ACS Combinatorial Science

RESEARCH ARTICLE

pubs.acs.org/acscombsci

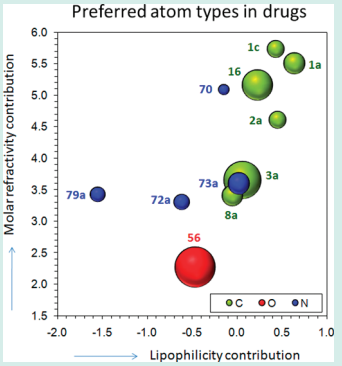
Atom Type Preferences, Structural Diversity, and Property Profiles of Known Drugs, Leads, and Nondrugs: A Comparative Assessment

Vellarkad N. Viswanadhan,* Hariharan Rajesh, and Vitukudi N. Balaji

Department of Computational Chemistry, Jubilant Biosys Limited, #96, Industrial Suburb, second Stage, Yeshwanthpur, Bangalore 560 064, India

Supporting Information

ABSTRACT: A new characterization of known drug, lead, and representative nondrug databases was performed taking into account several properties at the atomic and molecular levels. This characterization included atom type preferences, *intrinsic* structural diversity (Atom Type Diversity, ATD), and other well-known physicochemical properties, as an approach for rapid assessment of druglikeness for small molecule libraries. To characterize ATD, an elaborate united atom classification, UALOGP (United Atom Log P), with 148 atom types, was developed along with associated atomic physicochemical parameters. This classification also enabled an analysis of atom type and physicochemical property distributions (for calculated log *P*, molar refractivity, molecular weight, total atom count, and ATD) of drug, lead, and nondrug databases, a reassessment of the Ro5 (Rule of Five) and GVW (Ghose−Viswanadhan−Wendoloski) criteria, and development of new criteria and ranges more accurately reflecting the chemical space occupied by small molecule drugs. A relative druglikeness parameter was defined for atom types in drugs, identifying the most preferred types. The present work demonstrates that drug molecules are constitutionally more diverse relative to nondrugs, while being less diverse than leads.



KEYWORDS: druglikeness, drug properties, atom classification, ALOGP, AMR, lipophilicity, atom type diversity

1. INTRODUCTION

Since the seminal publications of Lipinski and co-workers,[1] library design and lead optimization efforts in small molecule drug discovery have included druglikeness considerations from the hit stage to candidate selection.[2−21] The objective of this work was to systematically analyze druglikeness at the atomic and molecular levels, including atom type preferences and *intrinsic* structural diversity or Atom Type Diversity (ATD), of drugs, leads, and nondrugs (commercially available small molecule collections). Lipinski's Rule of 5 (Ro5),[1] Ghose−Viswanadhan−Wendoloski (GVW) criteria,[3,4] and other drug/nondrug comparisons[7,8] were reassessed. Structural properties examined in detail include ATD, calculated log *P* (based on ALOGP method, ALOGP98),[22] molar refractivity (AMR89),[23] atom counts, and molecular weight, along with several others. Intrinsic structural diversity, also called atom type diversity (ATD), is shown to be a key differentiator of drugs and leads relative to nondrugs. A new and elaborate atom type representation, UALOGP (United Atom Log P) derived from ALOGP method,[22,23] is described along with corresponding atomic physicochemical parameters for a detailed characterization of the property ranges, constitutional makeup, and structural diversity of drugs.

2. MATERIALS AND METHODS

(a). Methodology. The present work describes the development of atomic and whole molecule parameters for characterizing and comparing the molecular databases. At the whole molecule level, ALOGP98 (calculated log *P*),[22] AMR89 (calculated molar refractivity),[23] NATS (number of atoms), MW (Molecular Weight), and a new parameter, ATD (atom type diversity), were calculated, along with a number of other descriptors[8] for chemical space characterization. At the atomic level, considering classification systems for physicochemical property estimation,[22−24] we created a united atom representation, United Atom Log P (UALOGP see Table 1) with 148 atom types, by extending the ALOGP representation[22,23] (113 types including hydrogens), thus enabling more elaborate definitions to reflect subtle differences among similar atom types, while de-emphasizing less important hydrogen types for characterizing druglikeness.

For all databases, several parameters were calculated from the property distributions: the mean, the standard deviation, the ranges of each property covering 95%, 80%, and 50% of the molecules in a database. At the atomic level, the following parameters were calculated based on the calculated atom type distribution in each database: the total frequency, the number of different compounds containing the type, % occurrence, mean occurrence per molecule, and its corresponding standard deviation.

A relative druglikeness parameter (RDP_i) was defined for each atom type, *i*, based on atom type distributions of drug and nondrug databases:

$$RDP_i = p_{i,d}/p_{i,n} \quad (1)$$

Received: January 26, 2011
Revised: April 10, 2011
Published: April 11, 2011

Table 1. Physico-Chemical Parameters (Group Contributions for log *P* and Molar Refractivity (MR) for the UALOGP Atom Types)[a,b]

