Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds

N. Baurin,* R. Baker, C. Richardson, I. Chen, N. Foloppe, A. Potter, A. Jordan, S. Roughley, M. Parratt, P. Greaney, D. Morley, and R. E. Hubbard

Vernalis (Cambridge) Ltd., Granta Park, Abington, Cambridge, CB1 6GB, UK

Received November 11, 2003

We have implemented five drug-like filters, based on 1D and 2D molecular descriptors, and applied them to characterize the drug-like properties of commercially available chemical compounds. In addition to previously published filters (Lipinski and Veber), we implemented a filter for medicinal chemistry tractability based on lists of chemical features drawn up by a panel of medicinal chemists. A filter based on the modeling of aqueous solubility (>1 μM) was derived in-house, as well as another based on the modeling of Caco-2 passive membrane permeability (>10 nm/s). A library of 2.7 million compounds was collated from the 23 compound suppliers and analyzed with these filters, highlighting a tendency toward highly lipophilic compounds. The library contains 1.6M unique structures, of which 37% (607 223) passed all five drug-like filters. None of the 23 suppliers provides all the members of the drug-like subset, emphasizing the benefit of considering compounds from various compound suppliers as a source of diversity for drug discovery.

INTRODUCTION

Most modern drug discovery projects start by identifying molecules with appropriate affinity for a target by assaying large numbers of available compounds. The resulting "hit" compounds are then assessed for their suitability as "leads" to enter the optimization process. To become a drug, a compound requires not only the desired binding affinity and selectivity for a particular target but also the incorporation of appropriate ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties.[1] There has been much interest recently in characterizing and defining the ADMET and physicochemical (MW, solubility) properties of compounds that make them suitable to be drugs ("drug-like") and the extent to which such properties need to be present in the initial lead molecules for optimization to be successful ("lead-like").[2,3] These analyses have generated a variety of in silico approaches to predict ADMET properties[4] and in addition focused attention on filtering the initial set of compounds used in hit generation to remove molecules with undesirable features. In this paper, we have implemented and extended a range of Oral Absorption prediction methods and applied them to large databases of compounds available from compound suppliers. These calculations analyze to what extent the appropriate combination of drug-like properties are present in commercially available compounds that can be screened for hit generation.

ADMET properties derive from a vast range of in vivo mechanisms, many of which are unknown or multicomponent. It is therefore currently impossible to predict such properties from first principles. Instead, knowledge based approaches have been developed, generating rules based on relatively simple compound properties, derived from known existing experimental ADMET data.[5] It is important to distinguish local from general ADMET modeling. Local modeling uses project specific data (often limited to a few chemotypes) to produce models that can be developed and applied in real time as a project progresses. In contrast, general ADMET modeling derives trends derived from many projects to produce rules, often known as drug-like filters.[6] Such filters can be used during compound optimization to derive an acceptable property space. Once the boundaries of this drug-like chemical space are crossed, the likelihood of adverse properties increases for a candidate during preclinical and clinical trials.

The first widely used drug-like filter was developed by Lipinski et al., who analyzed the trends in simple molecular descriptors for a reference set of ~2300 compounds[1] all passing phase I clinical trials. In this paper, we implement two "Lipinski" drug-like filters flagging compounds fulfilling either three out of four or all four of the following criteria: molecular weight $\leq$ 500, SlogP $\leq$ 5, number of H-bond donors $\leq$ 5, number of H-bond acceptors $\leq$ 10.

More recently, Veber et al. proposed a drug-like filter by analyzing a proprietary database containing ~1100 drug candidates.[7] In this paper, we implement a "Veber" drug-like filter flagging compounds that have a polar surface area less than 140 Å^2 and have less than 12 rotatable bonds.

While Veber and Lipinski rules were established by analysis of drug-like sets, Oprea et al. developed a so-called "ChemGPS" system using a set of 423 satellite compounds, specifically selected to be out of the drug space.[8] More complex models derived using neural networks training procedures have also been developed based on drug and nondrug sets (MDDR or CMC against ACD[9,10]). Although quite effective, like many models derived with neural networks, it is difficult to formulate any simple rules from the model derived. We chose to implement the Lipinski and Veber filters because they address drug-likeness using widely understood molecular properties. This is crucial to ensure

* Corresponding author e-mail: n.baurin@vernalis.com.

