Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs

Michal Vieth,* Miles G. Siegel, Richard E. Higgs, Ian A. Watson, Daniel H. Robertson, Kenneth A. Savin, Gregory L. Durst, and Philip A. Hipskind

Eli Lilly and Company, Lilly Research Laboratories, Lilly Corporate Center, DC 1513, Indianapolis, Indiana 46285

Received June 3, 2003

An increasingly competitive pharmaceutical market demands improvement in the efficiency and probability of drug candidate discovery. Usually these new drug candidates are targeted for oral administration, so a detailed understanding of the molecular-level properties that relate to optimal pharmacokinetics is a critical step toward improving the probability of selecting successful clinical candidates. Although the characteristics of druglike molecules have been previously discussed in the literature, the importance of this topic sustains a continued interest for additional perspective and further detailed statistical analyses. In this contribution, we approach the analysis from the perspective of profiling distinguishing features of orally administered drugs. We have compiled both structural and route-administration information for a total of 1729 marketed drugs to provide a solid basis for developing a new perspective on the characteristics of over 1000 orally administered drugs. The molecular properties and most commonly occurring structural elements are statistically analyzed to capture the differences between routes of administration, as well as between marketed drugs and SAR or clinical compounds. We find that, with respect to other routes of administration, oral drugs tend to be lighter and have fewer H-bond donors, acceptors, and rotatable bonds than drugs with other routes of administration. These differences are particularly pronounced when comparing the mean values for oral vs injectable drugs. We also demonstrate that the mean property values for oral drugs do not vary substantially with respect to launch date, suggesting that the range of acceptable oral properties is independent of synthetic complexity or targeted receptor. Finally, we note that, while these properties are descriptive of each class, they are not necessarily predictive of what class any particular drug will reside in, since there is significant overlap in the acceptable ranges found for each drug class.

Introduction

Pharmaceutical researchers invoke the term "druglike"[1−4] to describe molecules with properties that fall within the boundaries delineated by the wide majority of pharmaceutical agents. In general, druglike properties are viewed as those that convey desirable pharmacokinetic and pharmacodynamic (PK/PD) properties,[1−8] independent of pharmacological target or indication. With the recent upsurge in library synthesis, the emphasis on understanding what constitutes druglike properties has increasing significance. Molecules from libraries that emphasize druglike properties are more likely to produce viable structure−activity relationship (SAR) starting points because in a typical SAR effort, optimization of these properties is critical to the success of delivering a clinical candidate and usually independent of the specific pharmacological target.

Understanding which of the many possible *molecular* properties that most directly influence the *druglike* properties of a molecule has been the focus of significant research. Lipinski defined the so-called rule of 5[9] in an effort to address this question. More recently, workers have examined parameters such as the number of rotatable bonds, polar surface area, log *D,* and counts of nitrogen and oxygen atoms in an effort to define easily calculated properties that will be predictive of a favorable PK/PD outcome.[1−3,10−12] These studies typically compare ranges of molecular properties for known drugs with pharmacologically active molecules that are not yet drugs such as those in the MDDR database[13] or for compounds that are not likely to possess biological activity such as those in the ACD database.[14]

In this paper we compare properties of known drugs to two other sets of nondrugs: clinical candidates (compounds not yet approved but in clinical trials) and SAR compounds (compounds known to possess biological activity but not of clinical interest). In addition, we compare not just the properties of drug molecules known to possess good PK/PD properties with nondrug molecules but also those of drug molecules known to exhibit poor PK/PD properties. As a first approximation, we have taken the set of approved drugs with an oral formulation to represent drugs with good PK/PD properties and the set of approved drugs with no oral formulation to represent drugs with poor PK/PD properties. This focuses any property differences on oral bioavailability. We further partitioned the set of nonoral drugs into three categories (injectable, topical, and absorbent) and then compared all nonoral groups independently to each other as well as to the set of oral drugs.

A potential complication with the analysis of oral drugs is the steady evolution of the industry's ability to

* To whom correspondence should be addressed. Phone: (317) 277-3959. Fax: (317) 276-6545. E-mail: m.vieth@lilly.com.

10.1021/jm030267j CCC: $27.50
Published on Web 11/26/2003

synthesize molecules of increasing complexity, coupled with a shift over time in the nature and location of drug targets in the body. These factors alone may cause trends in some molecular properties. We have therefore analyzed our set of oral drugs with respect to launch date to determine, as a first approximation, if any properties show a time dependence that is more likely reflective of chemistry and business realities rather than PK/PD properties.

Finally, we performed an analysis of both oral and nonoral drug sets using Molecular Slicer (MS). This is an in-house tool for computationally "slicing" molecules into their component fragment and scaffold pieces. We then examined the results to determine if certain fragments or scaffolds occur at significantly higher rates in one or the other group of compounds and attempt to correlate the property trends of these fragments and scaffolds with the observed property trends between compound classes.

Methods

Data Gathering. The data were gathered with the goal of compiling the largest possible set of marketed drugs for which the route of administration could be systematically assigned from the FDA's orange book and the Micromedex database.[15] Both sources allow for a methodical assignment of the primary routes of administration based on keywords or information in the databases. The chemical structure was assigned to each drug using the electronic MDL databases (MDDR,[13] CMC3D[16]) or when necessary by querying SciFinder.[17]

An initial list of 1477 single-ingredient FDA-approved drugs was compiled from the 22nd edition of the FDA Orange Book.[18] We were able to assign chemical structures and classify the route of delivery for 1082 of these drugs. A total of 856 were assigned structures using an exact trade name/alias match between orange book and MDL databases (731 coming from CMC3D database), and the structures of the remaining 226 drugs were retrieved from SciFinder.[17] An additional 647 drugs with entries in the 151st volume of the Drugdex[15] database were added to this list of 1082 drugs. From the set of 647, 525 of these had structures in the CMC3D or MDDR[13,16] databases and the remaining 122 structures were retrieved from Scifinder.[17] By use of all of these sources, the number of marketed drugs in our data set for which the chemical structures and route of delivery information could be compiled is 1729. Throughout the remainder of this paper we will term this set of 1729 molecules as "marketed drugs".

A secondary set of 1817 molecules was extracted from the 2002 edition of the MDDR database.[13] These 1817 molecules all had a clinical phase or clinical keyword in the PHASE field of the MDDR database, and for this study, these molecules will be termed "clinical compounds". Finally, 113 937 molecules were extracted from the MDDR database in which the PHASE was classified as "biological testing". These 113 937 molecules will be termed "SAR compounds" in this analysis. These molecules retrieved from the MDDR database contain a reasonably comprehensive collection of patent and literature compounds in early phase development and represent a comparison set that captures some of the breadth of medicinal chemistry SARs.

