J Comput Aided Mol Des (2007) 21:113–119
DOI 10.1007/s10822-007-9105-3

ORIGINAL PAPER

Lead-like, drug-like or "Pub-like": how different are they?

Tudor I. Oprea · Tharun Kumar Allu · Dan C. Fara · Ramona F. Rad · Lili Ostopovici · Cristian G. Bologa

Received: 27 November 2006 / Accepted: 10 January 2007 / Published online: 28 February 2007

Abstract Academic and industrial research continues to be focused on discovering new classes of compounds based on HTS. Post-HTS analyses need to prioritize compounds that are progressed to chemical probe or lead status. We report trends in probe, lead and drug discovery by examining the following categories of compounds: 385 leads and the 541 drugs that emerged from them; "active" (152) and "inactive" (1488) compounds from the Molecular Libraries Initiative Small Molecule Repository (MLSMR) tested by HTS; "active" (46) and "inactive" (72) compounds from Nature Chemical Biology (NCB) tested by HTS; compounds in the drug development phase (I, II, III and launched), as indexed in MDDR; and medicinal chemistry compounds from WOMBAT, separated into high-activity (5,784 compounds with nanomolar activity or better) and low-activity (30,690 with micromolar activity or less). We examined Molecular weight (MW), molecular complexity, flexibility, the number of hydrogen bond donors and acceptors, LogP—the octanol/water partition coefficient estimated by ClogP and ALOGPS), LogSw (intrinsic water solubility, estimated by ALOGPS) and the number of Rule of five (Ro5) criteria violations. Based on the 50% and 90% distribution moments of the above properties, there were no significant difference between leads of known drugs and "actives" from MLSMR or NCB (chemical probes). "Inactives" from NCB and MLSMR were also found to exhibit similar properties. From these combined sets, we conclude that "Actives" (569 compounds) are less complex, less flexible, and more soluble than drugs (1,651 drugs), and significantly smaller, less complex, less hydrophobic and more soluble than the 5,784 high-activity WOMBAT compounds. These trends indicate that chemical probes are similar to leads with respect to some properties, e.g., complexity, solubility, and hydrophobicity.

Keywords Chemical probes · Combinatorial chemistry · Computer chemistry · Database filtering · Druglike · Drug research · Hydrogen bonds · Lead discovery · Leadlike · MLSCN · MLSMR · Molecular complexity · Property distribution · Pub-like · PubChem · "Rule of 5" test · WOMBAT

Abbreviations

ALOGPS	Program available from vcclab.org, Germany
ClogP	LogP calculated with the Biobyte program
HAC	Number of H-bond acceptors
HDO	Number of H-bond donors
LogP	The logarithm of the octanol-water partition coefficient

Dedicated to Yvonne C. Martin on her 70th birthday.

T. I. Oprea (✉) · T. K. Allu · D. C. Fara · C. G. Bologa
Division of Biocomputing, Department of Biochemistry and Molecular Biology, University of New Mexico School of Medicine, MSC11 6145, Albuquerque, NM 87131, USA
e-mail: TOprea@salud.unm.edu

T. I. Oprea
Sunset Molecular Discovery LLC, 1704 B Llano St., Suite 140, Santa Fe, NM 87505, USA

R. F. Rad · L. Ostopovici
Romanian Academy Institute of Chemistry, Mihai Viteazul nr 24, Timisoara 300223, Romania

MDDR	MDL Drug Data Report
MLI	Molecular Libraries and Imaging initiative
MLSCN	The MLI Screening Centers Network
MLSMR	The MLI Small Molecule Repository
MW	Molecular weight
NCB	Nature Chemical Biology
NIH	National Institutes of Health
RNG	Number of rings
Ro5	Lipinski's Rule of Five
RTB	Number of non-terminal flexible bonds
SMCM	Simple Molecular Complexity Metric
SMILES	Simplified Molecular Input Line Entry Specification
SumNO	Sum of nitrogen and oxygen atoms
TlogP	Tetko's LogP, calculated with ALOGPS
TlogSw	Tetko's logarithm of the (molar) aqueous solubility, calculated with ALOGPS
WOMBAT/WB	WOrld of Molecular BioAcTivity database