10.1021/ci034260m CCC: $27.50
Published on Web 02/10/2004

Table 1. Summary of Suppliers Whose Compound Catalogues Were Used in This Study[a]

supplier name	address	date/version of library	initial number of compounds	library size[a]
Analyticon	www.ac-discovery.com	Jul 2002	765	765
Asinex	www.asinex.com	Aug 2003	202 665	202 542
Aurora	www.aurora-feinchemie.com/kutyrev/	Jan 2003	2252	2245
Biofocus	www.biofocus.co.uk	Jan 2003	10 486	10 486
Bionet	www.keyorganics.ltd.uk	Jun 2003	34 708	34 699
ChemDiv	www.chemdiv.com	Jun 2003	362 299	362 229
ChemOvation	www.chemovation.com	Jan 2003	1049	1049
ChemT&I	www.chemti.com	Jun 2003	165 113	165 054
Chembridge	www.chembridge.com	Aug 2003	319 720	319 537
Comgenex	www.comgenex.com	Jun 2003	164 666	164 663
Enamine	www.enamine.relc.com	Jul 2003	123 072	122 859
IBS	www.ibscreen.com/	Jul 2003	284 271	283 825
IFLab	www.iflab.kiev.ua	Jun 2003	150 050	149 862
MCL	www.mosmedchemlabs.com	Aug 2002	105 891	105 868
MDPI	www.mdpi.org	Jun 2002	9067	8734
Maybridge	www.maybridge.com	Aug 2003	56 485	56 405
Menai	http://www.ryansci.com/meninfon.shtml	Jun 2002	3778	3777
Specs	www.specs.net	Aug 2003	220 589	220 558
TOSLab	www.toslab.com	Jun 2003	6682	6669
TimTec	www.timtec.net/home.htm	Jun 2003	128 472	128 181
Tripos	www.tripos.com	Jul 2003	92 920	92 920
UKR	03187, Kyiv-187, Zabolotnogo St., 82, a. 60, Ukraine	Sep 2002	118 473	118 426
Vitas-M	www.vitasmlab.com	Jun 2003	113 817	113 669

[a] Number of compounds after removal of stereoisomers within each supplier's catalog.

the drug-like filters can be used on a daily basis by as broad an audience as possible within a drug discovery organization.

Flagging compounds containing chemical groups incompatible with drug development[11] can be considered as a drug-like "MedChem tractability" filtering approach. This requires the input of medicinal chemistry experience, to build substructure queries that identify chemical features of compounds which are intrinsically reactive or which make them difficult to synthesize or modify. In this paper, we implement two MedChem tractability filters. The MedChem-1 filter flags compounds on the basis of a list of 50 groups considered as undesirable by a consensus of the Vernalis medicinal chemists (e.g. aldehyde). Similarly, the MedChem-2 filter flags compounds on the basis of a list of 14 groups considered as potentially undesirable (e.g. guanidinium).

The Lipinski drug-like filters reflect indirectly that a good drug candidate should show both a reasonable aqueous solubility and a membrane permeation profile[1,12] (lipophilicity and H-bond capabilities cutoffs, respectively). Bergstrom et al. recently derived solubility and membrane permeation models and, by combining the two predictions, they rapidly obtained a reasonably accurate absorption profile of drug-like molecules.[12] In this paper, we implemented aqueous solubility and membrane permeation models, based on publicly available data. We then used these models to calculate distributions of properties for a reference set of 1141 small organic drug-like compounds. From these distributions, we derived a solubility drug-like cutoff and a Caco-2 membrane permeation drug-like cutoff.

There are many suppliers of chemical compounds around the world, with catalogues totalling many millions of structures. With virtual libraries, billions of compounds are theoretically available from enumeration of multibranched chemical scaffolds. For both types of databases, drug-like filters can dramatically reduce the library size to focus on the area of interest very rapidly, as such filters can handle millions of structures per day on a standard workstation.[13,14] In addition, such drug-like alerts can be used to guide medicinal chemistry prioritization for synthesis of smaller batches of compounds.

We have assembled a database of the 2.7 million compounds available from 23 suppliers. This database has been assessed using these five drug-like filters, the largest analysis since the reference paper by Voigt et al.[15] and a more recent analysis by Mozziconacci et al.[16] In addition to the drug-like filter results, we report a duplicate analysis in the drug-like space of these 23 suppliers.

METHODS

Library Preparation. For all calculations described in this paper, the chemical structures were processed to protonate basic groups and deprotonate acidic groups. Structures of compounds supplied by the 23 compound suppliers (Table 1) were assembled into an in-house Oracle database, linked to structures via ISIS/Host.[17] The SMILES representation (Simplified Molecular Input Line Entry System (SMILES[18]) has been used to define and process molecules. The analyses described in this paper are based on 1D/2D descriptors that do not distinguish stereoisomers. Therefore, a 2D SMILES string representation of each compound was used to remove the very few duplicates within the initial set of compounds available from each supplier.

Software. The Molecular Operating Environment (MOE, Chemical Computing Group, version 2002) was used to calculate all 1D or 2D descriptors and derive all drug-like filters, except the MedChem tractability filters which are defined by the ISIS atoms and bond querying tools, and applied with a combination of PL/SQL scripts[19,20] using MOL files. The PLS and cross-validation procedures were encoded in Scientific Vector Language (SVL) to produce the solubility model and Caco-2 permeation model.

