagonist	pendetide	scintigraphy (agent)
betamicin	antibacterial	icatibant	bradykinin antagonist	argiprestocin	testicular androgen biosynthesis (inhibitor)
amikacin	antibacterial	ambenonium chloride	cholinergic	terlipressin	vasopressor
apramycin	antibacterial	trimedoxime bromide	cholinesterase reactivator	thiamine	vitamin (cofactor), vitamin (provitamin)

molecular weight greater than 700 are collected in Table 6. A large portion of these compounds are simply the repetition of the unusually hydrophilic antibacterial, antifungal, antibiotic, or anticancer drugs. Unless one is interested in these classes of drugs, making compounds in this molecular weight range may be not be an efficient drug discovery process. The acceptable (qualifying) range here may be 160−480. The structures of the various "unusual drugs"', as given in Tables 3−6, are given in alphabetical order in *Comprehensive Medicinal Chemistry*[29] and may be useful as a reference.

We also studied the correlations of property distributions among these drug classes. These correlation matrices are given in Table 7. These matrices could be used to infer the similarity among drug classes. For example, with respect to log *P* distributions, (i) antiinflammatory drugs are highly correlated with antidepressants and least correlated with

Table 4. List of Unusually Hydrophobic Compounds in the CMC Database (ALOGP > 7.0)

generic name	drug class	generic name	drug class	generic name	drug class
quiflapon	5-lipoxygenase-activating-protein (inhibitor)	lecimibide	antihyperlipidemic	anipamil	calcium channel blocker
flumethrin	acaricide	melinamide	antihyperlipoproteinemic	dihydrotachysterol	calcium regulator
mebolazine	anabolic	salafibrate	antihyperlipoproteinemic	ubidecarenone	cardiovascular agent
bolazine	anabolic	tiafibrate	antihyperlipoproteinemic	steroidal cellobioside	
nandrolone decanoate	anabolic	tocofibrate	antihyperlipoproteinemic	carbamoyl derivs.	cholesterol absorption (inhibitor)
myrophine	analgesic (narcotic)	probucol	antihyperlipoproteinemic	stearylsulfamide	dermatologic
olvanil	analgesic	clinolamide	antihyperlipoproteinemic	flumethrin	ectoparasiticide
nabitan	analgesic	sitofibrate	antihyperlipoproteinemic	declaben	emphysema (therapeutic)
menabitan	analgesic	moctamide	antihyperlipoproteinemic	cholesterol	emulsifier
antrafenine	analgesic	octimibate	antihyperlipoproteinemic	stearyl alcohol	emulsion adjunct
testosterone ketolaurate	androgen	terbuficin	antihyperlipoproteinemic	furostilbestrol	estrogen
nandrolone decanoate	androgen	cetaben	antihyperlipoproteinemic	estradiol undecylate	estrogen
dymanthine	anthelmintic	hexachlorophene	antiinfective (topical)	cistinexine	expectorant
zilantel	anthelmintic	octenidine	antiinfective (topical)	benzquercin	for capillary fragility
bisbendazole	anthelmintic	thymol iodide	antiinfective	prednisolone steaglate	glucocorticoid
bunamidine	anthelmintic	olvanil	antiinflammatory	acebrochol	hypnotic
ly-290324	antiallergic	iralukast	antiischemic (platelet aggregation inhibitor)	gamma oryzanol	hypocholesterolemic
iralukast	antiallergic	halofantrine	antimalarial	orlipastat	hypolipidemic
symetine	antiamebic	lapinone	antimalarial	tereticornate-a	influenza and diarrhea (therapeutic)
butoprozine	antianginal	menoctone	antimalarial	iralukast	leukotriene (antagonist)
butoprozine	antiarrhythmic (cardiac depressant)	diathymosulfone	antimycobacterial	quiflapon	leukotriene-synthesis (inhibitor)
amiodarone	antiarrhythmic (cardiac depressant)	buclizine	antinauseant	tripalmitin	lung disorders (therapeutic)
declaben	antiarthritic	carzelesin	antineoplastic	glyceryltrierucate	multiple sclerosis (therapeutic)
iralukast	antiasthmatic	phenesterin	antineoplastic	duoperone	neuroleptic
eldacimibe	antiatherosclerotic	thalicarpine	antineoplastic	iralukast	phospholipase a2 (inhibitor)
chaulmosulfone	antibacterial (leprostatic)	verteporfin	antineoplastic	temoporfin	photosensitizer
clofazimine	antibacterial (tuberculostatic) (leprostatic)	bizelesin	antineoplastic	phytonadione	prothrombogenic
chloramphenicol palmitate	antibacterial	atrimustine	antineoplastic	triolein i 125	radioactive agent
clindamycin palmitate	antibacterial	aminoquinol	antiprotozoal (leishmaniasis)	iodocholesterol i 131	radioactive agent
clofoctol	antibacterial	fluphenazine enanthate	antipsychotic	adibenzylnor-spermidine	radioprotective
quindecamine	antibacterial	fluphenazine decanoate	antipsychotic	tocofenoxate	reverses aging of murine cells
biclotymol	antibacterial	bromperidol decanoate	antipsychotic	acebrochol	sedative
diathymosulfone	antibacterial	haloperidol decanoate	antipsychotic	isopropyl palmitate	vehicle (oleaginous)
gefarnate	anticholinergic (antispasmodic)	pipotiazine palmitate	antipsychotic	ergocalciferol	vitamin (antirachitic)
menatetrenone	anticoagulant	penfluridol	antipsychotic	ergocalciferol	vitamin (cofactor)
nabazenil	anticonvulsant	cholecalciferol	antirachitic	vitamin e	vitamin (cofactor)
butoprozine	antidepressant	chloramphenicol palmitate	antirickettsial	phytonadione	vitamin (cofactor)
amiodarone	antidepressant	teroxalene	antischistosomal	cholecalciferol	vitamin (cofactor)
naboctate	antiemetic	clofazimine	antitubercular (antimycotic)	ergocalciferol	vitamin (provitamin)
idoxifene	antiestrogen	dimyristoyl-phosphatidyl-azt	antiviral (aids therapeutic)	vitamin e	vitamin (provitamin)
thymol iodide	antifungal	cicloxolone	ntiviral (herpes genitalis)	phytonadione	vitamin (provitamin)
naboctate	antiglaucoma agent	avridine	antiviral	cholecalciferol	vitamin (provitamin)
eldacimibe	antihyperlipidemic	perfluamine	blood substitute	scarlet red	vulnerary