	UALOGP atom type description	UALOGP group contribution log *P*	UALOGP group contribution MR
	Type 1:CH_3R,CH_4		
1a	C^0sp^3 having no X attached to next C	0.6420	5.5021
1b	α − C	0.0099	5.4244
1c	C^0sp^3, having 1 X attached to next C	0.4377	5.7325
1d	C^0sp^3, having 2 X attached to next C	0.0513	5.8987
1e	C^0sp^3, having 3 X attached to next C	0.3411	5.2783
	Type: 2 CH_2R_2		
2a	C^0sp^3 having no X attached to next C	0.4562	4.6010
2b	α − C	0.0348	4.5492
2c	C^0sp^3, having 1 X attached to next C	0.3200	4.7546
2d	C^0sp^3, having 2 X attached to next C	0.0624	4.8654
2e	C^0sp^3, having 3 X attached to next C	0.2556	4.4518
2f	C^0sp^3, having 4 X attached to next C	−0.2880	
	Type 3: CHR_3		
3a	C^0sp^3 having no X attached to next C	0.0660	3.6475
3b	α − C	−0.1447	3.6216
3c	C^0sp^3, having 1 X attached to next C	−0.0021	3.7243
3d	C^0sp^3, having 2 X attached to next C	−0.1309	3.7797
3e	C^0sp^3, having 3 X attached to next C	−0.0343	3.5729
3f	C^0sp^3, having 4 X attached to next C	−0.3061	
	Type 5: CH_3X		
5	C^1sp^3, C^0sp^2	0.1023	5.6967
	Type 6: CH_2RX		
6a	C^1sp^3, C^0sp^2	0.0116	4.7122
6b	α − C	−0.2018	4.5620
	Type 7: CH_2X_2		
7	Any of C^2sp^3, C^1sp^2, C^0sp	0.0055	4.2339
	Type 8: CHR_2X		
8a	C^1sp3, C^0sp^2	−0.0504	3.3979
8b	α − C	−0.1571	3.3228
8c	C^0sp^3, having 1 X attached to next C	−0.0139	3.4255
8d	C^0sp^3, having 2 X attached to next C	−0.1433	3.4809
8e	C^0sp^3, having 3 X attached to next C	−0.0467	3.2741
8f	C^0sp^3, having 4 X attached to next C	−0.3185	
	Type 9: $CHRX_2$		
9a	C^2sp^3, C^1sp^2, C^0sp	0.1322	3.1775
9b	attached H's when C is α − C	0.1576	3.1958
	Type 10: CHX_3		
10	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.7184	3.3879
	Type 15: $=CH_2$		
15a	C^1sp^3, C^0sp^2	0.4736	4.9879
15b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	−0.8237	4.8641
	Type 16: =CHR		
16a	C^1sp^3, C^0sp^2	0.2339	5.1593
16b	attached H's when C is an α − C	0.1272	5.0842
16c	C^0sp^3, having 1 X attached to next C	0.2698	5.1869
16d	C^0sp^3, having 2 X attached to next C	0.1410	5.2423
16e	C^0sp^3, having 3 X attached to next C	0.2376	5.0355
	Type 18: =CHX		
18a	C^2sp^3, C^1sp^2, C^0sp	−0.2871	4.4010
18b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	−0.8422	4.4325
	Type 21: ≡CH		
21a	C^2sp^3, C^1sp^2, C^0sp	0.9877	4.2830
21b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.4326	4.3145
	Type 24: R- -CH- -R		
24a	C^1sp^3, C^0sp^2	0.3050	4.3433
24b	attached H's when C is α − C	0.1983	4.2682
	Type 27: R- -CH- -X		
27a	C^2sp^3, C^1sp^2, C^0sp	0.5785	3.3014
27b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	−0.0366	3.3329
27c	attached H's when C is α − C	0.5239	3.3197
	Type 30: X- -CH- -X		
30	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.1443	3.3329
	Type 33: X- -CH- -X		
33a	C^2sp^3, C^1sp^2, C^0sp	1.0322	4.2676
33b	attached H's when C is α − C	0.1511	4.2859
	Type 36: Al-CH=X		
36a	C^2sp^3, C^1sp^2, C^0sp	0.3860	4.8195

dx.doi.org/10.1021/co2000168 |*ACS Comb. Sci.* 2011, 13, 327–336

Table 1. Continued

	UALOGP atom type description	UALOGP group contribution log *P*	UALOGP group contribution MR
36b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.1691	4.8510
36c	α − C	0.3914	4.8378
	Type 37: Ar-CH═		
37a	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	−0.0615	5.6090
37b	α − C	0.4990	5.5958
	Type 42: X- -CH···X		
42	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	−0.2518	3.6104
	Types 56, 57: −OH		
56	aliphatic −OH	−0.4603	2.2646
57	phenol, enol, carboxyl OH	−0.1163	2.2778
	Type 64: Se-any-Se		
64	-SeH	0.5565	11.9366
	Type 66: Al-NH_2		
66	-NH_2	−0.7499	4.2221
	Type 67: Al_2NH		
67	-NH	−0.4204	3.3000
	Type 69: Ar-NH_2, X-NH_2		
69	-NH_2	−0.5955	5.2841
	Type 70: Ar-NH-Al		
70	H attached to heteroatom	−0.1425	
	Type 72: RCO-N, N-X═X		
72	H attached to heteroatom	−0.6149	3.3000
	Type 73: Ar_2NH, Ar_3N		
73	H attached to heteroatom	0.0223	3.5956
	Type 74: R≡N, R═N-		
74	H attached to heteroatom	0.0313	3.5000
	Type 79: N^+		
79	H attached to heteroatom	−1.5475	
	Type 106: R-SH		
106	H attached to heteroatom	0.5110	8.6916
	Type 118: PX_3 (phosphite)		
118	phosphite	−0.9002	
	Type 119: PX_3 (phosphine)		
119	H attached to heteroatom	0.5669	

[a] The original ALOGP atom types[22] are identified, for each corresponding subset of UALOGP types. [b] R represents any group linked through carbon; X represents any heteroatom (O, N, S, P, Se, and halogens); Al and Ar represent aliphatic and aromatic groups, respectively; "═" represents a double bond; "≡" represents a triple bond; "- -" represents a aromatic bonds as in benzene or delocalized bonds such as the N−O bond in a nitro group; "·· " represents aromatic single bonds as the C−N bond in pyrrole. The C- -N bond order in pyridine may be considered as 2 while we have one such bond and 1.5 when we have two such bonds.

where *i* refers to the atom type, $p_{i,d}$ is the % occurrence of type *i* in the drug database and $p_{i,n}$ is the % occurrence of type *i* for the nondrug database (an approximation for expectation value, based on a typical distribution in commercial small molecule collections). Values of $RDP_i > 1$ indicate preferred types in drugs, while values <1 indicate the opposite.

(b). Molecular Databases. For the present analysis, three sources of drug molecules were considered: (i) a database of FDA approved drugs, available as part of ZINC databases;[25] (ii) the DrugBank,[26] a database of nearly 4800 entries including over 1,350 FDA-approved small molecule drugs, 123 FDA-approved biotech (protein/peptide) drugs, and 71 nutraceuticals and over 3,243 experimental drugs; and (iii) JBLDrugDB,[42] a small, well-annotated proprietary database of FDA approved drugs. These databases were filtered to exclude highly lipophilic (calculated log $P > 8.0$) and highly hydrophilic compounds (calculated log $P < -5.0$), similar to earlier analyses.[3,4] We also excluded compounds such as nutraceuticals and protein/peptide drugs from our analysis, as well unusually small (<100 MW or <14 atoms or <10 heavy atoms) or large (>800 MW or >100 atoms) molecules. Also removed were polymers, peptides, quaternary ammonium, multiple acids, and phosphates. This additional filtering excluded entries not of particular interest as small molecule drugs.

Two sets of 470 lead-drug pairs generated by Hann and coworkers[28] were also analyzed to delineate differences between lead molecules and the corresponding drugs. The lead database was not subjected to filtering. The present work also required an analysis of a "nondrug" database for an atomic level druglikeness assessment. Of several commercially available compound databases, we chose the Chembridge[27] database, which is quite large (the latest version contains over 700,000 compounds, which includes both target class and general libraries) and extensively used for screening in the pharmaceutical industry. Two independent random sets of 10000 molecules each were picked from an earlier version of the data set and subjected the data set to the filtering procedure mentioned in the methodology. This was necessary to reduce the size of nondrug data set for practical purposes. This marginally reduced the number of compounds to 9969 and 9965 for the two sets. One of the two sets was used as the nondrug data set. We also analyzed the drug and nondrug databases employed by Hutter[7] for comparison.