Lipinski Filter. The original "rule of 5"[1] was derived by analysis of the compounds in the World Drug Index that

Table 2. List of 50 Chemical Features Used by the MedChem-1 Tractability Filter[a]

CHO	inorganics	S−Cl, Br, F	quaternary nitrogen	cyclic alkylating agents
COCl, Br, F	SO2Cl (Br, F)	N-halogen	alkyl *N*-oxides	anthracenes
anhydrides	C═O−S	disubstituted sulfates, sulfonates	dinitrobenzenes	phenanthrene
OCOCl, Br, F	C═S−O	carbodiimides	radicals	ortho-quinones
alkyl chloride and alkyl bromide		imines	B,P,Si,Sn	fluoranthrene
thiol (SH)	C═S	azo	polyaryl-sulfonyl	binaphthyl
isonitriles	epoxides & aziridines	linear thioureas	bis-guanidinium	ketenes
NCO	S−S	adamantanes	complex dyes	triptycene
NCS	O−O (peroxide)	hydroxylamines	quinone-like	1,4-dianiline
metals	N═O (nitroso)	positively charged	aromatic precursors	beta-naphthylanilines

[a] Compounds containing any of these features fail MedChem-1 filter.

had progressed to at least Phase 2 clinical trials. This analysis established that poor human absorption or permeation is more likely when there are more than 5 H-bond donors, more than 10 H-bond acceptors, a molecular weight greater than 500, and a ClogP greater than 5. We used an analogous approach to derive our "own" Lipinski drug-like filters based upon the calculation of SlogP and the count of H-bond donors and H-bond acceptors. SlogP was calculated with the MOE software (version 2002). The identification of H-bond donors/acceptors was through the definitions based on SMILES implemented by default within MOE.

Veber Filter. The original "rule of 2"[7] was derived from an analysis of oral bioavailability measurements in rats for over 1100 drug candidates studied at GlaxoSmithKline. The rule predicts a high probability of good oral bioavailability in the rat if a compound has 10 or fewer rotatable bonds and a polar surface area equal to or less than 140 Å^2. The Veber drug-like filter used in this paper is based on the same method for calculation of polar surface area and the same definition of rotatable bonds (terminal bonds not taken into account, cyclic bonds not taken into account, amide bond not taken into account).

MedChem Tractability Filters. Following extensive discussions with medicinal chemists within our company, we established a list of 50 undesirable chemical features (MedChem-1 filter defined in Table 2). Fourteen potentially undesirable chemical features were also listed (MedChem-2 filter defined in Table 3).

Statistical Parameters. The mean absolute error (MAE) and the correlation coefficient r^2 are used in this paper to describe the quality of the fit for the Caco-2 permeation and aqueous solubility linear regression model.

$$r^2 = \frac{\left(\sum_{p=1}^{n}(y_p^{real} - \overline{y^{real}}) \times (y_p^{pred} - \overline{y^{pred}})\right)^2}{\left(\sum_{p=1}^{n}(y_p^{real} - \overline{y^{real}})^2\right) \times \left(\sum_{p=1}^{n}(y_p^{pred} - \overline{y^{pred}})^2\right)}$$

Aqueous Solubility Modeling. A wide variety of models have been built for aqueous solubility using different training sets of organic molecules and different types of data mining protocols.[21−27] Our objective was to derive a water solubility model from large, publicly available, compiled data sources, with a mean absolute error within the estimated overall experimental uncertainty of 1 log unit. A secondary desirable objective was for the model to be expressed as chemically understandable 1D/2D descriptors. The training set contains 3041 structures along with their solubility in pure water, collected from two publicly available sources, the PhysProp database[28] and a list of structures and aqueous solubilities extracted from an article by Huuskonen et al. [29] Compounds containing nonorganic elements were removed using a SMILES[18] based script.

Table 3. List of 14 Chemical Features Used by the MedChem-2 Tractability Filter[a]

allenes	highly chiral compounds
alpha-naphthylalanines	hydrazones
aminals and acetals	oximes and oxime ethers
anilines	pyridine N-oxides
barbiturate	steroid like
cyclic thioureas	guanidinium or amidine
cyclic alkylating agents	binaphthyl

[a] Compounds containing any of these features fail MedChem-2 filter.

Each molecule was described by five general descriptors (MWeight, SlogP,[30] Polar Surface Area,[31] PEOE_RPC+, PEOE_RPC−) and 40 group-count descriptors. PEOE_RPC+ is the relative positive partial charge on the molecule, obtained by dividing the largest positive q_i by the sum of all the positive q_i. The partial charges q_i were calculated with the Partial Equalization of Orbital Electronegativities (PEOE) method.[32] PEOE_RPC− is the corresponding relative negative partial charge. The 40 group-count descriptors are defined according to the MOE input language,[33] which is an implementation of the SMILES representation: [CH3], [CH2!i], [CH!i], [CQ4!i], [CH2iQ1], [CHiQ2], [CiQ3], [cH], [cQ3], [c$(c(a)(a)(a))], [C$(C(=*)(=*))], [CQ1H$(C#*)], [CQ2$(C#*)], [NQ1!i], [NQ2!i], [NQ3!i], [NQ1i], [NQ2i!$(N(=*)(=*))], [nH0Q2], [nHQ2], [nQ3], [NQ3$(N(∼O)(∼O))], [N$(N#*)], [OQ1!i], [OQ2!i], [*$(*=[OQ1i])], [o], [S+0Q1!i], [SQ2!i], [*$(*=[SQ1i])], [*$(*[S-1Q1])], [s], [SQ3$(S(∼O)(*)(*))], [SQ4$(S(∼O)(∼O))], [*-*$(*P)], [F], [Cl], [Br], [I], [NQ4!i]. The choice of structural groups was mainly inspired by the solubility modeling work published by Klopman et al.[34] We nonetheless use a smaller set of structural groups, since we did not observe any significant predictive improvement when using more group descriptors.