The quest for high-quality lead/probe compounds

The NIH (National Institutes of Health) Molecular Libraries and Imaging initiative (MLI) [1] has assembled a small-molecule chemical library known as the MLI Small Molecule Repository, or MLSMR (using ''MLSMR'' in PubChem Substance [2] retrieves this subset). Focused on the early stages of lead discovery, with emphasis on target identification, assay development, biomolecular screening, hit-to-probe analysis, the MLI is expected to derive ''chemical probes'' [1] from (some of) the screened assays. The MLI bridges the cultural divide between the public and private sectors by increasing the availability of small molecules as chemical probes for basic research. These MLI activities are likely to influence the industrial and academic drug discovery enterprise, in particular with respect to the intellectual property aspects, and in some cases are anticipated to yield to lead identification and optimization, followed by clinical trials.

The MLI efforts influence less the large pharmaceutical houses, which advocate the ''megabrand'' concept [3], i.e., focused efforts on drugs that yield significant income, e.g., in excess of 800 million USD [4, 5], instead of marketing a variety of products to smaller markets. Based on often-flawed marketing research [6], this type of non-scientific evaluation has a decisive influence on what compounds are progressed to clinical trials and submitted for approval to regulatory agencies. Some of the scientific aspects of the decision making process, in most of the industrial sector, are based on cheminformatics [7]. Cheminformatics is likely to play a major role in the decision making process of the MLSCN (the MLI Screening Centers Network), which is expected to progress compounds to the ''chemical probe'' status for a variety of targets and, perhaps, for some phenotypic screens.

The cheminformatics-driven process of choosing the appropriate leads and chemical probes is becoming important in both the industrial and academic environment. In this paper, we are revisiting our earlier work [8], where we contrasted some computed properties of high-activity (better than 1 nM) molecules from literature with those of leads. This work is actually inspired by Lipinski's rule of five (Ro5) [9], which sets the upper (ninety percentile) limits as follows: 500 Da for molecular weight (MW); 5 for CLogP, the calculated logarithm of the octanol/water partition coefficient, 5 for the number of hydrogen bond donors (HDO); and 10 for the sum of hydrogen bond acceptors (HAC), respectively. The Ro5 criteria, derived from marketed drugs, are applicable to the selection of candidate drugs for oral delivery, not to leads [10], nor are they applicable to ''pub-like'' compounds, as we conclude later in this text. We suggested more restrictive values for leads [11], as leads tend to exhibit lower molecular complexity [12]. Not surprisingly, such restricted values are also applicable to the active molecules from MLSMR, as the property space for chemical probes is much closer to the one for leads, which is more restrictive than the property space for drugs [13].

Materials and methods

Datasets

To examine current trends in the property space of leads, drugs, and chemical probes, we compiled a list of known leads and drugs, and compared these to compounds of interest for pharmaceutical development, to ''active''/''inactive'' compounds from PubChem [2], as well as to high- and low-activity compounds from WOMBAT, a medicinal chemistry database. The following datasets were investigated:

- The 385 leads and 541 drugs which emerged from these leads, which resulted by combining previously described datasets [10, 12, 14]; in this work, these sets are referred to as *Leads* and *Drugs*
- Compounds in pharmaceutical development, extracted from the MDDR (MDL Drug Data Report) 2005.2 database [15], categorized according to their clinical testing phase, in the following manner:
- 1,147 launched drugs;
- 301 compounds in phase III clinical trials, referred to as *Phase III*;
- 1,047 compounds in phase II clinical trials referred to as *Phase II*;
- 801 compounds in phase I clinical trials, referred to as *Phase I*;
- Compounds of current interest extracted from the PubChem database system [2], categorized according to their source and PubChem activity label, as follows:
- 152 ''actives'' from MLSMR and MLSCN, referred to as *MLSMR Act*;
- 46 ''actives'' from Nature Chemical Biology, tested in MLSCN, referred to as *NCB Act*;
- 1,488 ''inactives'' from MLSMR and MLSCN, referred to as *MLSMR Inact*;
- 72 ''inactives'' from Nature Chemical Biology, tested in MLSCN, referred to as *NCB Inact*;