anticancer compounds; (ii) antidepressants are also least correlated with anticancer compounds; (iii) antipsychotics are generally less correlated with other drug classes and the least with antihypertensives; (iv) antihypertensives are well correlated with most drug classes and the most with hypnotics; (v) anticancer compounds are understandably less correlated with most drug classes in general, but have the best correlation with antiinfectives. With respect to molar refractivity distributions, (i) antiinflammatory drugs are highly correlated with anticancer compounds and least with antipsychotics; (ii) antidepressants are least correlated with antiinfectives; (iii) antipsychotics are least correlated with hypnotics. With respect to molecular weight distributions, (i) antiinflammatory drugs are highly correlated with anticancer compounds and least correlated with antipsychotic compounds; (ii) antidepressants are least correlated with antipsychotic compounds; (iii) antipsychotics are generally less correlated with other drug classes and the least with

Table 5. List of Unusually Low Molecular Weight Compounds in the CMC Database (MW < 150)

generic name	drug class	generic name	drug class	generic name	drug class
glutamic acid	acidifier (gastric)	metacresol	antifungal	imexon	immunostimulant
racemethionine	acidifier (urinary)	ornithine	antihyperlipoproteinemic	ethohexadiol	insect repellent
fumaric acid	acidifier	oxiniacic acid	antihyperlipoproteinemic	deferiprone	iron chelating agent
cyclopentamine	adrenergic	oxibetaine	antihyperlipoproteinemic	racemethionine	lipotropic
phenpromethamine	adrenergic	γ-aminobutyric acid	antihypertensive	methionine	lipotropic
tyramine	adrenergic	carbamide peroxide	antiinfective (topical)	choline chloride	lipotropic
tuaminoheptane	adrenergic	dimethyl sulfoxide	antiinflammatory (topical)	mecysteine	mucolytic
octodrine	adrenergic	phenylethyl alcohol	antimicrobial agent (ophthalmic)	fampridine	multiple sclerosis (therapeutic)
fomepizole	alcohol dehydrogenase (inhibitor)	sorbic acid	antimicrobial agent	γ-aminobutyric acid	neurotransmitter
methyl isobutyl ketone	alcohol denaturant	carzolamide	antineoplastic	tetrazolylglycine	nmda agonist
tromethamine	alkalizer	alanosine	antineoplastic	dimiracetam	nootropic
diethanolamine	alkalizing agent	cycloleucine	antineoplastic	rolziracetam	nootropic
diisopropanolamine	alkalizing agent	urethane	antineoplastic	methionine	nutrient (amino acid)
trolamine	alkalizing agent	diazouracil	antineoplastic	lysine	nutrient (amino acid)
trichloroethylene	analgesic (inhalation)	lycidyl methacrylate	antineoplastic	isoleucine	nutrient (amino acid)
picolamine	analgesic	fluorouracil	antineoplastic	aspartic acid	nutrient (amino acid)
gaboxadol	analgesic	thiocarzolamide	antineoplastic	leucine	nutrient (amino acid)
trolamine	analgesic	imexon	antineoplastic	serine	nutrient (amino acid)
salicylamide	analgesic	butanediol cyclic sulfite	antineoplastic	threonine	nutrient (amino acid)
acetanilide	analgesic	dianhydrogalactitol	antineoplastic	alanine	nutrient (amino acid)
trichloroethylene	anesthetic (inhalation)	imidazopyrazole	antineoplastic	valine	nutrient (amino acid)
cyclopropane	anesthetic (inhalation)	guanazole	antineoplastic	proline	nutrient (amino acid)
ethylene	anesthetic (inhalation)	methyl methanesulfonate	antineoplastic	glycine	nutrient
fluroxene	anesthetic (inhalation)	methylformamide	antineoplastic	cinnamaldehyde	perfume agent
norflurane	anesthetic (inhalation)	aminothiadiazole	antineoplastic	isaxonine	peripheral neuropathies (therapeutic)
ether	anesthetic (inhalation)	hydroxyurea	antineoplastic	cysteamine	radioprotective