Partial Least Square (PLS[35]) regression was applied and a model built with 25 components. This optimal number of components was chosen on the basis of a cross-validated Leave-Group-Out-5 (LGO-5) protocol (the training set is randomly split into five subsets of equal size: each subset is then predicted with a model built using the remaining 4 groups). For the training set, the model has a correlation coefficient r^2 of 0.82, and a mean absolute error MAE of

0.70. The predictive performance of the model was assessed in two ways. First, the LGO-5 cross-validation procedure gave a q^2 value (0.81) very similar to the r^2 value (0.82). Second, solubility was calculated for a reference test set of 21 drug-like molecules[29] of known drug-like molecules not present in the training set. The r^2 value is 0.82 and MAE value is 0.67.

The predictive ability of our PLS model is equivalent to that of other linear models (e.g. Huuskonen et al., $r^2 = 0.80$, for a multilinear regression model based on a 675 molecule training set[29]) but less predictive than nonlinear models (e.g. Tetko et al., $r^2 = 0.91$, for a neural-network model based on a training set of 1291 molecules[36]). However, given that the error on experimental solubility values could reach 1 log unit,[23] the accuracy of our model is acceptable, particularly as it is based on chemically understandable descriptors. As with the Caco-2 model described below, these features are important to gain acceptance with the chemistry end-user and also to allow rapid calculation of properties for many millions of compounds.

Caco-2 Passive Membrane Permeation Modeling. In contrast to solubility there is no major, publicly available data source for cell permeability measurements, and the interlaboratory variability of Caco-2 measurements[37,38] makes it very difficult to compare data from multiple literature sources. Instead we have focused on a recent paper[12] where a limited number of diverse compounds were assayed in the same laboratory with the same protocol. The training set contains 13 structurally diverse, orally administered drugs for which the apparent permeability (P_{app}) was measured. The molecules were also carefully selected so that they show similar "apical to basolateral" P_{app} compared to "basolateral to apical" P_{app}. This provides a small data set with which to build a model of the passive component of intestinal absorption (diffusion through the intestinal barrier). Each molecule was described by 3 descriptors: MWeight, SlogP,[30] and radius.[39] A Partial Least Square regression was applied and a model was built with 2 components. The optimal number of components was selected here by a Leave-One-Out (LOO) cross-validation procedure. The Caco-2 membrane permeability model we obtained is

$$\text{Log } P_{app} = 0.49\ \text{SlogP} - 0.10\ \text{Radius} - 0.002\ \text{Weight} - 4.06$$

It is interesting to note that the above equation reflects the common sense assumption that a compound will diffuse more easily through a phospholipidic cell membrane if it is more lipophilic (SlogP with a positive coefficient) but not too big (negative coefficients for Weight and Radius parameters). The model is simpler than the one published by Bergstrom et al. [12] (3 descriptors vs 6 descriptors) but is as predictive in terms of the statistical parameters on the training, test, and external test set. The predictive ability of the model was assessed with a 10 molecule test set,[12] measured in the same laboratory as the training set and a 22 molecule external test set[40] from a different laboratory. The PLS model of Caco-2 membrane passive permeation we obtained is characterized in Table 4. The performance of our model on the training set is good with a correlation coefficient r^2 of 0.92 and a mean absolute error MAE of 0.30. Although this is expected with such a small training set, it should be stressed that only 3 descriptors were used. The predictive performance assessed by the leave-one-out cross-validation procedure is good, with a q^2 value of 0.81.

Table 4. Characteristics of the Intestinal Membrane Permeability 2D Model

	n	r^2	MAE	q^2
training set	13	0.92	0.30	0.81
test set	10	0.74	0.56	
external test set	22	0.53	0.86	

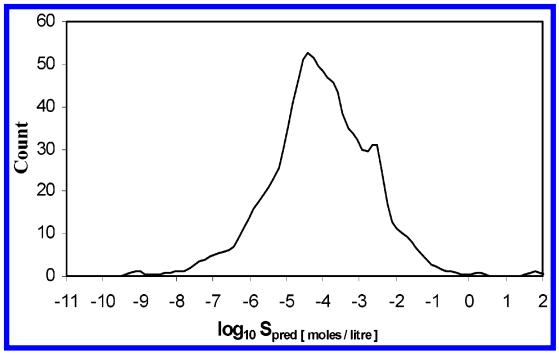

Figure 1. Distribution of the aqueous solubility predictions (moles/L), for the reference set of 1,141 small drug-like molecules.