Route of Delivery Assignment for Drugs. From within the entire set of 1729 marketed drugs, 1193 were assigned to the "ORAL" category based on a clear specification of oral or sublingual route of administration in the Orange Book[18] or in the dosage forms of the Drugdex[15] database. Sublingual, which accounts for 11 drugs, is distinctly different from oral and technically should be a separate group. However, in practice it is difficult to assess, without detailed studies, how much is absorbed through the bucal membranes sublingually vs the amount accidentally absorbed orally. Of the remaining marketed drugs that did not have oral dosage forms, 112 have been assigned to the "TOPICAL" category when there was a clear indication of a topical route of delivery as defined in Orange Book and Drugdex databases. An additional 116 drugs that did not have oral or topical dosage forms and had ophthalmic, otic, nasal, inhalation, vaginal, or rectal dosage forms were assigned to the "ABSORBENT" category in which it is anticipated that the drugs are absorbed through membranes. The remaining drugs that were not previously assigned and had injectable formulations (intramuscular, intravenous, subcutaneous) were assigned to the "INJECTABLES" category. These 308 drugs classified as "INJECTABLES" contain only those that did not have any other specific oral, topical, or absorbent specifications.

Description of Physical Property Calculations. Although we have defined our sets of compounds for cross-comparison more broadly than in previous studies, the physical properties examined in our study have been used in previously reported work.[5,9,12,19] Using a consistent set of properties allows us to easily cross-check our results derived from a larger data set with the previous studies[19] while still providing a comprehensive profile of the physical characteristics of drug groups, clinical compounds, and SAR molecules. These computed properties include molecular weight (MW), atom counts (NATOM), computed logarithm of octanol−water partition coefficient (CLOGP),[20] number of rotatable bonds (ROT), number of rings (NRING), counts of nitrogens and oxygens (ONs), counts of OHs and NHs (OHNHs), rule-based counts of hydrogen-bond acceptors (ACC), counts of hydrogen-bond donors (DON), polar surface area (PSA),[21] total surface area (SA), and number of halogens (halogen).

Chemical Fragments. Molecular Slicer. The structural fragments analyzed in this paper were generated using an internally developed tool, Molecular Slicer (MS), which we routinely use to analyze the common fragments of biologically targeted compounds. MS is similar to other retrosynthetic algorithms previously described in the literature, such as RECAP[22] and REOS.[23] Our tool deconstructs compounds into scaffolds and side chain fragments using a sequential set of 15 preassigned rules. Side chain fragments are characterized by having only one "break point", while scaffolds have two or more. The "break points" are not always based on a logical retrosynthetic step but are frequently defined in a manner that allows us to analyze the composition of large volumes of compounds from a pharmacophore perspective.[24] The rules encoded and used in this study to generate these fragments are detailed in Table 1. After the molecules are processed, the resulting fragments (with the addition of explicit hydrogens) are then used to perform substructure searches to determine the frequency of occurrence of these specific fragments in each set of analyzed molecule sets. Example fragments, substructure queries, and matching molecules are shown in Figure 1.

Statistical Analysis. All statistical analyses were performed within the JMP statistical package.[25] Hypothesis tests were conducted using one parametric (two-sample *t*-test) and two nonparametric (Wilcoxon and median test) tests to compare the means of molecular properties across the various groups of molecules. We have used a *p*-value significance level of 0.05 in at least two of the three tests as a guideline for determining statistical significance. For count-based properties (e.g., number of rings), a $(\text{count} + 0.5)^{1/2}$ transformation was done prior to the *t*-test. Additionally, because of the large sample sizes associated with these comparisons, we also provide some discussion about scientifically meaningful differences in group means to supplement the statistical testing.

Results and Discussion

Physical Property Selection Based on Descriptor Correlation. Although we will present the distribution data for all the properties in this study, we will limit most of our discussion to what is most relevant in differentiating the categories. The most relevant drug-like properties were selected by comparing their correlations with each other.

Table 1. Sequentially Applied SMARTS[29] Queries Used in Molecular Slicer Algorithm[a]

SMARTS	query atoms bond break
a-[CH,CH2;R0;0*]	0 1
[C0*D>1,c0*D>1]-[NH,ND2,ND3,OD2;R0;0*]	0 1
[R;0*]-[CH2R0,NHR0,OR0;0*]-[R]	0 1
-[CD3H,ND2;R0;0](-a)-a	0 1
[a]-&!@[a]	0 1
[NR;0*]-[CD3R0;0*](=O)-[R]	0 1
[NR;0*]-[CD2R0;0*]-[R]	0 1
[N0*,n0*;!H2]-[SR0;0*]-[CD>1,cD>1,ND>1]	0 1
[NR;0*]-[CD2R0;0*]-[CD2,CD3,OD2,ND2,ND3,aD2,aD3]	0 1
a-[NHR0]-[CR0;0*](=O)-[OR0,NR0;0*]	0 1
[CRD>1;0*]-[NH,O;0*]-[CR0;0*](=O)-[NH,O;0*]	1 2
[OD1H0]=[CD3R0;0*](-[ND2,ND3;0*]-[CD>1,aD>1;0*])-[CD>1,OD2,aD>1,ND>1;0*]	1 2
[OD1H0]=[CD3R0;0*](-[CD>1;0*])-[CH2;R0;0*]	1 3
[a;0*]-[CD3R0;0*](=O)-[D2,D3,D4;0*]-[D<4]-[D<4]	0 1
[CR,NR]=[CR]-&!@[a]	1 2

[a] The bond breakages occur between the indicated pair of query atoms. The SMARTS contain locally developed extensions to the SMARTS language, most notably the relational operator. Isotopic labels are applied to designate previously perceived bond breakages, so the queries require nonisotopic atoms for new breakages.