vinyl ether	anesthetic (inhalation)	ammonium lactate	antipruritic (topical)	glycerin	reduces intraocular and intracranial pressure
salicyl alcohol	anesthetic (local)	acetanilide	antipyretic	methyl nicotinate	rubifacient
ethyl chloride	anesthetic (topical)	metacresol	antiseptic (topical)	monoethanolamine	sclerosing agent
cathinone	anorexic	isoniazid	antitubercular (antimycotic)	paraldehyde	sedative
phentermine	anorexic	pyrazinamide	antitubercular (antimycotic)	meparfynol	sedative
levamfetamine	anorexic	cycloserine	antitubercular (antimyotic)	ethchlorvynol	sedative
piperazine	anthelmintic (as citrate)	cyacetacide	antitubercular	dextroamphetamine	stimulant (central)
metyridine	anthelmintic	cysteamine	antiurolithic	methamphetamine	stimulant (central)
cyacetacide	anthelmintic	amitivir	antiviral	ampyzine	stimulant (central)
thiouracil	antianginal	fosfonet	antiviral	piracetam	stimulant (central)
parachlorophenol	antibacterial (topical)	dimepranol	antiviral	amphetamine	stimulant (central)
isoniazid	antibacterial (tuberculostatic)	foscarnet	antiviral	pentylenetetrazole	stimulant (central)
pyrazinamide	antibacterial (tuberculostatic)	kethoxal	antiviral	histamine	stimulant (gastric secretory)
cycloserine	antibacterial (tuberculostatic)	creatinine	bulk agent for freeze-drying	cathinone	stimulant
methenamine	antibacterial (urinary)	heptaminol	cardiotonic	thiouracil	thyroid (inhibitor)
bacitracin a	antibacterial	phenylpropanol	choleretic	aminothiazole	thyroid (inhibitor)
fosfomycin	antibacterial	timonacic	choleretic	methimazole	thyroid (inhibitor)
taurultam	antibacterial	piracetam	cognition enhancer	mipimazole	thyroid (inhibitor)
benzyl alcohol	antibacterial	allyl isothiocyanate	counter-irritant	methylthiouracil	thyroid (inhibitor)
ornithine	anticholesteremic	captamine	depigmentor	pimagedine hcl	tooth discoloration (inhibitor)
vigabatrin	anticonvulsant (tardive dyskinesia)	hydroquinone	depigmentor	cetohexazine	tranquilizer
valpromide	anticonvulsant	cysteine	detoxicant	emylcamate	tranquilizer
trimethadione	anticonvulsant	ethyl nitrite	diaphoretic	valnoctamide	tranquilizer
milacemide	anticonvulsant	isosorbide	diuretic	acetohydroxamic acid	urease (inhibitor)
dimethadione	anticonvulsant	urea	diuretic	tiformin	uremic diabetes (therapeutic)
valproic acid	anticonvulsant	ethyl nitrite	diuretic	cyclopentamine	vasoconstrictor
ethosuximide	anticonvulsant	niacinamide	enzyme cofactor	octodrine	vasoconstrictor
levcycloserine	anticonvulsant	niacin	enzyme cofactor	nicotinyl alcohol	vasodilator (peripheral)
milacemide	antidepressant	levcycloserine	enzyme gaucher's disease (inhibitor)	betahistine	vasodilator
phenelzine	antidepressant	guaiacol	expectorant	aminoethyl nitrate	vasodilator
octamoxin	antidepressant	cinnamaldehyde	flavor agent	niacinamide	vitamin (cofactor)
tranylcypromine	antidepressant	bacitracin a	food additive	niacin	vitamin (cofactor)
mebanazine	antidepressant	amogastrin	gastric secretion stimulant	niacinamide	vitamin (provitamin)
metformin	antidiabetic	aminocaproic acid	hemostatic	niacin	vitamin (provitamin)
cysteamine	antidote (acetaminophen)	nafarelin acetate	hormone agonist (gonadotrophin releasing)	adenine	vitamin b4
dimercaprol	antidote (heavy metal)	pidolic acid	humectant (as Na salt)	allopurinol	vulnerary
taurultam	antifungal	mequinol	hyperpigmentation (therapeutic)	trientine	wilson's disease (therapeutic)
octanoic acid	antifungal	paraldehyde	hypnotic	allopurinol	xanthine oxidase (inhibitor)
benzoic acid	antifungal	meparfynol	hypnotic		
flucytosine	antifungal	cetohexazine	hypnotic		

Table 6. List of Unusually High Molecular Weight Compounds in the CMC Database (MW > 700)