Table 5. Analysis of the Reference Set of 1141 Drug-like Molecules When the Drug-like Filters Are Applied

	pass	fail
Lipinksi 3 rules out of 4	1128 (99%)	13
Lipinksi 4 rules out of 4	1037 (91%)	104
Veber 2 rules	1043 (91%)	98
MedChem-1 filter	1052 (92%)	89
MedChem-2 filter	930 (82%)	211
MedChem-1+MedChem-2 filter	841 (74%)	300
Solubility filter	1062 (93%)	79
Caco-2 permeation filter	1073 (94%)	68

The predictive performance on the first test set is satisfactory with an r^2 of 0.74. The predictive performance on the external test set is lower (r^2 of 0.53). This may reflect the interlaboratory variability of the caco-2 data, but the overall mean absolute error on these data set remains under 1 log unit.

Reference Set of Small Organic Drug-like Molecules. These were selected by applying the same selection criteria as Lipinski et al.[1] to the World Drug Index. A search for compounds having a United States Adopted Name (USAN) or International Nonproprietary Name (INN) identifies molecules that have passed preclinical stages and Phase I evaluation. An additional molecular weight filter ($250 \leq MW \leq 550$) and organic filter (SMILES code based) produced the final 1141 compounds.

Aqueous Solubility Drug-like Filter. Our aqueous solubility model predicts that the solubility of the reference molecules is between 7.25 pM and 215 M (-11.14 and $+2.33$ in log units). The distribution for this set is plotted in Figure 1. As can be seen in Table 5, most (93%) of the compounds are predicted to have solubility above 1 μM (-6 in log units), so this was taken as the cutoff to be applied as the solubility filter for drug-like molecules. This is not an unreasonable value as any drug candidate which was not

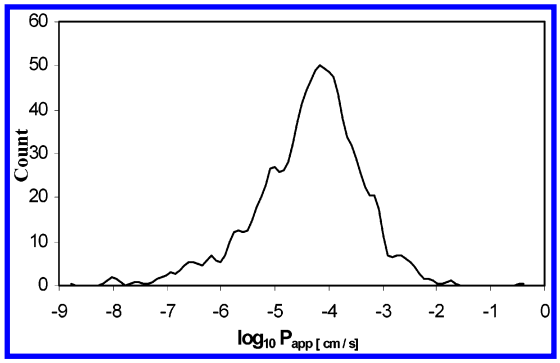

Figure 2. Distribution of the Caco-2 permeation predictions (cm/s) for the reference set of 1,141 small organic drug molecules.

water soluble at 1 μM would potentially be too problematic for oral absorption.

It could be argued that solubility is one of the properties that is routinely improved during the lead optimization process, and so the cutoff is too stringent for filtering compounds at the early stages of drug discovery, particularly hit identification. This has to be balanced with the desire for inclusion of at least some drug-like properties in lead compounds to allow scope for optimizing other properties. For the purposes of analyzing the compound suppliers' databases, we have maintained this 1 μM solubility cutoff.

Caco-2 Membrane Permeation Drug-like Filter. Figure 2 shows the distribution of Caco-2 membrane permeability predicted with our model for the reference set of compounds. The predictions (in log units) vary from a minimum of −8.76 (1.74×10^{-9} cm/s) to a maximum of −0.40 (0.40 cm/s) for the most permeable. As can be seen in Table 5, most (94%) of the permeation predictions for the 1141 reference molecules are above 10 nm/s (−6 in log units). This was therefore chosen as the cutoff to be used in analyzing the larger database of compounds. As for the solubility drug-like filter, we chose a Caco-2 permeation drug-like cutoff that should ensure a drug-like margin for borderline compounds.

RESULTS

Profiling of the Reference set of 1141 Drug-like Molecules. Table 5 summarizes how many of the reference drug-like molecules pass the drug-like filters. More than 90% of the reference molecules pass each of the individual drug-like filters, except for the MedChem-2 tractability filter flagging compounds containing what we consider potentially undesirable features.

Thirteen of the reference compounds do not fulfill at least 3 of the 4 Lipinski rules, while 104 do not fulfill 4 of the rules. The Veber drug-like filter has a very similar attrition rate (98 molecules fail). In general, the Veber filter discards molecules that are very flexible or polar, whereas the Lipinski 4-rules filters discards very large or hydrophobic compounds. Only 34 molecules in the reference set fail both Veber and Lipinski 4-rules filters being very large, flexible, and hydrophobic. It is interesting to note that 89 reference compounds are flagged by the "undesirable" MedChem tractability filter, while the "potentially undesirable" MedChem tractability filter has the highest attrition rate with 211 reference compounds failing. Altogether, some 300 marketed drugs are failing our MedChem tractability filters. Inspection of a representative set of compounds shows that these failures would be considered difficult compounds with which to begin a lead optimization program. For example, lindane (hexachlorocyclohexane) is a drug but if this appeared as a hit in a screen, there is little scope for medicinal chemistry on such a structure. Interestingly, the solubility and Caco-2 permeation filters select approximately the same percentage of compounds from the complete reference set (93% and 94%) as from the reduced MedChem-1+2 filtered subset (94% and 96%) emphasizing that, for this set of compounds, MedChem tractability does not correlate with solubility or membrane permeability.