(c). Atom Types and Atomic Properties. The new united atom type representation (UALOGP) introduced here was derived from the ALOGP representation,[22,23] with greater elaboration and distinctions of heavy atom types, grouped with bonded hydrogen atoms. This led to a total of 148 heavy atom types, based on the number and types of added hydrogens for each heavy atom. For example, a methyl carbon (ALOGP type 1) belongs to one of five united atom types (1a−1e), which were created based on five different types of attached hydrogens. Shown in Table 1 are UALOGP group contributions derived from atomic constants for lipophilicity (log *P*) and molar refractivity (MR) developed by Ghose et al[22] and Viswanadhan et al.[23]

(d). Atom Type Diversity. Characterization of intrinsic structural diversity was based on the concept that atom classification is hierarchical, with elemental types at the primary level. ALOGP[22] and UALOGP classifications constituted secondary and tertiary levels of finer differentiation. Hydrogen atoms were considered implicit, and only heavy atoms were used for assessment. On the basis of these, three structural diversity measures were defined for a molecule with NHATS heavy atoms.

$$P_1 = \text{Number of element types/NHATS} \quad (2)$$

dx.doi.org/10.1021/co2000168 |ACS Comb. Sci. 2011, 13, 327–336

Table 2. Average Physico-Chemical Properties:[a] Molecular Weight (MW), Calculated log *P* (ALOGP98), Calculated Molar Refractivity (AMR89), Atom Type Diversity (ATD), and Number of Atoms (NATS) Identified from Several Drug Databases, a Lead Database, and a Nondrug Database

database		number of compounds	MW	ALOGP98	AMR89	ATD	NATS
1. Drugs from Leads		404	314.4(87.3)	2.6(1.6)	86.4(22.9)	5.06(3.32)	42.9(11.8)
1a	charged subset	214	322.1(77.0)	3.0(1.5)	92.4(20.3)	4.32(2.33)	47.2(10.0)
1b	neutral subset	190	305.8(97.1)	2.2(1.6)	80.0(24.0)	5.88(4.00)	38.0(12.2)
2. Leads		206	272.0(90.6)	2.0(1.8)	75.2 (24.4)	6.49(5.11)	37.5(12.1)
2a	charged subset	100	274.7(86.1)	2.6(1.7)	80.7(24.2)	5.59(4.67)	41.5(12.0)
2b	neutral subset	106	269.3(95.0)	1.5(1.7)	70.0(23.5)	7.34(5.38)	33.8(11.1)
3. JBLDrugDB		777	331.1(105.6)	2.4(1.9)	88.6(25.7)	5.33(4.06)	44.1(14.2)
3a	charged subset	318	320.9(89.9)	2.5(1.8)	90.1(23.4)	5.13(3.18)	46.0(11.8)
3b	neutral subset	459	338.1(114.7)	2.3(1.9)	87.7(27.1)	5.46(4.57)	42.8(15.5)
4. Drug Bank		2380	315.4 (107.5)	2.0(2.2)	83.9(29.5)	5.98(4.73)	41.1(14.9)
4a	charged subset	899	318.0(103.6)	2.0(2.2)	87.9(29.3)	5.94(4.68)	44.4(14.0)
4b	neutral subset	1481	313.8(109.7)	1.9(2.2)	81.4(29.3)	6.01(4.76)	39.2(15.1)
5. FDA approved drugs from ZINC		943	322.4(101.9)	2.3(1.9)	86.8(25.8)	5.35(4.05)	43.0(13.8)
5a	charged subset	407	326.9(91.8)	2.6(2.0)	82.8(26.6)	4.86(3.19)	46.5(11.8)
5b	neutral subset	536	319.0(109.0)	2.1(1.9)	81.4(26.6)	5.73(4.57)	40.2(14.6)
6. Non_Drugs		9969	403.2(59.6)	3.1(1.2)	113.7(17.1)	3.59(1.90)	56.6(9.0)
6a	charged subset	5012	400.3(58.7)	3.2(1.2)	114.6(16.7)	3.50(1.82)	58.8(8.65)
6b	neutral subset	4957	406.3(60.5)	3.0(1.2)	112.7(17.4)	3.68(1.98)	54.4(8.9)
7. Drugs_All		2880	318.7(107.8)	2.0(2.1)	84.6(28.8)	5.78(4.53)	41.8(14.9)
7a	charged subset	1058	319.0(99.8)	2.1(2.1)	88.3(28.0)	5.77(4.45)	44.8(13.4)
7b	neutral subset	1822	318.5(112.2)	1.9(2.1)	82.4(29.1)	5.78(4.57)	40.0(15.3)

[a] Standard deviation in parentheses.

$$P_2 = \text{Number of heavy atom types/NHATS} \quad (3)$$

$$P_3 = \text{Number of united atom types/NHATS} \quad (4)$$

Here, the *P*'s define the atom type diversity based on elemental types (eq 2, P_1), ALOGP[22] heavy atom types (eq 3, P_2), and united atom (UALOGP) types (eq 4, P_3). Atom type diversity (ATD) was defined as the product of P_1, P_2, and P_3, times 100 (a scale factor).

$$\text{ATD} = P_1 \cdot P_2 \cdot P_3 \times 10^2 \quad (5)$$

This definition ensured equal weight to each level of atom classification.

3. RESULTS AND DISCUSSION

The distributions of physicochemical properties in drug databases had been the subject of several studies.[3−5,28−31] Recent studies identified distinct physicochemical profiles of different drug classes, such as respiratory drugs, marketed versus development drugs.[29−31] Previous attempts at quantifying druglikeness using linear and nonlinear approaches[4−15,17,18,20,21] employed a multitude of molecular characteristics. These included, for example, atom pair frequencies,[7] whole molecule properties,[8] ALOGP atom types,[9] and ISIS keys[13,32] with the goal of computing druglikeness metric for a given molecule. These may be regarded as simplified representations of a multidimensional druglikeness concept encompassing a broad range of properties to differentiate drugs from nondrugs (e.g., refs 7−13,20). While linear approaches (e.g., refs 1−3,20) have the advantage of transparency, nonlinear approaches (e.g., refs 9,10) are likely to be more effective, as they can model descriptor interdependencies.