generic name	drug class	generic name	drug class	generic name	drug class
cp-331	analgesic	paulomycin a	antibacterial	atrimustine	antineoplastic
ivermectin b1a	anthelmintic (oncocerchiasis)	clarithromycin	antibacterial	vinrosidine	antineoplastic
abamectin b1b	anthelmintic	scopafungin	antibacterial	peplomycin	antineoplastic
fuladectin a3	anthelmintic	pristinamycin	antibiotic	lanreotide	antineoplastic
fuladectin a4	anthelmintic	ardacin	antibiotic	talisomycin	antineoplastic
iralukast	antiallergic	penimocycline	antibiotic	toyomycin	antineoplastic
berythromycin	antiamebic	hamycin a	antibiotic	vinylglycinate	antineoplastic
iralukast	antiasthmatic	salinomycin	anticoccidial	vincristine	antineoplastic
pamaqueside	antiatherosclerotic	maduramicin	anticoccidial	dactinomycin	antineoplastic
rifapentine	antibacterial (antitubercular)	seglitide	antidiabetic	bleomycin a2	antineoplastic
rifampin	antibacterial (antitubercular)	lypressin	antidiuretic	rodorubicin	antineoplastic
rifamide	antibacterial (antitubercular)	desmopressin	antidiuretic	vinfosiltine	antineoplastic
rifabutin	antibacterial (antitubercular)	suramin	antifilarial	docetaxol	antineoplastic
14-hydroxycl-arithromycin	antibacterial (broad spectrum)	candicidin (levorin a2)	antifungal (topical)	nogalamycin	antineoplastic
maridomycin	antibacterial (gram-positive)	levorin a2	antifungal agent	bryostatin-1	antineoplastic
chaulmosulfone	antibacterial (leprostatic)	basifungin	antifungal agent	vinformide	antineoplastic
gramicidin	antibacterial (topical)	lucimycin	antifungal	eprinomectin b1b	antiparasitic
berythromycin	antibacterial	cilofungin	antifungal	eprinomectin b1a	antiparasitic
mocimycin	antibacterial	nystatin a1	antifungal	partricin a	antiprotozoal
virginiamycin factor s	antibacterial	amphotericin b	antifungal	pipotiazine palmitate	antipsychotic
coumermycin a1	antibacterial	rutamycin	antifungal	cp-331	antipyretic
rifamexil	antibacterial	scopafungin	antifungal	octreotide	antisecretory (gastric)
phenyracillin	antibacterial	partricin a	antifungal	suramin	antitrypanosomal
mirosamicin	antibacterial	fungimycin	antifungal	rifapentine	antitubercular (antimycotic)
penimepicycline	antibacterial	itraconazole	antifungal	rifampin	antitubercular (antimycotic)
rifametane	antibacterial	tiqueside	antihyperlipidemic	rifamide	antitubercular (antimycotic)
lexithromycin	antibacterial	pantenicate	antihyperlipoproteinemic	rifabutin	antitubercular (antimycotic)
primycin	antibacterial	glunicate	antihyperlipoproteinemic	sucrosofate	antiulcerative (K salt)
josamycin	antibacterial	dextrothyroxine	antihyperlipoproteinemic	ilatreotide	antiulcerative
flurithromycin	antibacterial	etiroxate	antihyperlipoproteinemic	dimyristoyl-phosphatidyl-azt	antiviral (aids therapeutic)
pristinamycin	antibacterial	ditekiren	antihypertensive (rennin inhibitor)	streptovarycin c	antiviral
paldimycin b	antibacterial	saralasin	antihypertensive	acemannan	antiviral
rokitamycin	antibacterial	sr-43845	antihypertensive	1 731723	antiviral
amphomycin	antibacterial	fk-744	antihypertensive	ritonavir	antiviral
azithromycin	antibacterial	bietaserpine	antihypertensive	palinavir	antiviral
erythromycin stinoprate	antibacterial	protoveratrine a	antihypertensive	deltibant	bradykinin antagonist
rifaximin	antibacterial	zankiren	antihypertensive	icatibant	bradykinin antagonist
streptonicozid	antibacterial	mipragoside	antiinflammatory	gitoformate	cardiac glycoside
erythromycin ethylsuccinate	antibacterial	proglumetacin	antiinflammatory	digitoxin	cardiotonic
carbomycin	antibacterial	cp-331	antiinflammatory	acetyldigitoxin	cardiotonic
midecamycin	antibacterial	iralukast	antiischemic (platelet aggregation inhibitor)	lanatoside c	cardiotonic
kitasamycin	antibacterial	lividomycin	antimicrobial (broad spectrum)	digoxin	cardiotonic
erythromycin	antibacterial	streptovarycin c	antimicrobial	deslanoside	cardiotonic
vancomycin hcl	antibacterial	triptorelin	antineoplastic (prostatic carcinoma therapeutic)	pengitoxin	cardiotonic
erythromycin propionate	antibacterial	leuprolide	antineoplastic (prostatic carcinoma)	metildigoxin	cardiotonic
dirithromycin	antibacterial	vinorelbine	antineoplastic alkaloid	gitaloxin	cardiotonic
diproleandomycin	antibacterial	carzelesin	antineoplastic	ubidecarenone	cardiovascular agent
troleandomycin	antibacterial	aclarubicin	antineoplastic	sincalide	choleretic
colistin	antibacterial	vindesine	antineoplastic	steroidal cellobioside	
ramoplanin a2	antibacterial	vinzolidine	antineoplastic	carbamoyl derivs.	cholesterol absorption (inhibitor)
megalomicin	antibacterial	vinblastine	antineoplastic	11-ketotigogenin cellobioside	cholesterol absorption (inhibitor)
relomycin	antibacterial	plicamycin	antineoplastic	demecarium bromide	cholinergic (ophthalmic)
spiramycin	antibacterial	echinomycin	antineoplastic	salinomycin	coccidiostat
aspartocin	antibacterial	taxol	antineoplastic	semduramicin	coccidiostat
quinupristin	antibacterial	didemnin b	antineoplastic	sulbutiamine	coenzyme precursor
polymyxin b1	antibacterial	vinepidine	antineoplastic	alfadex	complexing agent
piridicillin	antibacterial	olivomycin a	antineoplastic	truxipicurium iodide	curariform
tylosin	antibacterial	vinleucinol	antineoplastic	iobitridol	diagnostic agent
erythromycin acistrate	antibacterial	ditercalinium chloride	antineoplastic	pentagastrin	diagnostic aid (gastric secretion indicator)
ristocetin	antibacterial	vapreotide	antineoplastic	sulfobromophthalein	diagnostic aid (hepatic function)
roxithromycin	antibacterial	verteporfin	antineoplastic	ioxilan	diagnostic aid
avilamycin-a	antibacterial	bizelesin	antineoplastic	iofratol	diagnostic aid