Profiling of the 23-Supplier Database. The results of applying the Lipinski, Veber, Medchem, solubility, and Caco-2 drug-like filters to 2 675 022 compounds from 23 suppliers are summarized in Table 6. Each library of compounds was analyzed by the five drug-like filters. Results are reported as the number of compounds passing through the filter and the size of this subset (as a percentage) compared to the number of compounds in the library from each supplier.

It is not appropriate to discuss the details of the analyses for each individual supplier. However, some general trends can be seen. A high percentage of compounds pass the Caco-2 permeability filter, which indicates that most of the compounds are relatively small and lipophilic (with the expected exception of those libraries of compounds that are predominantly made up of natural products). This is also the reason a high proportion of compounds pass the Veber filter, as most of the compounds are relatively small and lipophilic while not being highly polar and not very flexible. However, a substantially higher percentage of compounds fail to satisfy at least 3 of the 4 Lipinski rules, and much more fail to satisfy all of the 4 rules. This can mainly be attributed to the generally high lipophilicity of the compounds. Libraries with many compounds failing at least one of the 4 Lipinski rules have a mean SlogP value above 4.3, whereas libraries with a much higher percentage of compounds fulfilling the 4 Lipinski rules have a mean SlogP value below 3.3. The same trend is seen in the solubility filter attrition rates, with a greater number of soluble compounds in libraries fulfilling all 4 Lipinski rules compared to those libraries fulfilling just 3 of the 4 rules. This high lipophilicity was commented on by Lipinski et al.[1] and associated with the prevalence of DMSO as a solvent for storage of compounds. It is interesting to see that 5 years after the Lipinski paper, this same tendency is still observed for some of the supplier libraries.

The final row of Table 6 summarizes the analysis of the total library of 2 675 022 compounds from all 23 suppliers. There are 1 622 763 unique structures. It is informative to compare the properties of this set of structures with the reference set of 1141 drug-like molecules.

The solubility drug-like filter flags a substantially higher percentage of unique structures compared to the reference set (41% vs 7%, respectively). In complete contrast, the Caco-2 permeation filter flags fewer structures in the unique set (1%) compared to the reference set (6%). This analysis highlights that the Solubility and Caco-2 permeation drug-like filters can be considered as complementary, flagging two extreme molecular profiles, either very insoluble/lipophilic or very impermeable/hydrophilic structures, respectively. The thresholds for the solubility and Caco-2 permeation filters were set from an analysis of the reference

Table 6. Analysis of the Drug-like Properties Computed for Compounds from 23 Suppliers[a]