The above subsets were retrieved from PubChem on August 9, 2006. The ''active'' and ''inactive'' labels are defined by the depositors of the data and were considered as such, without further verification. Preference was given to MLSCN compounds as opposed to other PubChem sources for two reasons: (a) MLSMR compounds were tested under HTS conditions in the MLSCN, which is indicative of the manner in which chemical probes are likely to be identified and (b) biological activities reported in *Nature Chemical Biology* papers were not available for the NCB compound sets.

- Compounds extracted from papers published in mainstream medicinal chemistry journals [16], split in two categories according to the their biological activity:
- 30,690 compounds for which the biological activity is above 1 μM, or below 6 units on the $-\log_{10}$ (activity) scale, on all of the documented literature assays (*WB6*);
- 5,784 compounds for which the biological activity is below 1 nM, or above 9 units on the $-\log_{10}$ (activity) scale, in one of the documented literature assays (*WB9*). Of these, only 127 were launched drugs.

These subsets were extracted from WOMBAT 2006.1, which covers the 1991–2005 papers from Journal of Medicinal Chemistry (77.6%), the 2002–2004 papers from Bioorganic Medicinal Chemistry Letters (15.4%) 2002–2003 and the 2002–2003 papers from Bioorganic Medicinal Chemistry (5.6%), as well as some other journals (1.4%); the percentages in brackets indicating the relative contribution of these journals in the 2006.1 release of WOMBAT. WOMBAT compounds that fall in the mid-range (between 6.01 and 8.99 on the $-\log_{10}$ (activity) scale) were not included in this study because they cannot be distinguished on a bioactivity basis. Overall, their property distribution mimics the distribution of medicinal chemistry compounds [13] (data not shown).

Calculated properties

The property space of the above datasets was compared by following the 50% (median) and 90% (''tail'') distribution moments (except for solubility—see below) of the following (calculated) properties:

- MW (molecular weight);
- Molecular complexity – as monitored by the number of rings (RNG) and by SMCM, or the simple molecular complexity metric, a rule-based system detailed elsewhere [17];
- The number of non-terminal flexible bonds (RTB), as shown in Equation 1:

$$\mathrm{RTB} = N_{\mathrm{nt}} + \sum_i (n_i - 4 - \mathrm{RGB}_i - \mathrm{ShB}_i) \qquad (1)$$

where N_{nt} is the number of non-terminal freely rotatable bonds (but single bonds observed in groups like, e.g., sulfonamides (N–S) or esters (C–O), are excluded); n_i is the number of single bonds in any non-aromatic ring i with 6 or more bonds; RGB_i is the number of rigid bonds in ring i; ShB_i is the number of bonds shared by ring i with any other ring. The substructural SMARTS [18] query used for N_{nt} is as follows: [!$(*=,#*)&!D1&!r3&!r4&!R3&!R4]-@[!$(*=,#*)&!D1&!r3&!r4&!R3&!R4]

- The number of hydrogen bond donors (HDO) based on the following SMARTS query: [!H0;#7,#8;!$([*,–,–2,–3]),!$(*–*=,#*)]
- The number of hydrogen bond acceptors (HAC) based on the following SMARTS query: [!$([#6,F,Cl,Br,I,o,$([#8](-[#6a])–[#6a]),s,nX3,$([#7]C = O),$([#7]–[#6a]),#7v5,#15v5,#16,*+1,*+2,*+3])]