To identify distinct attributes of drugs at atomic and molecular levels, a comprehensive assessment of structural diversity and property distributions of drugs, leads, and nondrugs was performed, re-examining physicochemical property ranges and distributions for druglikeness proposed earlier.[2−5]

UALOGP Atom Classification and Its Validation. The ALOGP atom classification was originally proposed by Ghose and Crippen[33] to quantify molecular lipophilicity (as assessed by 1-octanol to water partition coefficient - log *P*) and hydrophobic interactions in QSAR studies. Extensive revisions[22,23] of atomic parameters were shown to successfully predict log *P* and molar refractivity of organic molecules and were used for molecular fingerprinting and QSAR studies (e.g., refs 35−39). ALOGP atom types were employed for characterizing druglikeness,[9] and deriving GVW criteria.[3,4]

In the context of the current, more elaborate UALOGP system, it is pertinent to consider Leo's critique of the ALOGP types[34] and statistically derived atomic lipophilicity values. Leo[34] noted that the fitting procedure yields negative atomic lipophilicity values for carbons and relatively large, positive values for attached hydrogens, which Leo considered contrary to intuitive expectations. This, however, would not be true for the current UALOGP representation. He noted that atomic lipophilicity values in the ALOGP98[22] version differ significantly from the older version,[23] for carbons and hydrogens. For example, rms deviation for ALOGP lipophilicity parameters from the two publications of 1989[23] and 1998[22] was computed to be 0.42 (for the atom types in Table 1). However, when these two sets of ALOGP parameters were used to derive and compare UALOGP parameters, the corresponding rms deviation dropped to 0.05 because of

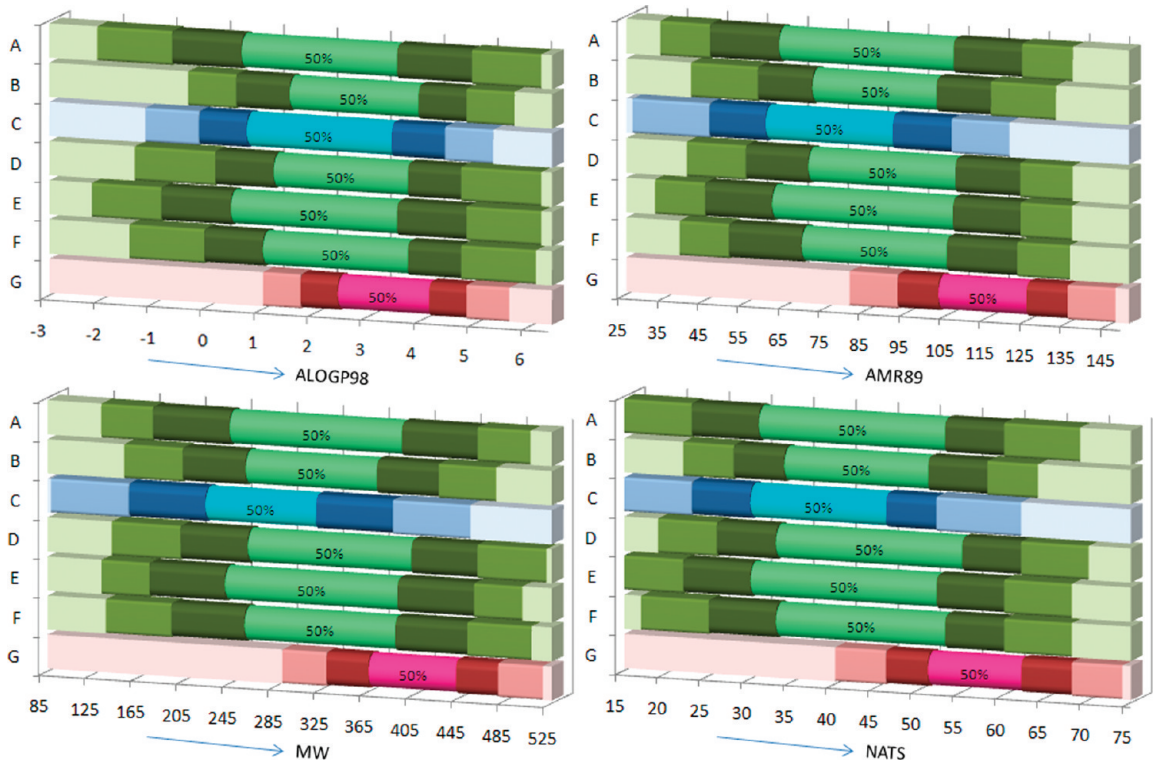


Figure 1. (a) ALOGP98, (b) AMR89, (c) MW, and (d) NATS ranges covering different fractions of each database. The middle bright colored part covers 50% of each database. Dark colored extensions on either side constitute another 30%, covering 80% range, and another 15% is added by further light colored extensions, covering 95% range. Databases shown are (A) Drugs_All, (B) Drugs from leads, (C) leads, (D) JBLDrugDB, (E) DrugBank, (F) FDA approved drugs, and (G) nondrugs.

smaller differences between the sets, though with a greater number of parameters. This confirms that the UALOGP parameters shown in Table 1 are close to convergence and addresses Leo's point.

Whole Molecule Properties. Table 2 shows the physicochemical property ranges for different sets of drugs, leads, and nondrugs, described earlier. We evaluated the ranges separately for charged and neutral molecules, an aspect known to be significant,[2,40] but not considered explicitly in previous assessments.[1,3,7] Contribution to drug absorption from charged form is significant, and it is often preferable to replace or remove charged groups in a lead molecule to improve PK properties such as hERG activity or to improve brain penetration.[40] The log *P* value is, by definition, associated with the neutral form of a molecule, though the charged form could be relevant under physiological conditions. For log *P* (ALOGP98)[22] calculations we used the neutral form. For example, −COOH oxygens were assigned the types 58 (=O) and 61 (−OH). For all other calculations, the representation (and hydrogen addition or deletion for charged groups) reflects the net charge as calculated by the EPIK[41] module of the Schrodinger software. Thus, type 62 was assigned for oxygens of $-COO^-$ and type 79 for nitrogen of $-NH3^+$. Table 2 shows the average calculated values of log *P* (ALOGP98), molar refractivity (AMR89), molecular weight (MW), number of atoms (NATS), and atom type diversity (ATD) for different databases and their subsets.

For nondrugs the average values of ALOGP98, AMR89, molecular weight, and number of atoms are higher than for drugs. The average difference in calculated log *P* (ALOGP98) between drugs (Drugs_All) and nondrugs is 1.1 units (Table 2). Atom type diversity (ATD) is, on the average, higher for drugs (5.78 for Drugs_All vs 3.59 nondrugs). Table 2 also shows the property differences between drugs and leads from the drug-lead pair database. On the average, drugs are heavier (ΔMW ~ 42, ΔNATS ~ 5), bulkier (ΔAMR89 ~ 11), and more lipophilic (ΔALOGP98 ~ 0.6) relative to leads. These data shown here are largely consistent with the conclusions of Hann et al,[28] but have some differences with those of Oprea et al.[43] and Rishton.[44] Interestingly, leads have greater atom type diversity (ATD ~ 6.5 for leads vs 5.78 for drugs). The database of leads considered here is relatively small. Nevertheless, this analysis suggests that, on the average, lead optimization process adds greater bulk but lowers atom type diversity.