Table 6 (Continued)

generic name	drug class	generic name	drug class	generic name	drug class
teprotide	enzyme angiotensin-converting (inhibitor)	geclosporin	immunosuppressive	cagutocin	oxytocic
bisbutitiamine	enzyme cofactor	eledoisin	lacrimal stimulant	atosiban	oxytocin antagonist
cistinexine	expectorant	iralukast	leukotriene (antagonist)	iralukast	phospholipase a2 (inhibitor)
sorbinicate	fibrinolytic	deslorelin	lhrh agonist	goserelin	prostatic carcinoma (therapeutic)
flavin adenin dinucleotide	flavin coenzyme	lutrelin	lhrh agonist	triolein i 125	radioactive agent
ahn-683	fluorescent ligand for analysis of benzodiazepine receptors	histrelin	lhrh agonist	thyroxine i 125	radioactive agent
		detirelix	lhrh antagonist	pendetide	scintigraphy (agent)
benzquercin	for capillary fragility	tripalmitin	lung disorders (therapeutic)	petrichloral	sedative
somatostatin	gastro-duodenal ulcers (therapeutic for severe hemorrhage)	cetrorelix	luteinizing-hormone-releasing-hormone antagonist	zaragozic acid	squalene synthase (inhibitor)
gonadorelin	gonad-stimulating principle	glyceryltrierucate	multiple sclerosis (therapeutic)	ceruletide	stimulant of gastric secretion
buserelin	gonad-stimulant	truxipicurium iodide	muscle relaxant (general)	argiprestocin	testicular androgen biosynthesis (inhibitor)
ganirelix	gonad-stimulating principle	pipecuronium bromide	muscle relaxant (general)	thyromedan	thyroid hormone
fertirelin	gonadotropin-releasing hormone (vet.)	dimethyltubo-curarinium chloride	muscle relaxant (skeletal)	levothyroxine	thyroid hormone
examorelin	growth hormone releaser	doxacurium chloride	muscle relaxant (skeletal)	thyromedan	thyromimetic
ritonavir	hiv-1 and hiv-2 (inhibitor)	alcuronium chloride	muscle relaxant (skeletal)	levothyroxine	thyromimetic
l 731723	hiv-1 protease (inhibitor)	atracurium besilate	muscle relaxant (skeletal)	lypressin	vasoconstrictor
vasopressin	hormone (antidiuretic)	gallamine triethiodide	neuromuscular blocker	felypressin	vasoconstrictor
petrichloral	hypnotic	metocurine iodide	neuromuscular blocker	angiotensin ii	vasoconstrictor
pamaqueside	hypocholesteremic	mivacurium chloride	neuromuscular blocker	ornipressin	vasoconstrictor
eledoisin	hypotensive	pancuronium bromide	neuromuscular blocker	angiotensin amide	vasoconstrictor
acemannan	immunomodulator	laudexium methyl sulfate	neuromuscular blocker	inositol niacinate	vasodilator (peripheral)
thymoctonan	immunomodulator	cisatracurium besylate	neuromuscular blocking agent	terlipressin	vasopressor
		ebiratide	nootropic	troxerutin	venous disorders (therapeutic)
romurtide	immunostimulant	oxytocin	oxytocic	bisbentiamine	vitamin (cofactor)
sirolimus	immunosuppressant	carbetocin	oxytocic	bisbentiamine	vitamin (provitamin)
tacrolimus	immunosuppressant	demoxytocin	oxytocic	bisbentiamine	vitamin b1 source
oxeclosporin	immunosuppressive agent				
cyclosporine	immunosuppressive				

hypnotics; (iv) antihypertensives are well correlated with most drug classes and the most with antiinfectives; (v) anticancer compounds have the best correlation with antihypertensives. At this point we are not very sure whether such correlations can be used to design interchangeable compound libraries.

Composition of Functional Groups. The knowledge of functional groups, rings, and any interesting structural moieties in specific drug classes or the whole CMC data set may be an important aid in the design of directed libraries or universal drug libraries. Figure 1 shows the set of organic functional groups considered in the present analysis. While not exhaustive, they cover most of the commonly occurring functional groups used to classify organic compounds, common aromatic rings, and some of the structural moieties that Rishton[40] mentioned as reactive groups that should be avoided in drugs. Table 8 shows the frequency of occurrence of these groups in seven drug classes that we studied along with that in the whole CMC database.

The benzene ring (XXVI) appears to be the most abundant structural unit in the whole data set as well as in any of the drug classes that we studied. It constituted around 70% of the whole data set and often more than 50% in the various drug classes that we studied. It is interesting to note that this number is slightly larger than the total number of heterocyclic compounds (aromatics and nonaromatics combined: 4314). This is not unexpected because of the easy chemistry around a benzene ring and the absence of stereochemical isomers. The phenyl ring has a relatively low

Table 7. Correlation Matrices Comparing the Eight Databases of Table 1

Part A

1	2	3	4	5	6	7	8
1.00	0.90	0.96	0.71	0.93	0.89	0.82	0.89
0.96	1.00	0.96	0.86	0.71	0.67	0.63	0.66
0.93	0.93	1.00	0.85	0.82	0.77	0.66	0.75
0.76	0.62	0.60	1.00	0.43	0.37	0.45	0.44
0.93	0.90	0.76	0.83	1.00	0.98	0.78	0.94
0.72	0.81	0.67	0.17	0.63	1.00	0.74	0.96
0.97	0.97	0.89	0.62	0.90	0.87	1.00	0.83
0.75	0.66	0.52	0.73	0.82	0.58	0.77	1.00