- LogP, the octanol/water partition coefficient [19] estimated by Leo's ClogP [20] and by Tetko's ALOGPS, referred here as TLogP [21];
- LogSw, the intrinsic water solubility [22, 23], as implemented in Tetko's ALOGPS [24], and referred here as TLogSw. For this property, we used the 10%, rather than the 90% distribution moment.
- Finally, the number of Ro5 violations, based on the Ro5 criteria described above, using the original implementation of Lipinski et al. [9]. For example, we used the sum of oxygen and nitrogen atoms (SumNO) as the H-bond acceptor count instead of the above HAC descriptor, and ClogP instead of TLogP.

Our earlier work used the LogP and LogSw estimators from the EPI (Estimation Program Interface) Suite [25], because these were freely available (and hence, data could be reproduced). However, we decided to use ALOGPS and ClogP for the following reasons: ALOGPS is also freely available [26] and has shown reliable performance on drug-like compounds from in-house pharmaceutical databases [27, 28], whereas ClogP is considered by many the standard LogP estimation program.

Results and discussion

Property distribution for MLSMR and NCB actives

As previously mentioned [8], lead structures are, on average, smaller and more polar than the final structures of the marketed drugs, and significantly different from high-activity molecules. The earlier analysis [8] was based on a lower number of compounds: 62 leads and 75 drugs, and a few thousand medicinal chemistry actives and inactives. This was been expanded to 385 leads and 541 drugs, to which we added tens of thousands of literature compounds. Before re-evaluating the trends noted earlier in a significantly larger dataset, we were interested in comparing these leads and drugs with the chemical probes from the set of MLSMR and NCB ''actives'', as tested under HTS conditions in the MLSCN and available from the PubChem system (''pub-like'' compounds).

We found no significant differences in the property distribution of the leads, when compared to the MLSMR and NCB actives, both at the median and 90% statistical moment for MW, RNG, and Ro5—as shown in Table 1. Some differences were observed for HDO (1 donor less in the MLSMR_Act subset), HAC & SumNO (smaller 90% values for NCB actives), RTB, as well as LogP and TLogSw values. However, the NCB subset has roughly 8 times smaller population size when compared to Leads. This limited dataset indicates that, according to LogP and TLogSw, leads are more hydrophobic and less soluble (at the 10% distribution moment) when compared to MLSMR actives. However, the small population size for the ''pub-like'' actives to date does not warrant significant conclusions in this direction. Therefore, further analyses reported below refer to the combined set of 385 Leads, 152 MLSMR actives and 46 NCB actives as the ''Actives'' dataset (a total of 569 unique structures).

Trends in chemical probe discovery

The 569 ''Actives'' were compared with the following sets: The ''Drugs'' set, which combines the 541 drugs resulting from the 385 Leads with the 1,147 MDDR Launched Drugs, to yield 1,651 (unique) Drugs; the WB9 set; Phases I, II and III from MDDR; the ''Inactives'' set, which combined 1488 MLSMR & the 72 NCB inactives for a total of 1,551 unique structures; and, finally, the 30,960 WB6 structures. Median values are compared in Table 2, whereas the 90% distribution moments (except for LogSw, at 10%) are given in Table 3. While they may carry equivalent information, data is presented at two different distribution moments because, for larger population sizes, the ''tail-end'' values (10% and 90%) may differ even when the median values are similar.