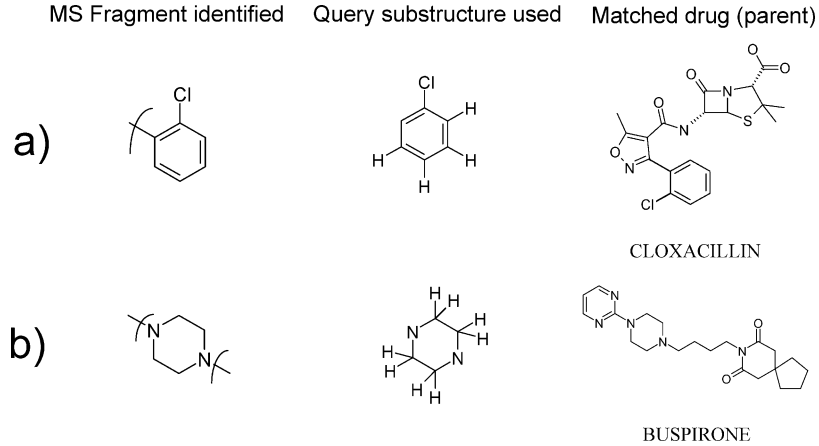

Figure 1. Example of molecular slicer fragments and substructure queries for side chains (a) and scaffolds (b). The MS identified fragment (right) is converted to a substructure query (middle), which is then used to identify the number of molecules containing the fragment (right).

Table 2. Property Correlations for Marketed Drugs[a]

	MW	CLOGP	ONs	OHsNHs	**NATOM**	NRING	ROT	**total SA**	**PSA**	ACC[b]	**DON**[b]	HALOGEN
MW		0.18	0.45	0.12	**0.96**	0.51	0.50	0.88	0.33	0.39	0.13	0.15
CLOGP	−0.03		−0.55	−0.40	0.23	0.20	0.09	0.33	−0.60	−0.51	−0.38	0.16
ONs	0.82	−0.44		0.43	0.41	0.04	0.36	0.28	**0.93**	0.79	0.42	−0.18
OHsNHs	0.66	−0.44	0.78		0.11	−0.07	0.12	0.06	0.54	0.34	**0.99**	−0.11
NATOM	**0.97**	0.01	0.82	0.65		0.59	0.49	**0.92**	0.28	0.32	0.12	0.01
NRING	0.55	0.20	0.34	0.21	0.62		−0.29	0.38	−0.06	0.07	−0.05	−0.03
ROT	0.77	−0.10	0.72	0.62	0.77	0.16		0.70	0.25	0.17	0.11	−0.09
total SA	**0.96**	0.05	0.78	0.64	**0.98**	0.54	0.84		0.14	0.18	0.07	−0.08
PSA	0.74	−0.53	**0.96**	0.82	0.72	0.24	0.67	0.68		0.81	0.53	−0.18
ACC[b]	0.70	−0.46	0.87	0.64	0.67	0.26	0.54	0.62	0.88		0.32	−0.10
DON[b]	0.66	−0.42	0.77	**1.00**	0.66	0.22	0.62	0.64	0.81	0.62		−0.11
HALOGEN	0.08	0.15	−0.13	−0.09	−0.04	−0.08	−0.05	−0.06	−0.13	−0.08	−0.09	

[a] On the basis of the correlation threshold of $R = 0.9$, the initial list of 12 descriptors can be narrowed down to 8 representative properties, and these are shown in bold. Correlation coefficient R(25) is shown. Number of atoms (NATOM), number of donors (DON, computed by a set of donor rules), total surface area (SA), and polar surface area (PSA) are highly correlated with simpler properties. Correlations for all 1719 marketed drugs for which all properties were computed are displayed in the lower diagonal, while the correlations for 1384 drugs that satisfy 10−90% MW coverage (i.e., 196−563 Da) are shown in the upper diagonal. [b] Donors (DON) and acceptors (ACC) are defined according to a set of substructure-based rules derived on the basis of group Pka values and conversations with experienced medicinal chemists. One example of how these classifications differ from simple enumeration of OH, NH groups is the case of a phenol with *tert*-butyl substituents adjacent to the OH. In this case, the OH group is considered spatially hindered from participating in hydrogen bonding.

Table 2 shows the correlation matrix of physical property descriptors for the set of 1729 marketed drugs used in this study. On the basis of the correlation threshold of $R = 0.9$, the initial list of 12 descriptors can be narrowed down to 8 representative properties. The following descriptors were deemed secondary on the basis of their strong correlation with a more fundamental property: PSA(21) based on a 0.96 correlation with ON count; total SA based on a 0.96 correlation with MW; NATOM based on a 0.97 correlation with MW; DON based on a 0.995 correlation with NHOH count (there are very few donors in drugs including ionized guanidinum groups that differ from OHNH counts). The remaining core set of physical descriptors that will be