Part B

1	2	3	4	5	6	7	8
1.00	0.98	0.93	0.77	0.84	0.81	0.99	0.82
0.87	1.00	0.96	0.65	0.73	0.83	0.99	0.73
0.96	0.86	1.00	0.53	0.59	0.84	0.97	0.64
0.85	0.55	0.81	1.00	0.91	0.34	0.66	0.82
0.98	0.79	0.90	0.88	1.00	0.62	0.74	0.94
0.78	0.93	0.78	0.33	0.67	1.00	0.84	0.68
0.98	0.87	0.91	0.77	0.97	0.81	1.00	0.75
0.46	0.20	0.31	0.31	0.55	0.34	0.59	1.00

[a] In part A, the upper triangle gives the correlation coefficients of calculated log *P* distributions (using ALOGP method) for each pair of the databases (numbered according to Table 1). The lower triangle is the corresponding correlation matrix for calculated molar refractivity (using AMR method). In part B, the upper triangle gives the correlation coefficients of calculated molecular weight distributions for each pair of the databases (numbered according to Table 1). The lower triangle is the corresponding correlation matrix for the distributions of molecular size, as mesaured by the number of atoms.

Table 8. Composition of Functional Groups (Identified in Figure 1) among Drugs Classified by Disease State: (a) Antiinflammatory, (b) Antidepressants, (c) Antipsychotic, (d) Antihypertensive, (e) Hypnotics, (f) Anticancer, (g) Antiinfectives, (h) CMC[a]

no.	description	(a) 293	(b) 222	(c) 110	(d) 368	(e) 75	(f) 431	(g) 37	(h) 6454
I	carboxyl	114	6	2[b]	87	0	39	7	972
II	alcohol	69	25	20	82	5	161	5	1668
III	aldehyde	1[c]	0	0	0	0	2[c]	0	34[d]
IV	aliphatic primary amine	3	15	0	17	0	43	1	367
V	aliphatic secondary amine	1	46	1	63	0	16	3	587
VI	aliphatic tertiary amine	21	116	102	66	9	59	5	1910
VII	amino acid	2	0	0	5	0	5	0	96
VIII	aromatic primary amine	6	3	2	15	2	35	4	350
IX	aromatic secondary amine[e]	38	10	13	30	0	53	1	462
X	aromatic tertiary amine	14	41	40	39	6	44	7	663
XI	carboxamide	60	53	23	106	44	109	1	1752
XII	keto	91	8	26	21	4	109	5	1014
XIII	*N*-oxide	0	0	0	3	0	1	0	12
XIV	nitro	2	1	0	13	2	11	5	170
XV	phenolic OH	19	4	3	25	1	58	7	660
XVI	epoxy	0	0	0	0	0	10	0	49
XVII	C−O−O−C	0	0	0	0	0	0	0	4
XVIII	C−N−O−C	2	0	0	1	0	1	0	16
XIX	C−N−N−C	17	5	0	14	1	4	0	100
XX	C−N−S−C[f]	0	0	0	0	0	0	0	4
XXI	C−S−S−C	0	0	0	1	0	5	0	42
XXII	C−S−O−C[g]	0	0	0	0	0	0	0	0
XXIII	nucleoside	0	0	0	0	0	33	0	66
XXIV	pyridine	29	20	5	30	2	27	5	521
XXV	pyrimidine	2	2	0	15	0	27	0	158
XXVI	pyrrole	24	16	12	22	0	30	0	286
XXVII	benzene	224	205	107	292	34	189	24	4536
XXVIII	furan	5	3	1	5	0	2	4	128
XXIX	thiophene	13	4	3	9	2	6	1	124
XXX	imidazole	9	1	2	22	3	38	3	388
XXXI	ester	74	6	8	92	3	86	0	1174
XXXII	sulfonamide	16	3	3	32	0	4	0	291
XXXIII	sulfonic acid	0	0	0	1	0	4	0	42
XXXIV	aliphatic ether	8	0	0	1	0	0	0	46
	aromatic ether	2	4	1	27	1	1	0	81
	heterocyclic (any)	169	149	105	270	56	285	21	4314
	heterocyclic (aromatic)	103	52	27	118	12	127	13	1832

[a] The numbers in the column label show the number of compounds found in the CMC database for the respective classes. [b] The acid counterpart of a salt. [c] Sugar or aromatic aldehydes. [d] Mostly antibiotics, sugar, and aromatic aldehydes. [e] Diaryl or aryl alkyl. [f] Excluding C−N−S(=O)−C. [g] Excluding C−O−S(=O)−C.

desolvation cost[41] and can serve as a scaffold for polar or hydrophobic functionalities, and it often contributes to hydrophobic interactions. The common heterocyclic aromatic rings, when each type of ring is considered separately, had a fairly low occurrence in the CMC data set as well as in the various drug classes studied here. Among nitrogen-containing heterocycles, the pyridine ring is most common, which may be understood by its somewhat chemical inertness. In addition it can act as an H-bond acceptor. The number of compounds where there was a carbon heteroatom bond in a ring was 4314. The number of compounds which had both a benzene ring and a carbon−heteroatom ring bond was 3163, indicating that this combination is likely to create successful drug candidates.