Given the *Actives* subset as baseline, it becomes apparent that there are marked increases in MW, complexity and Ro5 violations, when comparing the Actives to high-activity molecules from literature (WB9), and less so—in decreasing order—for Phase I, II, III, Drugs and Inactives. Surprisingly, the 1,551 *Inactives* from MLSMR and NCB appear to be smaller, less complex, less hydrophobic and more soluble at both the 50% and 90% distribution moments, when compared to *Actives*. The WB6 subset of low-activity molecules appears to be more similar to the *Drugs* subset, and not to the set of high-activity molecules. The only subset that violates at least 1 Ro5 criterion at the 50% distribution level is WB9; by contrast, only the *Actives* set (one Ro5 violation) and the *Inactive* set (no Ro5 violations) are within the Ro5 ''cube'' at the 90% distribution moment. The data in Tables 2 and 3 continues to support our earlier observations [8] regarding the trends in lead discovery: Namely, that high-activity molecules (WB9) are significantly more complex, larger, more flexible, more hydrophobic, and less soluble when compared to any other set in this analysis, in particular to *Actives* and to the structurally related WB6 set of low-activity compounds.

Table 1 Property distribution moments (median and 90%) for leads, MLSMR and NCB active compounds

Type	Count	Distr.	MW	SMCM	RNG	HDO	HAC	SumNO	RTB	ClogP	TLogP	TLogSw	Countb	Ro5
Leads	385	50%	287.4	37.58	3	1	4	5	4	2.30	2.28	–3.25	384	0
Leads	385	90%	436.9	62.00	4	3	8.6	11	10	5.08	4.66	–5.24	384	1
MLSMR Act	152	50%	273.9	29.96	3	0	4	5	4	2.50	2.48	–3.32	150	0
MLSMR Act	152	90%	426.5	50.06	4	2	6	8	8	4.40	4.30	–4.78	150	1
NCB Act	46	50%	273.9	37.19	3	1	3	4	2	3.07	3.21	–3.82	45	0
NCB Act	46	90%	419.7	61.86	4.5	3	5	6	9	5.47	4.94	–5.25	45	1

Distr. indicates the distribution moment. Countb indicates the number of molecules for which ALOPGS returned numerical values (TlogP and TLogSw)

Table 2 Median property values for Actives, Drugs, WB9, Phase I–III, Inactives and WB6 compounds

Type	Count	MW	SMCM	RNG	HDO	HAC	SumNO	RTB	ClogP	TLogP	TLogSw	Countb	Ro5
Actives	569	284.3	35.22	3	1	3	4	4	2.45	2.40	–3.32	566	0
Drugs	1,651	340.4	42.87	3	1	4	5	6	2.38	2.43	–3.72	1,636	0
WB9	5,784	463.6	56.73	4	1	5	7	10	3.82	3.60	–4.67	5,716	1
Phase I	801	422.5	51.39	3	1	5	7	8	3.00	2.92	–4.26	788	0
Phase II	1,047	400.4	49.25	3	1	5	6	8	3.17	2.88	–4.27	1,033	0
Phase III	301	379.5	46.42	3	1	5	6	7	2.70	2.60	–4.02	297	0
Inactives	1,551	260.2	28.49	2	0	3	5	4	2.00	1.98	–3.03	1,532	0
WB6	30,690	364.4	42.62	3	1	4	6	6	3.00	2.85	–4.19	30,583	0

Countb indicates the number of TlogP and TLogSw values

Table 3 Ninety percent property values for Actives, Drugs, WB9, Phase I–III, Inactives and WB6 compounds (except for TlogSw, values given for 10%)

Type	Count	MW	SMCM	RNG	HDO	HAC	SumNO	RTB	ClogP	TLogP	TLogSw	Countb	Ro5
Actives	569	432.5	59.77	4	2	6	8	9	5.05	4.61	–5.12	566	1
Drugs	1,651	538.0	71.23	5	3	9	12	14	5.35	4.88	–5.50	1,636	2
WB9	5,784	760.9	94.54	6	3	10	14	25	6.49	5.61	–6.03	5,716	2
Phase I	801	709.7	94.45	5	3	11	14	20	5.93	5.19	–5.77	788	2
Phase II	1,047	730.3	91.64	5	3	11	14.4	21	6.09	5.22	–5.74	1,033	2
Phase III	301	638.8	79.48	5	4	12	14	19	6.36	5.36	–5.51	297	2
Inactives	1,551	370.5	44.37	3	1	5	7	8	3.12	3.20	–3.95	1,532	0
WB6	30,690	589.8	70.89	5	3	8	11	18	6.10	5.27	–5.87	30,583	2