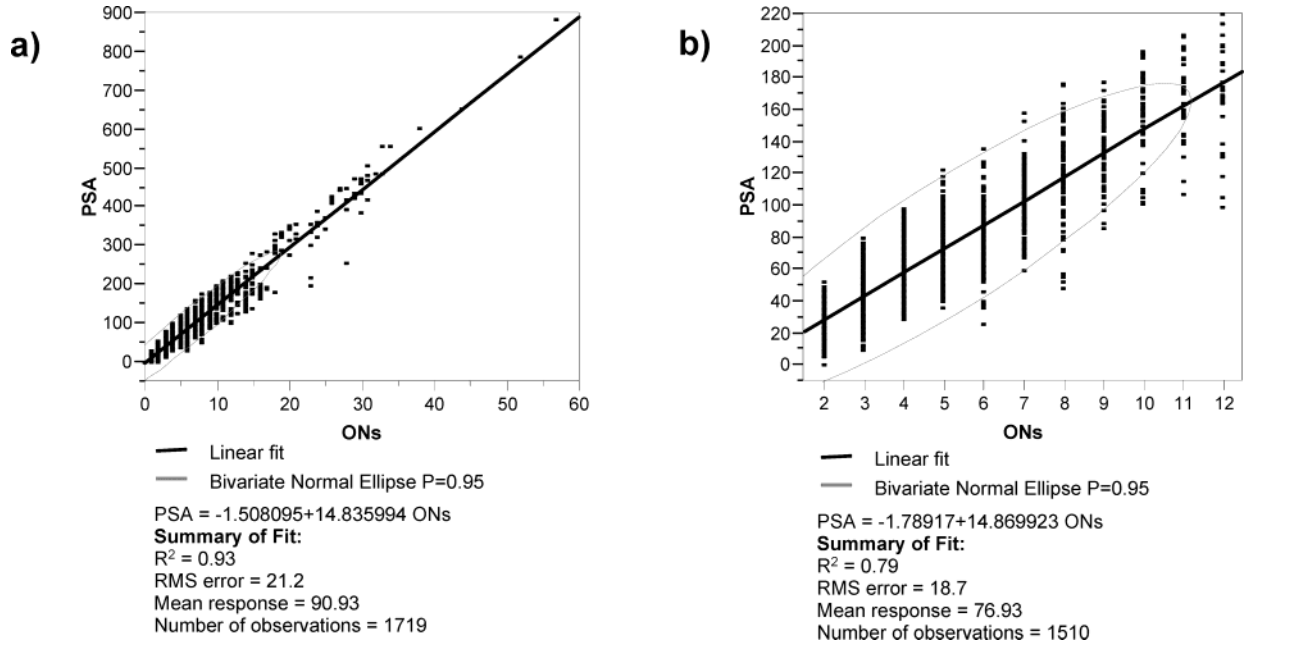

Figure 2. Correlation between selected physical properties for marketed drugs. (a) Polar surface area (PSA)[21] correlation[25] with the count of oxygens and nitrogens (ON) for 1719 marketed drugs used in this study (some drugs failed PSA computation because of their large size). The R^2 is 0.93, indicating that 93% of variance of PSA can be explained by the simple counts of oxygens and nitrogens. (b) PSA correlation with ON count for 1510 drugs with ON count between 2 and 12 (corresponding to 10−90% range of ON for all drugs). R^2 of 0.79 indicates that 79% of variance in total PSA for this group can be explained by the ON count. The difference between (a) and (b) is indicative of the fact that PSA might be nonredundant with ON count for narrow ranges of ON, even though it is not adding value when compounds with wide ranges of ON count are considered. For all comparisons presented in this paper, PSA gives trends consistent with ON count.

used in further discussions includes MW, ON, OHNH, ACC, NRING, ROT, HALOGEN, and CLOGP. Two representative graphs of these correlations are shown in Figure 2a (all drugs) and Figure 2b (drugs with ON count between 2 and 12) for PSA vs count of ONs. It is worth noting that the polar surface distribution for marketed oral drugs is qualitatively similar to the results from a study of 1590 oral drugs that reached at least phase II efficacy studies.[26] The above-mentioned correlations for PSA indicate that more fundamental properties can in principle be used for explaining bioavailability[12] and brain penetration of sets of compounds.[26] However, in practice these properties have limited usage for general groups of compounds where metabolism could be a significant factor in oral bioavailability or brain penetration.

The significance of these computed physical properties can also be examined by studying how the means of different physical properties vary with respect to their FDA approval date. For this analysis, we used a subset of 1082 FDA-approved oral drugs for which the approval date was readily available in the Orange Book.[18] The relationship is depicted for the 691 ORAL drugs in Figure 3a for molecular weight and in Figure 3b for the CLOGP. No meaningful trend with time is observed for either MW or CLOGP. Similar results were obtained for the entire set of 1082 FDA-approved drugs as well as for all other physical descriptors. By contrast, the nature of drug and disease targets has varied considerably over this same time period. We examined the distribution of indications for all oral drugs approved in the years 1982, 1992, and 2002 (Figure 4). These could be divided broadly into six classes: inflammation/asthma, cardiovascular, central nervous system, hormones, infectious disease, and metabolic disease. In 1982, 8 of the 14 approved oral drugs targeted inflammation or cardiovascular pathways (typically peripheral targets). In 1992, 7 of 15 approved oral drugs targeted infectious disease (exogenous pathogens). In 2002, 4 out of the 10 approved drugs targeted the central nervous system. Despite this dramatic variation in the nature and location of the drug target, the mean molecular properties of the drugs do not significantly vary with respect to launch date. The necessity to maintain these physical properties within a particular range highlights the need of all oral drugs to be permeable and absorbed independently of receptor target and strengthens the case for using this subset of physical properties to characterize molecule sets as druglike.

Physical Properties of Oral Drugs. A number of researchers have examined the distribution of computed physical properties of drugs.[5,12,19,27,28] We examined the means and percentiles of physical properties for marketed drugs (grouped by delivery route) as well as clinical and SAR compounds. These results are presented in parts a, b, c, and d of Table 3 for the means, 0−100th percentiles, 10−90th percentiles, and 5−95th percentiles, respectively. On the basis of the means and percentiles of these properties, oral drugs have the lowest MW of all groups while injectable drugs have the highest MW and lowest mean CLOGP. In addition, CLOGP 90th percentiles of all groups, except SAR compounds, are within 1 log unit (the approximate accuracy in the CLOGP computation) of 5, with clinical compounds on average higher and injectables/absorbents lower. The halogen content of all drugs is nearly