Among the common organic functional groups considered, the alcoholic hydroxyl (II) and the carboxamide group (XI) occur with a high frequency among the CMC database as well as in the various drug classes studied here. This is also expected, since they have both hydrogen-accepting and -donating abilities in a hydrogen bond. They are hydrophilic and chemically stable, neutral groups. Aliphatic tertiary amine had a comparable presence in the CMC database. It outnumbered aliphatic primary or secondary amines by 3- to 5-fold. Both basicity and biochemical stability may be the reasons for the success of tertiary amines. The keto (XII) and the carboxy esters (XXXI) were found in comparable occurrence. We did find compounds with a single bond between heteroatoms; however, they are mostly stabilized by a conjugated C=X bond. The antineoplastic compounds often have reactive functionalities, and such drugs should be considered as special brute force drugs and should be avoided in the regular drug design process. The carboxylic acid, aromatic secondary and tertiary amines, and aliphatic secondary amines also constituted moderately in the drug database. It is interesting to note that carboxyl acid groups are virtually absent among antipsychotic, antidepressant, and hypnotic drug classes, all of which act on the central nervous system. As is well known, CNS acting drugs have to cross the blood brain barrier, requiring them to be sufficiently lipophilic, thus disfavoring acid groups.

From the analysis presented here we may provide the following consensus definition of a drug-like molecule: (i) an organic compound having a calculated log *P* (ALOGP) between −0.4 and 5.6, a molar refractivity (AMR[25,31,32])

between 40 and 130, a molecular weight between 160 and 480, and the total number of atoms between 20 and 70; (ii) structurally a combination of some of the following groups: a benzene ring, a heterocyclic ring (both aliphatic and aromatic), an aliphatic amine (preferably tertiary), a carboxamide group, an alcoholic hydroxyl group, a carboxy ester, and a keto group; (iii) chemically stable in the physiological buffer, as obvious by the absence of a reactive functional group or structural moiety.

4. Conclusion

We provided here an analysis of some computable physicochemical properties and chemical constitutions of known drug molecules available in the CMC database and seven known drug classes. Our study showed that the qualifying range (the chance of missing good compounds is less than 20%) of calculated log *P* (ALOGP[24−26]) for drug-like molecules is −0.4 to 5.6. The mean ALOGP[24−26] is 2.3, and the preferred range (most populated for an interval having 50% of the drugs) is 1.3 to 4.1. For molar refractivity the qualifying range is 40 to 130. The mean is 97, and the preferred range is 70 to 110. For molecular weight the qualifying range is 160 to 480. The mean molecular weight is 360, and the preferred range is 230 to 390. For the total number of atoms the mean value is 48, and the qualifying range is 20 to 70. The preferred range for the total number of atoms is 30 to 55. For different drug classes the ranges may be considerably tighter than what we stated above.

Benzene is the most abundant structural unit found in the drug database. It outnumbered any other structural unit by a few folds. Tertiary aliphatic amine is the most frequent functional group in the drug molecules. Carboxamides and alcohols are the two other groups whose frequency of occurrence was close to the aliphatic tertiary amine group. The frequency of occurrence of the common heterocyclic aromatic rings is far less than that of the aliphatic heterocyclic rings.

The consensus definition of a drug-like molecule obtained here will help to streamline the design of combinatorial chemistry libraries for drug design as well as in developing a more efficient corporate medicinal chemistry library.

References and Notes

(1) Drews, J. Genomic sciences and the medicine of tomorrow. *Nature Biotechnol.* **1996**, *14*, 1516−1518.

(2) Drews, J. Sciences towards the Medicine of Tomorrow. *Chimia* **1996**, *30*, 507−510.

(3) Armstrong, R. W.; Combs, A. P.; Tempest, P. A.; Brown, S. D.; Keating, T. A. Multiple-component condensation strategies for combinatorial library synthesis. *Acc. Chem. Res.* **1996**, *29*, 123−131.

(4) Chaiken, I. M.; Janda, K. D. *Molecular Diversity and Combinatorial Chemistry: ACS Conference Proceeding Series*; American Chemical Society: Washington, DC, 1996

(5) Caflisch, A.; Karplus, M. Computational combinatorial chemistry for de novo ligand design: Review and assessment. *Perspect. Drug Discovery Des.* **1995**, *3*, 51−84.

(6) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Adapting structure-based drug design in the paradigm of combinatorial chemistry and high-throughput screening: An overview and new examples with important caveats for newcomers to combinatorial library design using pharmacophore models or multiple copy simultaneous search (MCSS) fragments. In *Rational Drug Design*; Parrill, A., Reddy, M. R., Eds.; American Chemical Society: Washington, DC, 1998; in press.

(7) Moran, E. J.; Sarshar, S.; Cargill, J. F.; Shahbaz, M. M.; Lio, A.; Mjalli, A. M. M.; Armstrong, R. W. Radio frequency tag encoded combinatorial library method for the discovery of tripeptide-substituted cinnamic acid inhibitors of the protein tyrosine phosphatase PTP1B. *J. Am. Chem. Soc.* **1995**, *117*, 10787−10788.

(8) Joseph-McCarthy, D.; Hogle, J. M.; Karplus, M. Use of the multiple copy simultaneous search (MCSS) method to design a new class of picornavirus capsid binding drugs. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 32−58.

(9) (a) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078−1082. (b) Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J. A.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. Structure-Based Design and Combinatorial Chemistry Yield Low Nanomolar Inhibitors of Cathepsin D. *Chem. Biol.* **1997**, *4*, 297−307. (c) See also the web page *www.combinatorial.com.*

(10) Kollman, P. Molecular-Dynamics and Free-Energy Perturbation Calculations − What Role Do They Play In Computer-Assisted Molecular Design. *FASEB J.* **1995**, *9*, A1253−A1253.