Countb indicates the number of TlogP and TLogSw values

Looking at the MLSMR and NCB subsets; one notices that the *inactives* are smaller than the *actives*. This is perhaps best explained by examining the origin of these ''pub-like'' subsets: The NCB and MLSMR sets were both confirmed with HTS, without follow-up selection. However, high- and low-activity molecules from medicinal chemistry literature (WB9 and WB6, respectively) substantiate the influence of individual processing, since each molecule is designed with the thought of improving (mostly) target-binding affinity.

Trends in lead and drug discovery

Earlier observations by us [8] and others [29] that Leads exhibit lower MW, lower hydrophobicity and higher solubility when compared to Phase I–III compounds continue to hold true, if one compares data from Table 4. Here we illustrate the amount of change in the corresponding property values for MW, SMCM, RNG, SumNO, RTB, TLogP and TLogSw, when using the Leads subset as a baseline. This Table indicates that, on average, the *Leads* subset is significantly smaller, more soluble, less hydrophobic, less flexible and less complex than any other subset (except the MLSMR and NCB actives, as discussed earlier). While HTS hits from public datasets appear to be aligned to the known leads, we continue to note that high-activity molecules are, on average, more than 170 a.m.u. units larger, even though the progression from WB9 to Phase I-III and Drugs indicates a continued ''drop'' in MW. A similar ''drop'' can be noted for the number of rotatable bonds and for TLogP, whereas an increase can be noted for TLogSw. These property changes are

Table 4 Differences in the median (white background) and 90% distribution values (shaded; includes 10% values for TLogSw) for Phase I–III, Drugs, WB9 and WB6 datasets, compared to the Leads dataset

Type	Count	ΔMW	ΔMW	ΔSMCM	ΔSMCM	ΔRNG	ΔRNG	ΔNO	ΔNO	ΔRTB	ΔRTB	ΔTLogP	ΔTLogP	ΔTLogSw	ΔTLogSw
Leads	385	**287.4**	**436.9**	**37.58**	**62.00**	**3**	**4**	**5**	**11**	**4**	**10**	**2.28**	**4.66**	**–3.25**	**–5.24**
Phase I	801	*135.0*	*272.8*	*13.81*	*32.46*	*0*	*1*	*2*	*3*	*4*	*10*	*0.65*	*0.52*	*–1.01*	*–0.54*
Phase II	1,047	*112.9*	*293.4*	*11.67*	*29.64*	*0*	*1*	*1*	*3*	*4*	*11*	*0.61*	*0.55*	*–1.02*	*–0.50*
Phase III	301	*92.0*	*201.9*	*8.84*	*17.48*	*0*	*1*	*1*	*3*	*3*	*9*	*0.33*	*0.70*	*–0.77*	*–0.27*
Drugs	1,651	*53.0*	*101.1*	*5.29*	*9.23*	*0*	*1*	*0*	*1*	*2*	*4*	*0.16*	*0.22*	*–0.47*	*–0.26*
WB9	5,784	*176.1*	*324.0*	*19.15*	*32.54*	*1*	*2*	*2*	*3*	*6*	*15*	*1.33*	*0.95*	*–1.42*	*–0.79*
WB6	30,690	*76.9*	*152.9*	*5.04*	*8.89*	*0*	*1*	*1*	*0*	*2*	*8*	*0.58*	*0.61*	*–0.94*	*–0.63*

All Δ(property) values are in *italics*; actual values for Leads are in bold

further substantiated by the fact that no effort whatsoever was made to discriminate between orally available and intra-venous or intra-muscular drugs. In other words, these observations hold for drug discovery in general (1,651 unique drugs), and can be traced backwards through the MDDR subsets and WB9 (which on the average are the most distinct subset compared to *Leads*).