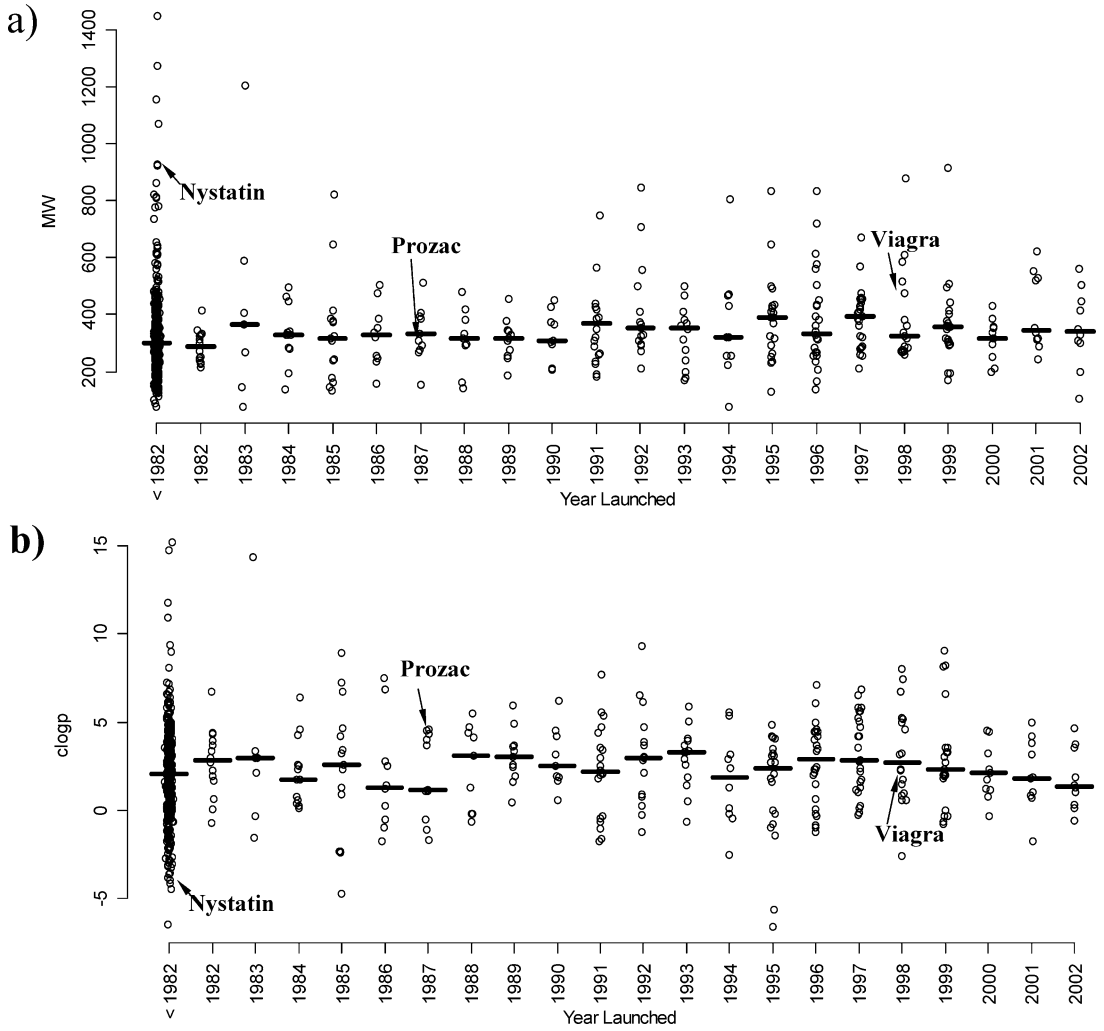

Figure 3. Median properties of FDA-approved ORAL drugs over time. The earliest approval year for each drug is used. Thick horizontal lines show the median for that year. Each point represents one drug's property. The mean values of properties show the same lack of trend with time: (a) molecular weight medians per FDA approval year; (b) CLOGP medians.

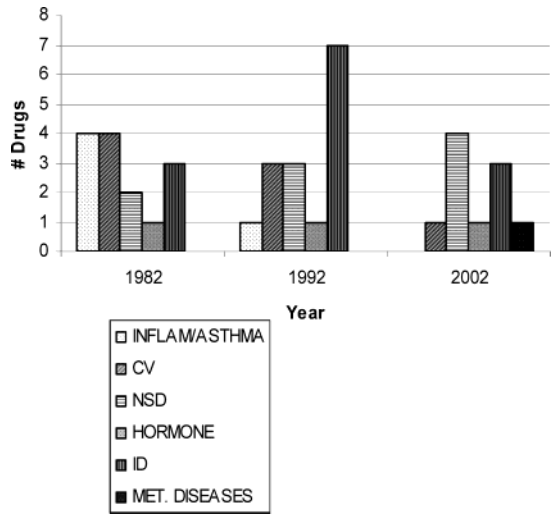

Figure 4. Distributions of drug classes approved by the FDA in 1982, 1992, and 2002. Only drugs present in this study are classified.

identical with the exception of the topicals, which possess a slightly higher halogen content.

Upon comparing these results to the breakthrough work of Lipinski[9] and the more recent work by Wenlock[19] and Blake,[28] we find that our set of 1193 marketed oral drug molecules show property distributions essentially identical to those of the set of 594 oral drugs examined by Wenlock, with minor differences in the distribution for the 90th percentiles for hydrogen-bond donors and acceptors (Table 4).

Comparison of Marketed Oral Drugs to Other Marketed Drugs. Physical Properties. One of the major goals of this study is to compare computed properties between oral drugs and drugs with different routes of administration. The percentile comparisons presented in Table 3, although informative, cannot be easily used to determine statistical significance of differences. We reduced the questions of meaningful difference between physical properties of oral drugs and other groups into a comparison of means of distributions. One parametric (*t*-test) and two nonparametric (Wilcoxon and median test) testing methods were used because of the skewness and outliers present in several of the groups. The statistical significance alone as determined by *p*-values for these large sample sizes is only a guide, since statistically significant differences is a necessary, but not sufficient, criterion for scientifically meaningful differences.

Table 3. Physical Property Means and Percentiles for All Compound Groups Used in This Study

route	MW	CLOGP	ONs	OHsNHs	NATOM	NRING	ROT	total SA	PSA	ACC	DON	HALOGEN
(a) Means												
oral (1193)	343.7	2.3	5.5	1.8	23.9	2.6	5.4	395	78	3.2	1.8	0.5
absorbent (116)	392.3	1.6	6.5	3	27.2	2.5	7.9	456.8	100.5	3.6	3	0.6
injectable (308)	558.2	0.6	11.3	4.7	37.7	3.2	12.7	532.4	143.6	6.2	4.7	0.4
topical (112)	368.5	2.9	5	1.9	25.4	2.9	5.3	412.4	75.4	3.2	1.8	0.9
clinical (1817)	422.5	2.8	7	2.2	29.8	3.3	8	486	98.3	3.9	2.2	0.5
SAR (113937)	447.5	3.4	7.1	2.1	31.5	3.5	8.4	511.5	96.7	4	2.1	0.6
(b) Minimum and Maximum Values (0−100% Percentiles)												
oral (1193)	74−1449	−7.6 to 20.2	0−33	0−18	4−101	0−10	0−40	99−1300	0−447	0−17	0−18	0−18
absorbent (116)	117−1324	−8.7 to 11.7	0−31	0−18	7−96	0−9	0−48	147−1492	0−505	0−16	0−18	0−7
injectable (308)	46−5826	−19.9 to 10	0−144	0−75	3−406	0−11	0−156	86−2387	0−879	0−75	0−75	0−14
topical (112)	60−1423	−11.3 to 10	0−33	0−21	4−100	0−6	0−35	100−1542	0−557	0−23	0−21	0−6
clinical (1817)	30−1456	−11.7 to 22	0−36	0−22	2−100	0−11	0−68	105−1903	0−627	0−22	0−22	0−8
SAR (113937)	59−2133	−19.5 to 30	0−58	0−26	4−100	0−34	0−82	105−2056	0−838	0−41	0−26	0−18
(c) (5−95% Percentiles)												
oral (1193)	164−589	−1.9 to 6.3	2−12	0−4	11−41	1−5	1−12	211−624	13−169	0−7	0−4	0−2
absorbent (116)	160−1007	−3.5 to 5.9	1−20	0−12	10−69	0−5	1−26	177−1073	12−343	0−11	0−12	0−3
injectable (308)	163−1297	−5.0 to 5.8	2−30	0−17	11−90	0−7	1−42	205−1230	20−416	0−15	0−16	0−3
topical (112)	130−505	−2.4 to 6.7	1−10	0−5	9−36	0−5	0−13	171−556	4−156	0−7	0−4	0−3
clinical (1817)	213−755	−1.7 to 7.2	2−15	0−6	15−53	1−6	1−19	265−824	28−215	1−9	0−6	0−3
SAR (113937)	242−776	−1.4 to 7.6	3−15	0−6	17−54	1−6	2−20	287−860	28−207	1−9	0−6	0−3
(d) (10−90% Percentiles)												
oral (1193)	200−475	−0.8 to 5.2	2−9	0−3	14−33	1−4	1−10	246−547	22−134	1−6	0−3	0−2
absorbent (116)	172−666	−2.3 to 4.8	2−14	0−7	11−43	0−4	2−16	225−704	20−219	1−7	0−7	0−2
injectable (308)	196−1085	−3.3 to 4.9	3−23	0−11	13−71	1−6	2−27	238−979	28−311	1−11	0−11	0−1
topical (112)	188−495	−0.6 to 6.0	2−8	0−3	12−35	1−5	1−9	227−531	21−114	0−5	0−3	0−3
clinical (1817)	250−614	−0.7 to 6.0	3−12	0−4	18−43	1−5	2−15	293−696	39−169	1−7	0−4	0−2
SAR (113937)	276−646	−0.0 to 6.6	3−12	0−4	19−45	2−5	2−16	324−726	38−167	1−7	0−4	0−2