Comparing the data for charged and uncharged drugs shows interesting differences. For FDA approved drugs, MW, ALOGP98, AMR89, and NATS differ by 8 Da, 0.5 log units, 1.4 units, and 6.3 atoms, respectively, for charged and uncharged molecules. As charged lead molecules are hydrophilic, they offer greater freedom to the medicinal chemist to add hydrophobic functionalities to improve potency, while maintaining a low log *D*. Not surprisingly, the corresponding property differences for charged versus neutrals in the case of a nondrug database are insignificant. For drugs from drug-lead pairs, the differences in MW, ALOGP98, AMR89, and NATS between charged and neutral molecules are 16 Da, 0.8 log units, 12 units, and 9, respectively, whereas for leads it is 5 Da, 1.1 log units, 11 units, and 8, respectively. Thus, the observed differences between charged and uncharged subsets appear consistent across all drug databases considered.

Figure 1 graphically shows the property ranges of ALOGP98, AMR89, MW, and NATS for different databases. The numerical values of these ranges are given in Supporting Information, Table S1. The ranges (at 80% and 50% coverage) for AMR, MW, and

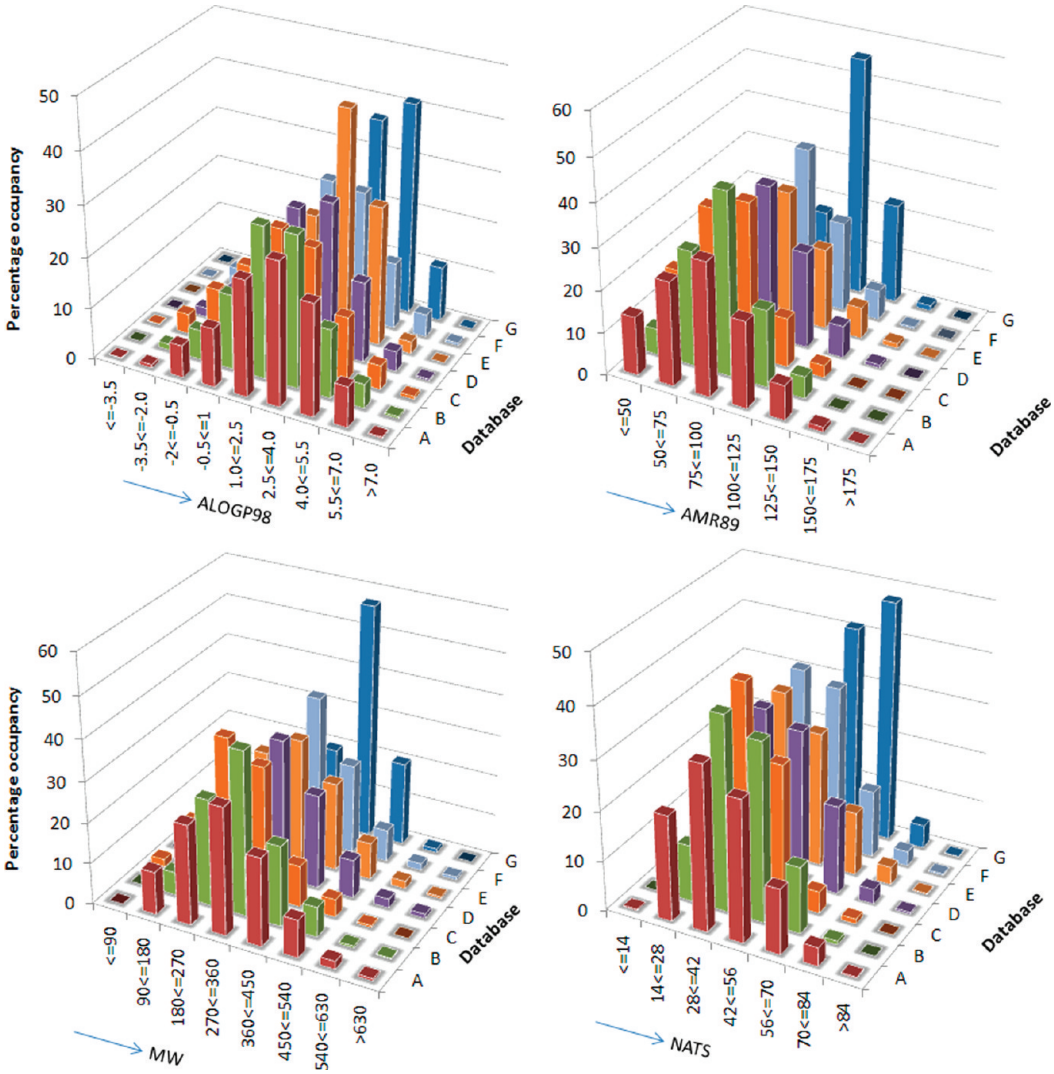


Figure 2. Histogram distributions of (a) ALOGP98, (b) AMR89, (c) MW, and (d) NATS for the six databases: A, (magenta) Drugs_All; B, (green) Drugs from Leads; C, (light brown) leads; D, (purple) JBLDrugDB; E, (dark brown) DrugBank; F, (light blue) FDA approved drugs, and G, (dark blue) nondrugs.