(11) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(12) Ghose, A. K.; Crippen, G. M. Use of physicochemical parameters in distance geometry and related three-dimensional quantitative structure−activity relationships: a demonstration using *Escherichia coli* dihydrofolate reductase inhibitors. *J. Med. Chem.* **1985**, *28*, 333−46.

(13) Ghose, A. K.; Crippen, G. M.; Revankar, G. R.; McKernan, P. A.; Smee, D. F.; Robins, R. K. Analysis of the in vitro antiviral activity of certain ribonucleosides against parainfluenza virus using a novel computer aided receptor modeling procedure. *J. Med. Chem.* **1989**, *32*, 746−756.

(14) Ghose, A. K.; Logan, M. E.; Treasurywala, A. M.; Wang, H.; Wahl, R. C.; Tomczuk, B.; Gowravaram, M.; Jaeger, E. P.; Wendoloski, J. J. Determination of Pharmacophoric Geometry for Collagenase Inhibitors Using a Novel Computational Method and Its Verification Using Molecular Dynamics, NMR and X-ray Crystallography. *J. Am. Chem. Soc.* **1995**, *117*, 4671−4682.

(15) Ghose, A. K.; Wendoloski, J. J. Pharmacophore Modeling: Methods, Experimental Verifications and Applications. In *3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity*; Kubinyi, H., Folker, G., Martin, Y. C., Eds.; Kluwer Academic: The Netherlands, 1998.

(16) Hansch, C. Quantitative Structure−Activity Relationships and the Unnamed Science. *Acc. Chem. Res.* **1993**, *26*, 147−153.

(17) Martin, Y. C. 3D QSAR: Current State, Scope and Limitations. In *3D QSAR in Drug Design: Recent Advances*; Kubinyi, H., Flokers, G., Martin, Y. C., Eds.; Kluwer/Escom: Dordrecht, 1998; Vol. 3, pp 3−23.

(18) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA) 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(19) Koltermann, A.; Kettling, U.; Bieschke, J.; Winkler, T.; Eigen, M. Rapid assay processing by integration of dual-color fluorescence cross-validation spectroscopy: high throughput screening for enzyme activity. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1421−1426.

(20) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. I. Molecular Frameworks. *J. Med. Chem.* **1996**, *37*, 2887−2893.

(21) McGregor, M. J.; Pallai, P. V. Clustering Large Databases of Compounds: Using MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(22) 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.

(23) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, Y.; Masushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.

(24) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. Prediction of Hydrophobic Properties of Small Organic Molecules Using Fragmental methods: An Analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.

(25) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(26) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.
(27) *C&H Dictionary of Pharmaceutical Agents*; available as a 3D UNITY database from Tripos, Inc.: St. Louis, MO, 1998.
(28) *Integrated Scientific Information System* (ISIS); available from MDL Information Systems, Inc.: San Leandro, CA, 1997.
(29) Craig, P. N. Drug Compendium. In *Comprehensive Medicinal Chemistry*; Hansch, C., Sammes, P. G., Taylor, J. B., Drayton, C. J., Eds.; Pergamon Press: Oxford, 1989; Vol. 6, p 237.
(30) Clark, T. *A Handbook of Computational Chemistry: A Practical Guide to Chemical Structure and Energy Calculations*; John Wiley: New York, 1985.
(31) *Galaxy v. 2.5*; available from AM Technologies, Inc.: San Antonio, TX, 1997.
(32) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure−activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.
(33) Mood, M. A.; Graybill, F. A.; Boes, D. C. *Introduction to the Theory of Statistics*; McGraw-Hill: New York, 1974.
(34) Hansch, C. In *Correlation Analysis in Chemistry*; Chapman, N. B., Shorter, J., Eds.; Wiley: New York, 1978.
(35) Hansch, C.; Leo, A. J. *Substituent Constants for Correlation Analysis in Chemistry*; Wiley: New York, 1979.
(36) Viswanadhan, V. N.; Ghose, A. K.; Weinstein, J. N. Mapping the binding site of the nucleoside transporter protein: a 3D-OSAR study. *Biochim. Biophys. Acta* **1990**, *1039*, 356−66.
(37) Viswanadhan, V. N.; Ghose, A. K.; Hanna, N. B.; Matsumoto, S. S.; Avery, T. L.; Revankar, G. R.; Robins, R. K. Analysis of the in vitro antitumor activity of novel purine-6-sulfenamide, -sulfinamide, and -sulfonamide nucleosides and certain related compounds using a computer-aided receptor modeling procedure. *J. Med. Chem.* **1991**, *34*, 526−32.
(38) Macdonald, C. M.; Turcan, R. G. Sites of Drug Metabolism, Prodrugs and Bioactivation. In *Comprehensive Medicinal Chemistry*; Hansch, C., Sammes, P. G., Taylor, J. B., Ramsden, C. A., Eds.; Pergamon Press: London, 1990; Vol. 5, pp 111−138.
(39) Gaillot, J.; Bruno, R.; Montay, G. Distribution and Clearance Concept. In *Comprehensive Medicinal Chemistry*; Hansch, C., Sammes, P. G., Taylor, J. B., Eds.; Pergamon Press: Oxford, 1990; Vol. 5, pp 71−109.
(40) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Trends* **1997**, *2*, 384−386.
(41) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

CC9800071