Table 4. Comparison of the Distribution of Rule of 5 Properties with the Previously Published Work of Lipinski[9] and Wenlock[19]

grouping	maketed oral[a]	selected USAN compounds[b]	marketed oral from Wenlock et al.[c]	Lipinski set recomputed[d]
number of molecules	1193	2245	594	1791
mean MW	344		337	300
mean CLOGP	2.3		2.5	2.5
mean acceptors/ON	5.5		4.9	4.5
mean donors/NHOH	1.8		2.1	3
90th percentile MW	474.6	500 (89%)	473	427.5
90th percentile CLOGP	5.2	5	5.5	5.3
90th percentile acceptors (ON)	9	10	8	8
90% donors (NHOH)	3	5	4	3

[a] Data from this study. [b] Data from Lipinski's original work.[9] [c] Data from Wenlock's paper.[19] [d] A set used in Lipinski's work[9] with only unique active ingredient organic structures present based on the 2001 edition of the WDI.[30] The recomputed data for Lipinski's USAN set show good agreement with our properties of oral drugs with the exception of having a smaller mean molecular weight. The differences in our distributions from those of the original report by Lipinski[9] are possibly due to slight differences in salt handling, duplicate structure removal, and other details in data processing.

Table 5 shows the property distribution means, median values, and *p*-values between the means of the marketed oral drugs and all other groups. On the basis of a significance of 0.05 in at least two of the three tests, oral drugs differ from injectables in all properties we evaluated. Specifically, we find that injectables have significantly higher mean MW, ON, OHNH, NRING, rotatable bonds, and acceptor counts and lower mean CLOGP and halogen counts than oral drugs. Generally the differences are quite large and indicate that injectable drugs are significantly heavier, more polar, and more flexible than oral drugs. Interestingly, in four of these seven properties (MW, OHNH, rotatable bonds, and acceptor count), the oral and injectable sets are at or near the extremes for mean values across all the sets examined. We can therefore look at the means of these two groups as two extreme cases of physical properties acceptable for xenobiotics.

Absorbent drugs show significant differences when compared to oral drugs in CLOGP and OHNH counts, in the same direction though not to the same degree as injectables. No significant difference was found for MW or the counts of ONs, halogens, rings, rotatable bonds, and acceptors.

Topical drugs are similar to oral drugs in acceptors, ONs, OHNHs, acceptors, and rotatable bonds, while showing statistically significant differences in MW, number of rings, halogens, and CLOGP. The difference in physical properties of oral drugs is most meaningful when compared to injectables but is significant in some properties when compared to absorbent and topical drugs as well. Of the three nonoral drug categories, topical and absorbent drugs appear most similar in their properties to oral drugs. Solely on the basis of these calculated properties, many topical drugs may in fact possess favorable PK/PD profiles but are administered topically to limit their distribution to the desired regions of the body.

We next compared the physical property means of oral drugs to results from the clinical and SAR sets. SAR molecules present in MDDR possess physical properties significantly (due to large sample sizes) and meaning-

Table 5. Differences in Means for Selected Properties between Oral and Nonoral Drugs[a]

descriptor	oral mean (median) $n = 1202$	absorbent mean (median) $n = 118$	p-value	injectable mean (median) $n = 328$	p-value	topical mean (median) $n = 113$	p-value	SAR mean (median) $n = 113\,937$	clinical mean (median) $n = 1817$
			0.0016		**<0.0001**		**0.092**		
MW	343.7	392.3	0.49	558.2	**<0.0001**	368.5	**0.0094**	447.5	422.5
	(322.5)	(332.4)	0.43	(416.4)	**<0.0001**	(379.1)	**0.017**	(414.6)	(390.5)
			0.0059		**<0.0001**		**0.032**		
CLOGP	2.3	1.6	**0.02**	0.6	**<0.0001**	2.9	**0.001**	3.4	2.8
	(2.3)	(2.0)	**0.18**	(0.7)	**<0.0001**	(3.3)	**0.0002**	(3.5)	(3)
			0.073		**<0.0001**		0.06		
ONs	5.5	6.5	0.99	11.3	**<0.0001**	5	0.02	7.1	7
	(5)	(5)	0.27	(8)	**<0.0001**	(4)	0.12	(6)	(6)
			<0.0001		**<0.0001**		0.76		
OHsNHs	1.8	3	**0.007**	4.7	**<0.0001**	1.9	0.25	2.1	2.2
	(1)	(2)	**0.03**	(2)	**<0.0001**	(1)	0.38	(2)	(2)
			0.055		**0.0002**		**0.2**		
NRING	2.6	2.5	0.053	3.2	**0.0007**	2.9	**0.026**	3.5	3.3
	(3)	(2)	0.65	(3)	**<0.0001**	(3)	**<0.0001**	(3)	(3)
			<0.0001		**<0.0001**		0.57		
rotbond	5.4	7.9	0.15	12.7	**<0.0001**	5.3	0.36	8.4	8
	(5)	(4.5)	0.89	(7)	**<0.0001**	(5)	0.62	(7)	(6)
			0.21		**<0.0001**		0.71		
ACC	3.2	3.6	0.48	6.2	**<0.0001**	3.2	0.74	4	3.9
	(3)	(3)	0.63	(5)	**<0.0001**	(3)	0.16	(3)	(3)
			0.38		**0.087**		**<0.0001**		
HALOGEN	0.5	0.6	0.84	0.4	**0.0003**	0.9	**<0.0001**	0.6	0.5
	(0)	(0)	0.64	(0)	**<0.0001**	(0)	**0.0002**	(0)	(0)