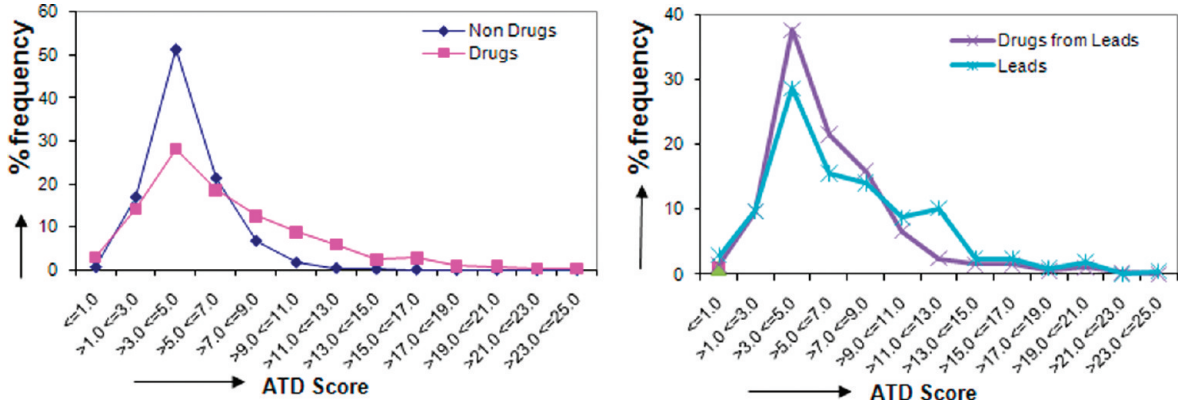
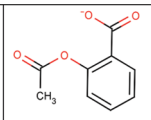
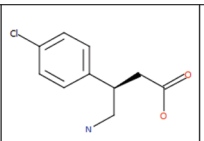
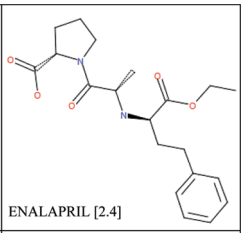
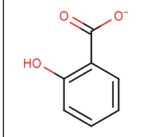
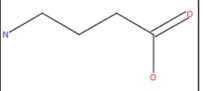
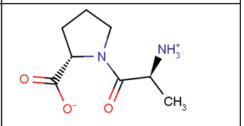


Figure 3. Distribution of ATD scores for (a) drugs (Drugs_All) and nondrugs (b) drug and lead pairs used in this study.

NATS (Figure 1) agree reasonably well with the previously reported ranges.[3] We see the 80% range for ALOGP98 is −0.8 to 4.7 (consistent with Ro5[1]), lower than the previously reported range[3] of −0.4 to 5.6. The 50% range for ALOGP98 shows a larger difference, with preferred range left-shifted by 0.5. The present analysis also shows that the ranges based on 95% coverage

dx.doi.org/10.1021/co2000168 |*ACS Comb. Sci.* 2011, 13, 327–336

Table 3. Examples of Drug-Lead Pairs[a]

Drugs from leads	 ASPIRIN [5.8]	 BACLOFEN [14.6]	 ENALAPRIL [2.4]
Leads	 SALICILIC_ACID [7.2]	 GABA [26.2]	 ALAPRO [11.1]

[a] ATD scores in parentheses.

are appropriate for a properly filtered database, so as not to exclude most of the known small molecule drugs. Use of 95% coverage of known drugs extends previously reported ranges.[3,4] Thus, the 95% ALOGP98 range is −2.2 to 6.1 (for Drugs_All and other drug databases), consistent with the elliptical filter proposed by Egan et al,[40] while the Ro5[1] has an upper limit of 5.0 for log *P*, which excludes a good number of orally absorbed known drugs. The absence of lower limit for log *P* in Ro5 may cause the inclusion of highly hydrophilic material as druglike. For nondrug databases, the calculated ranges necessarily depend on their sources. Thus, Hutter's[7] nondrug database has a different property distribution (Supporting Information, Figure S1).

Histogram plots of the distribution for these physicochemical properties are shown in Figure 2. Though the ALOGP98 distribution of nondrug database is different from other drug or lead databases, the range for the drug database subsumes the narrower nondrug range. On the whole, the drug database ranges are larger and strongly overlap the nondrug ranges. The peak of the ALOGP98 distribution in Drugs_All database is in the range of 1 to 2.5, though the range of 2.5 to 4 units is almost equally occupied. This contrasts with the peaks in the nondrug database where 44.4% of molecules occupy the 2.5 to 4.0 range. Leads peak around 1 to 2.5 as in Drugs_All, whereas drugs from lead-drug pairs peak in the 2.5 to 4 range (35% of the database), reflecting the addition of lipophilicity during lead optimization. We see a right-shift in the distribution for nondrugs with a greater peak in the log *P* range of 2.5 to 4.0. However, the nondrugs considered here are bulkier (the peak of the AMR89 histogram is in the range of 100 to 125, occupying over 50% of the database).

MW for the drug databases peak at 270−360 and for the lead database, peaks at 180−270 range. In Figure 2 (c), it is seen that the nondrug database has larger molecules, with a peak at the range of 360−450, covered by over half of the database. The histogram plot of NATS (Figure 2d) shows a peak in the range of 28−42 (~33% of molecules) for the drug databases, whereas the current nondrug database peaks in the 56−70 range. An examination of druglikeness of our drug and nondrug databases by the method of Xu and Stevenson[8] indicated that, for the drug database, the druglikeness index (DLI) peaks in the range of 75−85, while most nondrugs have DLI values below 75 (Supporting Information, Figure S2), though the DLI index ranges for the drug and the current nondrug databases overlap strongly. Xu and Stevenson[8] also reported a chemical space diversity analysis of drugs using 22 descriptors. A principal component analysis (PCA) of the same set of descriptors for our databases enabled an analysis of chemical space diversity for both drug and nondrug molecules. Plots of the first 2 PCs (Supporting Information, Tables S2 and S3, Figure S3) shows greater dispersion of nondrugs compared to drugs, though drugs occupy a part of the same chemical space occupied by nondrugs. This shows the need for a better druglikeness index to discriminate effectively drugs and nondrugs, when an appropriately filtered nondrug database is used.

In a pioneering study, Oprea[43] identified the differences in physicochemical characteristics of leads and drugs, which would be useful to identify compounds for lead generation and optimization efforts. Hann and others[28] identified 470 lead drug-pairs. This formed a useful database for the present analysis of the characteristics of leads for comparison with drugs, with the goal of providing a link between lead-like and drug-like chemical space. As part of the current study, we identify differences between leads and drugs, both at the whole molecule and at the atomic levels.

Figure 3 (a) shows the distribution of ATD scores calculated using eq 4 for the drug (Drugs_All) and nondrug databases. The profiles appear distinct for the two classes of molecules, though both distributions peak in the range of 3.0 to 5.0 of the ATD score. On the whole, drugs have higher ATD scores, and 45% of drugs have scores below 5.0. For nondrugs, this percentage is much higher (70%). Scores greater than 7.0 are much rarer for nondrugs (9% for nondrugs vs 36% for drugs). Thus, drugs are significantly richer in atom type diversity. Comparing drugs and corresponding leads (Figure 3b), it is seen that 45% of drugs and 41% of leads have ATD scores below 5.0. Greater than 70% of drug molecules have scores below 7.0 while coverage of leads for the same range is only 57%. This shows leads are constitutionally more diverse compared to drugs.

Some examples of drug-lead pairs are in Table 3. Atom type diversity and other properties of these molecules are given in Supporting Information, Table S4. As leads progress, final drug candidates have more atoms improving potency and optimizing other properties. This usually increases lipophilicity as noted earlier, entailing the addition of groups which decrease atom type diversity. For example, adding a phenyl increases atom count by 6, while lowering atom type diversity.