	library size	Lip-3	Lip-4	Veber	MedChem-1	MedChem-2	MedChem-1 + MedChem-2	Solubility	Caco-2	all-filtered drug-like
AnalytiCon	765	572/75%	208/27%	148/19%	640/84%	745/97%	620/81%	650/85%	509/67%	76/10%
ChemOvation	1049	1016/97%	767/73%	1034/99%	1044/100%	1035/99%	1030/98%	824/79%	1044/100%	721/69%
Aurora	2245	2186/97%	1974/88%	2135/95%	1735/77%	2063/92%	1553/69%	1966/88%	2210/98%	1325/59%
Menai	3777	3709/98%	3239/86%	3725/99%	3121/83%	2636/70%	1980/52%	3079/82%	3771/100%	1518/40%
TOSLab	6669	5788/87%	4422/66%	6325/95%	5853/88%	4770/72%	3954/59%	4181/63%	6607/99%	2402/36%
MDPI	8734	8295/95%	6947/80%	8135/93%	6625/76%	6840/78%	4731/54%	6658/76%	8632/99%	3744/43%
Biofocus	10486	10227/98%	8755/83%	10415/99%	10253/98%	9987/95%	9754/93%	8733/83%	10219/97%	7635/73%
Bionet	34699	33870/98%	26317/76%	34449/99%	31944/92%	28982/84%	26227/76%	25541/74%	34598/100%	18410/53%
Maybridge	56405	55237/98%	45868/81%	55422/98%	49252/87%	44708/79%	37555/67%	43463/77%	56031/99%	27736/49%
Tripos	92920	89400/96%	74685/80%	89242/96%	79467/86%	82624/89%	69171/74%	74873/81%	92207/99%	50870/55%
MCL	105868	88674/84%	57397/54%	100209/95%	82922/78%	77783/73%	54837/52%	45531/43%	105589/100%	24693/23%
Vitas-M	113669	105944/93%	84035/74%	108939/96%	94494/83%	86634/76%	67459/59%	71909/63%	113027/99%	42402/37%
UKR	118426	99313/84%	70645/60%	106954/90%	106226/90%	95351/81%	83151/70%	59050/50%	118184/100%	33179/28%
Enamine	122859	112516/92%	85808/70%	117107/95%	117200/95%	112702/92%	107043/87%	75610/62%	122189/99%	59618/49%
TimTec	128181	119889/94%	96230/75%	120582/94%	105343/82%	92364/72%	69526/54%	84588/66%	126803/99%	45500/35%
IFLab	149862	140446/94%	111746/75%	143309/96%	134705/90%	119526/80%	104369/70%	94610/63%	148815/99%	62225/42%
Comgenex	164663	120301/73%	63807/39%	129847/79%	142080/86%	155855/95%	133272/81%	66409/40%	164478/100%	37396/23%
ChemT&I	165054	144080/87%	111375/67%	159219/96%	147320/89%	140990/85%	123256/75%	89638/54%	164663/100%	60489/37%
Asinex	202542	191589/95%	158556/78%	193433/96%	173521/86%	163233/81%	134212/66%	142074/70%	200517/99%	94092/46%
Specs	220558	204541/93%	157936/72%	212515/96%	190465/86%	177899/81%	147806/67%	135105/61%	219307/99%	88776/40%
IBS	283825	262741/93%	200287/71%	269413/95%	254317/90%	223889/79%	194381/68%	174145/61%	280122/99%	108803/38%
Chembridge	319537	314309/98%	262965/82%	311830/98%	275327/86%	264474/83%	220264/69%	229916/72%	317508/99%	154800/48%
ChemDiv	362229	336308/93%	261553/72%	343289/95%	315420/87%	280319/77%	233510/64%	224476/62%	359038/99%	138647/38%
Sum	2675022									
Unique	1622763	1450529/89%	1077731/66%	1511941/93%	1423152/88%	1341231/83%	1141620/70%	949652/59%	1611920/99%	607223/37%
WDI	1141	1128/99%	1037/91%	1043/91%	1052/92%	930/82%	841/74%	1062/93%	1073/94%	708/62%

[a] To facilitate reading of the table, the suppliers are sorted according to the size of their library. Lip-3 is the number of molecules fulfilling 3 out of 4 Lipinski rules, and Lip4 is the number of molecules fulfilling all 4 Lipinski rules. MedChem-1 and MedChem-2 columns contain the number of compounds passing these MedChem tractability filters. MedChem-1+MedChem-2 compounds pass MedChem-1 and MedChem-2 filters. The all-filtered drug-like column totals the compounds passing Lip-4 + Veber + MedChem-1 + MedChem-2 + Solubility + Caco-2 filters.

Table 7. Number of Filters Passed by the 1 622 763 Unique Structures

pass	number	pass	number
0	67	3	414 387
1	15 040	4	413 600
2	172 446	5	607 223

set, so not surprisingly the percentage of molecules failing from that set are low and similar for both filters (7% and 6%, respectively). The analyses demonstrate that the unique structures are in general much more lipophilic, with almost all structures passing through the Caco-2 filter (99%) and much fewer passing the solubility filter (59%).

This difference in lipophilicity also explains differences in the Lipinski and Veber profiles between the unique set of supplier-available structures and the reference set of drug-like molecules. 11% of the unique structures do not satisfy 3, and 34% do not satisfy 4 of the Lipinski rules. However, only 7% fail the Veber criteria. This contrasts with the analysis of the reference set of molecules where essentially the same number of compounds failed both the sets of rules. The Veber filter does not penalize extreme lipophilicity, wherease Lipinski filters specifically address it through the SlogP cutoff.

Table 7 provides a breakdown of how many of the 1 622 763 unique structures pass a particular number of filters. This emphasizes that although some of the filters are based on the same molecular property (such as molecular weight, or SlogP), combinations of the filters are identifying a specific molecular profile forbidden to drugs. The 67 structures that fail all filters are large, flexible lipophilic structures that contain undesirable chemical features such as metals, sugars, or disulfide bonds. Table 8 analyses which

Table 8. Breakdown of Which Drug-like Filter Is Passed by the 15 040 Unique Structures Passing Just One Filter[a] and Which Filter Is Failed by the 413 600 Unique Structures Passing Four Filters[b]

filter type	number of pass-1 set passing[a]	number of pass-4 set failing[b]
Lip-4	1	29 437
Veber	3	19 129
Solubility	875	108 267
Caco-2	14 106	4740
MedChem-1 + MedChem-2	55	252 027

particular filter is passed by the 15 040 unique structures passing just one filter. These structures are again generally large, lipophilic compounds with undesirable chemical groups, but within the molecular weight limit established within the Caco-2 filter. Table 8 also analyses which filter is failed by the structures passing four of the drug-like filters. This analysis reiterates that there are a high number of structures of predicted low solubility. What is striking is the large number of structures (252 027 out of 413 600) that fail only on our MedChem-1 and MedChem-2 medicinal chemistry tractability filters.