Thus, high-activity compounds (WB9) continue to exhibit higher MW, higher LogP, higher complexity and flexibility, and lower LogSw when compared to Leads, Phase I-III compounds and Drugs: 63.9% of the WB9 compounds have MW >425, compared to 24% of the *Drugs*, and 10.7% of the *Actives* (and Leads), respectively. In the same category, however, fall 49.2% of the Phase I, 42.4% of the Phase II, and 35.9% of the Phase III candidate drugs from MDDR, as well as 32.4% of the WB6 compounds. Thus, the increase in MW as one steps backwards from launched drugs to Phases III–I and on to high-activity compounds is definite, and bears no relation to the historical set of *Leads*.

In addition to high MW, high-activity compounds also exhibit higher LogP values: 46.6% of the WB9 compounds have ClogP > 4.0, compared to only 20% of the Actives (18.7% of the Leads), and 24.3% of the *Drugs*, respectively. Among MDDR compounds, 33.5% in Phase I, 35% in Phase II, and 30.6% in Phase III have ClogP > 4.0, similar to 33.7% of the WB6 compounds, respectively. It can be concluded that, whereas 20–25% of the *Actives* and *Drugs* have an estimated ClogP above 4.0, a higher proportion of the high-activity compounds (almost 50%) are in the same category. There is a clear tendency towards synthesizing more hydrophobic structures in current medicinal chemistry literature. Similar trends were observed when using TLogP instead of ClogP.

As noted earlier [8], the ''intrinsic'' solubility in neutral state, indicative of a compound's solubility, can also serve as an indicator of how MW and LogP are distributed when observed simultaneously. We found that a significantly higher percentage of high-activity compounds, 57.85%, exhibit TlogSw < –4.5 values, compared to 21.3% of the *Actives* and 31.3% of the *Drugs*, respectively. In the same category (TLogSw < –4.5) fall ~40% of the MDDR and WB6 compounds: 42.1% of Phase I, 43.6% of Phase II, and 38.5% of Phase III clinical candidates and 41.5% of the low-activity compounds, respectively. Thus, there appears to be a definite improvement in solubility as one progresses backwards from WB9 (and WB6) to Phase I, II, III and on to launched drugs—trend which is not observed in the *Actives* set.

Conclusions

The set of MLSMR and NCB actives (198 compounds) was found to be similar to the set of documented (historical) *Leads* (385 compounds), as summarized in Table 1. This observation, applicable to the anticipated set of chemical probes, is even more remarkable when one compares the (significant) differences between the property distribution values of the 569 *Actives* and the 5,784 high-activity molecules from medicinal chemistry literature. The WB9 subset contains molecules that are, on average, larger, more hydrophobic and less soluble than any of the other datasets examined here. In the same manner, it can be noted that the *Actives* (pub-like) subset contains molecules that are, on average, smaller, less complex, less hydrophobic and more soluble than the other datasets.

As the academic community is embarking on chemical probe discovery, the issue of what constitutes high-quality probes is likely to be debated. Arguments such as ''historical bias'' are, and perhaps should be, used when it comes to defining property boundaries such as the Ro5 limits. And yet, the road to success is paved with multiple failures: Rather than increasing, the annual number of new approved drugs has decreased in the past decade, despite significantly larger numbers of molecules, and targets, being considered. Whether the boundary limits will be extend beyond the Ro5 ''cube'', only time will tell. We note, however, that over 55% of the top 200 oral drug products in the United States, Great Britain, Japan and Spain are ''high-solubility drugs'' [30], and that only 18 of the 133 active principles from these drugs have ClogP values greater than 4.0 (data not shown). If past experience is to be related to therapeutic and economic success, as defined by the Top 200 drugs, and by the *Actives* subset, one is advised to observe the boundaries of this property space, and to carefully decide which of these properties should be allowed to exceed these values.