[a] Within a p-value cell, the top p-value is from the two-sample t-test, the middle p-value is from the Wilcoxon test, and the bottom p-value is from the median test. For count-based descriptors, the t-test was performed on a $(\text{count} + 0.5)^{1/2}$ transformation. All p-values for the SAR group were <0.0001 and are not included in the table. All p-values, except for the halogen count, were <0.0001 for the clinical group and are not included in the table. Values in bold indicate that at least two of the three p-values are <0.05.

fully different from those of the marketed oral drugs (Table 5). The same is true for the majority of properties characterizing clinical compounds, with only the number of halogens showing a similar mean to oral drugs. Interestingly, all properties for clinical and SAR compounds have higher means in contrast to the other drug groups, which usually balance higher MW with lower CLOGP or other properties as shown in Table 3. This relative imbalance of properties of clinical and SAR compounds compared to that of drugs suggests one reason for the high attrition rate of drug candidates.

Unique Features of Marketed Oral Drugs. Common Fragments. Next, we considered whether molecular fragments commonly found in drugs exhibit molecular property trends similar to those of their parent molecules. To answer this question, we employed our pseudo-retrosynthetic tool, Molecular Slicer (MS), to identify the molecular fragments and count the occurrence of them in the different drug groups. Because the number of marketed nonoral drugs in each category was small, we decided to use INJECTABLE category only for this comparison. The 15 most common side chains (one MS break point) identified for each group are shown in Figure 5.

We then calculated the mean molecular properties for both the common ORAL and INJECTABLE fragments and looked for significant differences between the two groups. In contrast to the analysis of the drug molecules as a whole, none of the eight properties examined showed a statistically significant difference between the two groups; indeed, several of the most common side chains were found in both groups.

We next examined MS-derived scaffolds (fragments with two or more MS break points), depicted in Figure 6. These scaffolds can be thought of as linkers between side chains. The mean values for both ON count, rotatable bonds, and CLOGP showed statistically significant differences between the ORAL and INJECTABLE sets, with a magnitude and direction for the difference consistent with the trends observed for the whole molecule. A cursory visual comparison of the most frequent scaffolds reveals that the INJECTABLE scaffolds tend toward more polar character and flexibility than the ORAL scaffolds, and many would appear to be peptide in nature. This is consistent with a significant increase in the number of rotatable bonds for nonoral drugs; previous studies have shown that more rigid structures favor desirable PK/PD properties,[12] although we caution that it is difficult to assign causation using highly correlated observational data. Our analysis would suggest that this increased flexibility for nonoral relative to oral drugs tends to reside in the core of the molecule rather than in the side chains.

Conclusions

Differences between the marketed oral drugs and other marketed drug groups reveal the factors influencing oral bioavailability. Structurally, these differences tend to reside in the central scaffolds of drug molecules rather than in the appended side chains. The lower molecular weight, balanced CLOGP, and greater rigidity improve the likelihood of producing drugs candidates with an oral route of administration.

The differences between the marketed oral drugs and clinical or SAR compound properties, strengthened by the recent finding on the convergence of properties throughout development phases,[19,28] reveal possible reasons for the high attrition rates in development and

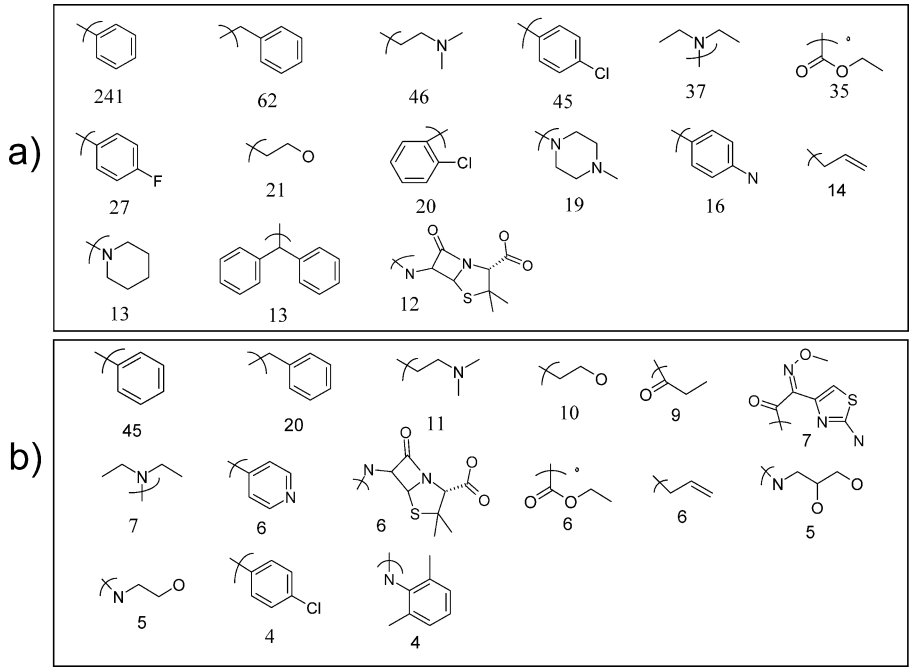

Figure 5. Comparison of the most frequent side chains: (a) oral; (b) injectable. The numbers indicate the count of the drugs containing that fragment. The means of properties are not significantly different for (a) and (b).