Analysis Drug Properties at the Atomic Level. For the Drugs_All database, we analyzed atom type distributions and several parameters (Supporting Information, Table S5) using both the united atom and the all atom representations. The parameters evaluated for each atom type are the frequency, percent occurrence, RDP_i (relative druglikeness, eq 1 of the Methods section), mean occurrence (number of atom types per molecule) with standard deviation. We examined the results using three other different random collections of 10,000 compounds each from the Chembridge database, to represent the nondrug database. The results are largely similar for different nondrug sets employed.

Atom type druglikeness (RDP_i) analysis provides insight on those types which are preferred in a drug molecule over a nondrug. Among heavy atoms, atom type 56 (hydroxyl oxygen) has the highest RDP_i value of 7 reflecting its important role as a donor or acceptor in drug molecules. United atom type 3a (C^0_{sp3} carbon having no heteroatom attached to its next C) occurs often among carbon atoms in drugs, with an RDP_i value 6. This saturated carbon type plays the role of bridgehead for many types of rings such as cyclohexane, piperidine, pyrrolidine, which are common

dx.doi.org/10.1021/co2000168 |ACS Comb. Sci. 2011, 13, 327–336

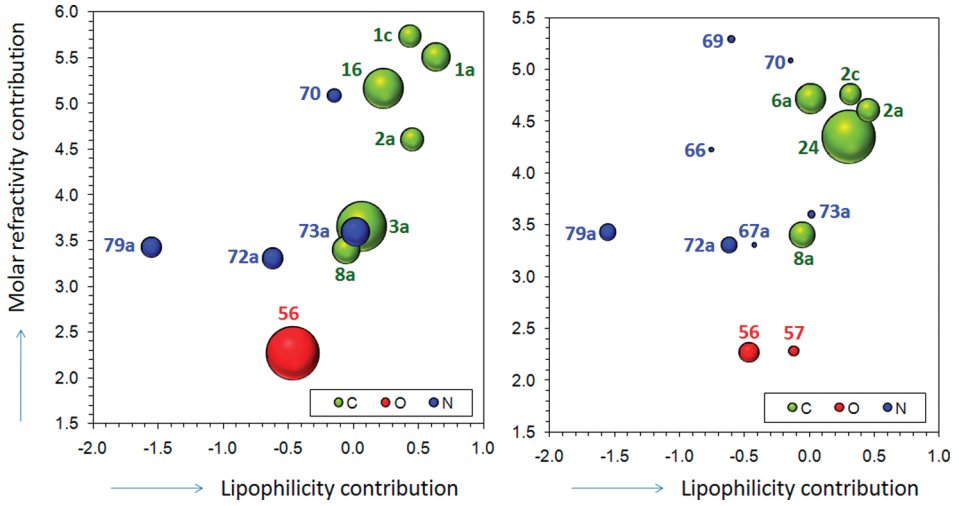


Figure 4. Most frequently occurring united atom types of carbon, oxygen, and nitrogen represented as colored spheres in the chemistry space of UALOGP property values, with atomic/group values of lipophilicity[22] along the *X*-axis and molar refractivity[23] along the *Y*-axis. Sphere size proportional to (a) RDP_i and (b) relative frequency of occurrence in drugs.

among drugs. We note that carbon atoms of type 17, 16, 41 have high RDP_i values of 5, 4, and 4, respectively. This underscores the importance of nonaromatic, unsaturated carbon types in drugs. Carbon atom type 4 (CR_4 type) is rare and modestly preferred in drugs ($RDP_i = 2.5$). Other less preferred carbon atom types are 27, 28, 33, and 34 with a low RDP_i value of 0.3. These represent aromatic carbon types (with or without an attached R-group) next to a heteroatom in the aromatic ring. Interestingly, in drugs 45 atom types are rare with near zero RDP_i value. Among heteroatoms, types 68 and 71 have a low RDP_i value (0.3). These represent trisubstituted nitrogen types with two or three attached aliphatic groups. In drugs, hydrogen attached to heteroatom (ALOGP type 50) is modestly preferred ($RDP_i = 2.5$).

Figure 4 shows atom types with relatively high RDP_i values and those occurring most often among known drugs, in the atomic property space, defined by ALOGP98[22] and AMR89[23] atomic values. Figure 4 shows colored circles for each atom type, proportional in size to (a) RDP_i or (b) occurrence frequency in the drug database. United atom types 3a (among carbon types), 56 (among oxygen types), and 73a and 79 (among nitrogen types) stand out with respect to RDP_i.

The UALOGP classification described here uses 148 types, ensuring adequate consideration of structural diversity in drugs. An earlier analysis by Hutter[7] of drug and nondrug databases used 47 atom types as implemented in the MM+ force field. A comparison of Hutter's results with the present study shows that the choice of the nondrug data set can have a strong influence on derived preferences of atom types in drugs. For example, the MM+ type C4 (ALOGP types 1−14), shows a strong preference for drugs, based on Hutter's[7] analysis and databases. We find that these atom types are common in both drugs and nondrugs. Notably, ~30% of Hutter's nondrug set does not pass our first filters, partly explaining the difference in results. Kutchukian et al[20] analyzed preferences of certain fragments to occur in drug database, using Hutter's data set,[7] which indicated that amides occur in drugs more often and esters in nondrugs. In contrast, our analysis shows that amides are more abundant in nondrugs and esters form a smaller percentage (~10%) of both drugs and nondrugs. These differences with earlier literature demonstrate the significance of using an elaborate atom classification as well as relevant nondrug data sets.

4. CONCLUSION

The present analysis of drug, lead, and nondrug databases reveals preferences of drugs with regard to atom types and their properties. A comparative assessment of various druglikeness criteria is described here, in terms of important physicochemical properties, based on a detailed characterization, that included multiple drug databases and commercially available screening databases to represent nondrugs realistically. The usefulness of a new united atom type representation, UALOGP, along with associated atomic physicochemical parameters, is demonstrated for the assessment of druglikeness. The consensus property ranges derived here, from the current comprehensive drug databases, differ from GVW80, Ro5, and other criteria, and ensure greater coverage (95%) of known drugs, including those from the past decade. Ranges for charged and uncharged drugs were shown to be distinct. Drugs were shown to be structurally more diverse than nondrugs, but less diverse than leads. These new findings were rationalized.

In summary, property preferences at atomic and molecular levels for drugs, leads, and nondrugs are evaluated and presented, to be considered for library design and lead optimization in drug discovery.