Acknowledgement This work was supported by National Institutes of Health grant U54 MH074425-01 (National Institutes of Health Molecular Libraries Initiative); and by the New Mexico Tobacco Settlement Fund (D.F. and T.I.O.). The calculated properties and SMILES for these 42,394 molecules will be available at the UNM Screening Center website (http://screening.health.unm.edu/). This paper is dedicated to Dr. Yvonne C. Martin, whose four decades of excellence in the areas of QSAR and computer-aided drug design has helped define the field of cheminformatics.

References

1. Austin CP, Brady LS, Insel TR, Collins FS (2004) Science 306:1138
2. The PubChem database is available online at the National Center for Biotechnology Information, Retrieved from http://pubchem.ncbi.nlm.nih.gov/ (09/01/07)
3. Anon (2002) Chemistry & Industry, vol. 4. London, p 9
4. According to the Tufts Center for the Study of Drug Development, the average cost to develop a new prescription drug is $802 million. Retrieved from, http://csdd.tufts.edu/NewsEvents/RecentNews.asp?newsid =6 (09/01/07)
5. Drews J (1998) Drug Discov Today 3:491
6. Horrobin DF (2000) J Royal Soc Med 93:341
7. Olsson T, Oprea TI (2001) Curr Opin Drug Discov Dev 4:308
8. Oprea TI (2002) J Comput Aided Mol Design 16:325
9. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Deliv 23:3
10. Oprea TI, Davis AM, Teague SJ, Leeson PD (2001) J Chem Inf Comput Sci 41:1308
11. Teague SJ, Davis AM, Leeson PD, Oprea TI (1999) Angew Chem Int Ed 38 (1999) 3743. German version: Angew Chem 111 (1999) 3962
12. Hann MM, Leach AR, Harper G (2001) J Chem Inf Comput Sci 41:856
13. Oprea TI (2000) J Comput Aided Mol Des 14:251
14. Proudfoot JR (2002) Bio Org Med Chem Lett 12:1647
15. Available from MDL Information Systems, Retrieved from, http://mdl.com/products/knowledge/drug_data_report/index.jsp, http://www.prous.com/index.html. (09/01/07)
16. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea, TI (ed) (2005) WOMBAT: World of Molecular Bioactivity, in Chemoinformatics in Drug Discovery., Wiley-VCH, New York, 2005, pp 223–239
17. Allu TK, Oprea TI (2005) J Chem Inf Model 45:1237
18. SMARTS tutorial, Retrieved from, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (09/01/07)
19. Leo A (1993) Chem Rev 93:1281
20. ClogP is available from the Biobyte Corporation, http://biobyte.com/index.html, (09/01/07)
21. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV (2005) J Comput Aided Mol Des 19:453
22. Ran Y, Jain N, Yalkowsky SH (2001) J Chem Inf Comput Sci 41:1208
23. Livingstone DJ, Ford MG, Huuskonen JJ, Salt DW (2001) J Comput Aided Mol Des 15:741
24. Tetko IV, Tanchuk VY, Villa AE (2001) J Chem Inf Comput Sci 41:1407
25. The EPI suite is available from the US Environmental Protection Agency, http://www.epa.gov/opptintr/exposure/pubs/episuite.htm (09/01/07)
26. ALOGPS is available from the Virtual Computational Chemistry Laboratory, http://vcclab.org/lab/alogps/ (09/01/07)
27. Tetko IV, Bruneau P (2004) J Pharm Sci 93:3103
28. Tetko IV, Poda GI (2004) J Med Chem 47:5601
29. Wenlock MC, Austin RP, Barton P, Davis AM, Leeson PD (2003) J Med Chem 46:1250
30. Takagi T, Ramachandran C, Bermejo M, Yamashita S, Yu LX, Amidon GL (2007) Mol Pharmaceutics 3:631