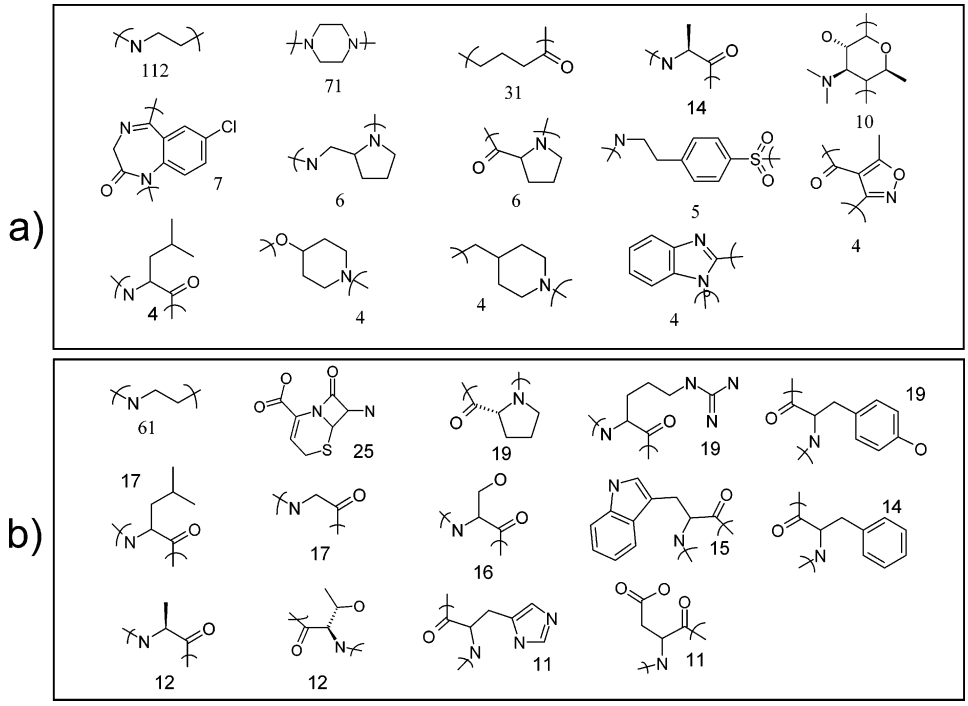

Figure 6. Comparison of most frequent scaffolds: (a) oral; (b) injectable. The numbers indicate the number of drugs containing the fragment. The means of physical properties (CLOGP, ON, rotbond) are significantly different for (a) and (b).

suggest the range of preferred physical properties and molecular building blocks to be targeted in the development of oral drugs. Although the pharmacological targets have changed over the years, the necessary characteristics for producing a drug candidate that is amenable to oral delivery have remained relatively stable. This implies that medicinal chemistry explorations of SARs with mean properties and scaffolds preferentially present in oral drugs should result in clinical candidates with favorable oral bioavailablity.

Finally, although comparing the means of physical properties between various drugs or drug fragments is informative, we caution against using these insights as predictors for "druglike" vs "non-drug-like" properties for individual substances. Because of the substantial overlap in the range of properties found between the different drug classes, we cannot accurately classify a particular drug as either oral or injectable on the basis of simple physical property calculations. These properties should instead be viewed as multivariate profiles that can be used to compare sets of molecules (i.e., synthetic libraries) to determine if a set is more or less like the oral drug group defined here. This work presents statistically motivated ideas and hypotheses

regarding characteristics of druglike molecule and fragment sets that we hope can be applied to early-phase research to decrease the time and increase the probability of delivering successful clinical candidates.

Acknowledgment. Fruitful discussions with Jim Wikel are acknowledged. We thank Dr. Michael Myers for critical comments and review of this manuscript and Dr. Abdelaziz Mahoui for help in the information gathering for drugs. We thank Chris Lipinski for pointing out the need to carefully examine the active ingredients for all drugs used in this study.

Supporting Information Available: Table with 1729 drug names, CAS numbers, computed properties, and the route of administration. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Proudfoot, J. R. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1647−1650.
(2) Muegge, I. Selection Criteria for Drug-like Compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.
(3) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.
(4) Walters, W. P.; Ajay, X.; Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384−387.
(5) Walters, W. P.; Murcko, M. A. Prediction of "drug-likeness". *Adv. Drug Delivery Rev.* **2002**, *54*, 255−271.
(6) Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(7) Egan, W. J.; Walters, W. P.; Murcko, M. A. Guiding molecules towards drug-likeness. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 540−549.
(8) Podlogar, B.; Muegge, I.; Brice, L. Computational methods to estimate drug development parameters. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 102−109.
(9) 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.
(10) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res. 14*, 568−571.
(11) Sakaeda, T.; Okamura, N.; Nagata, S.; Yagami, T.; Horinouchi, M.; et al. Molecular and pharmacokinetic properties of 222 commercially available oral drugs in humans. *Biol. Pharm. Bull.* **2001**, *24*, 935−940.
(12) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; et al. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(13) *MDL Drug Data Report*, version 2002.2; MDL Information Systems Inc.; San Leandro, CA, 2002.
(14) *Available Chemicals Directory (ACD)*, version 2002.2; MDL Information Systems Inc.: San Leandro, CA, 2002.
(15) *MICROMEDEX, T. MICROMEDEX(R) Healthcare Series*; Micromedex, Inc.: Englewood, CO, 2003; Vol. 115.
(16) *Comprehensive Medicinal Chemistry, CMC*; MDL Information Systems Inc.: San Leandro, CA, 2002.
(17) *SciFinder*, 2001 ed.; American Chemical Society: Washington, DC, 2002.
(18) *Food and Drug Administration Orange Book*, 22 ed., U.S. Food and Drug Administration: Washington, DC, 2003.
(19) 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.
(20) *CLOGP*; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA, 2002.
(21) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(22) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(23) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening—an overview. *Drug Discovery Today* **1998**, *3*, 160−178.
(24) Sheridan, R. P. Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037−1050.
(25) *JMP*, version 4.0.4; SAS Institute Inc.: Cary, NC, 1989−2001.
(26) Kelder, J.; Groothenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P.; Ploemen, J.-P. Polar Molecular Surface as a Dominating Determinant for Oral Absorption and Brain Penetration of Drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.
(27) Oprea, T. I. Current trends in lead discovery: Are we looking for the appropriate properties. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.
(28) Blake, J. F. Examination of the Computed Molecular Properties of Compounds Selected for Clinical Development. *BioTechniques* **2003**, *34*, S16−S20.
(29) James, C. J.; Weininger, D.; Delany, J. *Theory manual—SMARTS*; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA, 2001.
(30) *Derwent World Drug Index*, 2001−2002 ed.; Derwent Publications: London, U.K., 2001.

JM030267J