■ ASSOCIATED CONTENT

Supporting Information. Physico-chemical property ranges of various databases (Table S1), comparison of drug-lead pairs (Table S2), Descriptors for computing druglike index (Table S3), PCA components of descriptors (Table S4), and Relative druglikeness and other properties of atom types (Table S5). Histogram distributions of ALOGP98, AMR89, MW, and NATS for two sets of drugs and nondrugs (Figure S1), Comparison of DLI distributions for drugs and nondrugs (Figure S2), Scatter plot of PC's for drugs and nondrugs (Figure S3). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +91-80-6662 8908. Fax: +91-80-6662 8333. E-mail: vellarkad_viswanadhan@jubilantbiosys.com.

ACKNOWLEDGMENT

We thank Dr. M. C. Hutter and Dr. M. M. Hann for providing their databases.

REFERENCES

(1) 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.

(2) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature reviews: drug discovery* **2007**, *6*, 881–890.

(3) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. *J. Comb. Chem.* **1999**, *1*, 55–68.

(4) Viswanadhan, V. N.; Ghose, A. K.; Kiselyov, A.; Wendoloski, J. J.; Weinstein, J. N. Knowledge-based approaches for the design of small-molecule libraries for drug discovery. In *Combinatorial Library Design and Evaluation. Software Tools, and Applications in Drug Discovery.*; Ghose, A. K.; Viswanadhan, V. N.; Ed.; Marcel-Dekker, New York, USA, 2001, pp 267−289.

(5) Viswanadhan, V. N.; Balan, C.; Hulme, C.; Cheetham, J. C.; Sun, Y. Knowledge-based Approaches in the Design and Selection of Compound Libraries for Drug Discovery. *Curr. Opin. in Drug Design and Discovery* **2002**, *5*, 400–406.

(6) Ghose, A. K.; Herbertz, T.; Salvino, J. M.; Mallamo, J. P. Knowledge-based Chemoinformatics Approaches to Drug Discovery. *Drug Discovery Today* **2007**, *11* (23−24), 1107–1114.

(7) Hutter, M. C. Separating Drugs from Nondrugs: A Statistical Approach Using Atom Pair Distributions. *J. Chem. Inf. Model* **2007**, *47*, 186–194.

(8) 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.

(9) Sadowski, J.; Kubinyi, H. A. Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325–3329.

(10) Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert- Salvador, M. T.; Diaz-Villanueva, W.; Castro-Bleda, M. J. Drugs and Nondrugs: An Effective Discrimination with Topological Methods and Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1688–1702.

(11) Givehchi, A.; Schneider, G. Impact of Descriptor Vector Scaling on the Classification of Drugs and Nondrugs with Artificial Neural Networks. *J. Mol. Model* **2004**, *10*, 204–211.

(12) Biswas, D.; Roy, S.; Sen, S. A Simple Approach for Indexing the Oral Druglikeness of a Compound: Discriminating Druglike Compounds from Nondruglike Ones. *J. Chem. Inf. Model* **2006**, *46*, 1394–1401.

(13) 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.

(14) Walters, W. P.; Murcko, A. A.; Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3* (4), 384–387.

(15) eDrugSCAN, an online virtual screening tool, allowing stepwise search for drug-like compounds available at http://service.bioinformatik.uni-saarland.de/edrugscan/.

(16) (a) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. (b) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.

(17) Wang, J.; Ramnarayan, K. Towards Designing Drug-Like Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds. *J. Comb. Chem.* **1999**, *1*, 524–533.

(18) Wagener, M.; van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280–292.

(19) Rishton, G. M. Reactive Compounds and *in vitro* False Positives in HTS. *Drug Discovery Today* **1997**, *2*, 382–384.

(20) Kutchukian, P. S.; Lou, D.; Shakhnovich, E. I. FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space. *J. Chem. Inf. Model.* **2009**, *49* (7), 1630–1642.

(21) Frimurer, T. M.; Bywater, R.; Nærum, L.; Lauritsen, Leif. N.; Brunak, S. Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1315–1324.

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

(23) 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.

(24) Crippen, G. M.; Wildman, S. A. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

(25) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182.

(26) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, Jan 1; 34 (Database issue): D668−D672.

(27) www.chembridge.com.

(28) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.

(29) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A Comparison of Physiochemical Property Profiles of Development and Marketed Oral Drugs. *J. Med. Chem.* **2003**, *46*, 1250–1256.

(30) Ritchie, T. J.; Luscombe, C. N.; Macdonald, S. J. F. Analysis of the Calculated Physicochemical Properties of Respiratory Drugs: Can We Design for Inhaled Drugs Yet? *J. Chem. Inf. Model.* **2009**, *49* (4), 1025–1032.

(31) Lajiness, M. S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 470–477.

(32) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.

(33) (a) 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. (b) 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.

(34) Leo, A. J. The History of the development of CLOGP; available at http://www.daylight.com/meetings/mug98/Leo/clogp_history.html.

(35) Viswanadhan, V. N.; Ghose, A. K.; Weinstein, J. N. Mapping the Binding Site of the Nucleoside Transporter Protein: A 3D-QSAR Study. *Biochim. Biophys. Acta* **1990**, *1039* (3), 356–366.

(36) Viswanadhan, V. N.; Reddy, M. R.; Wlodawer, A.; Varney, M. D.; Weinstein, J. N. An Approach to Rapid Estimation of Relative Binding Affinities of Enzyme Inhibitors: Application to Peptidomimetic Inhibitors of the Human Immunodeficiency Virus Type 1 Protease. *J. Med. Chem.* **1996**, *39*, 705–712.

(37) Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. Prediction of Solvation Free Energies of Small Organic Molecules: Additive-Constitutive Models Based on Molecular Fingerprints and Atomic Constants. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (2), 405–412.

(38) Ghose, A. K.; Viswanadhan, V. N.; Sanghvi, Y. S.; Dee Nord, L.; Willis, R. C.; Revankar, G. R.; Robins, R. K. Structural Mimicry of Adenosine by the Antitumor Agents 4-methoxy- and 4-amino-8-(β-D-ribofuranosylamino)pyrimido[5,4-d]pyrimidine as Viewed by a Molecular Modeling Method. *Proc. Natl. Acad. Sci.* **1989**, *86*, 8242–8246.

(39) Ghose, A. K.; Grippen, 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.

(40) Egan, W. J.; Merz, K. M., Jr.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867–3877.

(41) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* **2007**, *21*, 681–691.

(42) A highly annotated database of approved drugs developed at Jubilant Biosys (www.jubilantbiosys.com).

(43) (a) 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* (5), 1308–1315. (b) Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. Leadlike, Druglike or Publike: How different are they? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 113–119.

(44) Rishton, G. M. Molecular Diversity in the Context of Leadlikeness: Compound Properties that Enable Effective Biochemical Screening. *Curr. Opin. Chem. Biol.* **2008**, *12*, 1–12.