Duplicate Analysis of the 23-Suppliers Database. The unique set of structures discussed above was derived from a duplicate analysis of the total database of compounds from all the suppliers. Table 9 analyzes the database of 2 675 022 compounds in terms of how many of the compounds are exclusive to a particular supplier and how many of these exclusive compounds are drug-like according to our filters. For example, the database contains 202 542 compounds from Asinex. Of these, 40 168 compounds were not available from any of the other 22 suppliers, of which 12 557 passed all the drug-like filters. Considering the 23 suppliers all together and unfiltered, 40% (1 077 001) of the initial 2 675 022

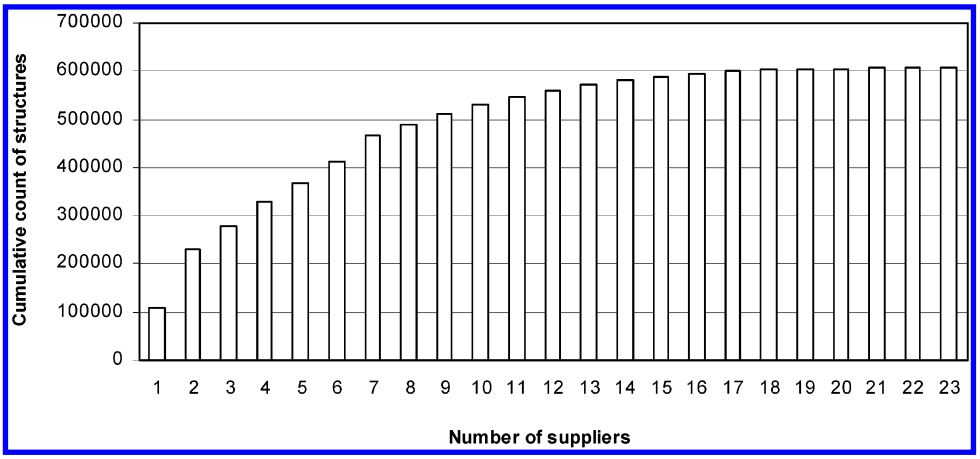

Figure 3. Plot of the cumulative number of all-filtered drug-like compounds obtained when adding suppliers in descending order of the number of exclusive structures (data in Table 9).

Table 9. Exclusivity Analysis of the Compounds Available from the 23 Suppliers[a]

	library size	unfiltered − exclusives	all-filtered drug-like − exclusives
AnalytiCon	765	765 (100%)	76 (10%)
ChemOvation	1049	1007 (96%)	704 (67%)
Aurora	2245	1620 (72%)	1006 (45%)
Menai	3777	2695 (71%)	970 (26%)
TOSLab	6669	3907 (59%)	1361 (20%)
MDPI	8734	7104 (81%)	2954 (34%)
Biofocus	10 486	10 485 (100%)	7635 (73%)
Bionet	34 699	32 162 (93%)	16 723 (48%)
Maybridge	56 405	42 865 (76%)	20 184 (36%)
Tripos	92 920	76 211 (82%)	42 923 (46%)
MCL	105 868	69 535 (66%)	11 737 (11%)
Vitas-M	113 669	17 297 (15%)	5858 (5%)
UKR	118 426	71 826 (61%)	17 876 (15%)
Enamine	122 859	83 291 (68%)	42 864 (35%)
TimTec	128 181	20 030 (16%)	5 619 (4%)
IFLab	149 862	16 906 (11%)	7 559 (5%)
Comgenex	164 663	162 504 (99%)	36 037 (22%)
ChemT&I	165 054	61 383 (37%)	19 269 (12%)
Asinex	202 542	40 168 (20%)	12 557 (6%)
Specs	220 558	61 106 (28%)	21 133 (10%)
IBS	283 825	126 303 (45%)	44 781 (16%)
Chembridge	319 537	98 875 (31%)	44 281 (14%)
ChemDiv	362 229	68 956 (19%)	20 487 (6%)
sum	2 675 022	1 077 001	384 594

[a] See text for the definition of the columns.

Table 10. Count of the Unfiltered and All-Filtered Drug-like Structures Shared by Multiple Suppliers[a]

unfiltered		all-filtered drug-like	
no. of suppliers	no. of shared structures	no. of suppliers	no. of shared structures
1[a]	1 077 001	1[a]	384 594
2	273 665	2	100 507
3	137 911	3	58 816
4	72 047	4	32 883
5	36 498	5	17 679
6	16 567	6	8091
7	6499	7	3221
8	1970	8	1040
9	462	9	289
10	112	10	79
11	25	11	18
12	6	12	6
sum	1 622 763	sum	607 223

[a] Equivalent to exclusive structures.

compounds are available from only one supplier, of which 384 594 (only 14% of the initial 2 675 022 compounds) pass all five drug-like filters.

Table 10 analyses how many of the structures are shared between several suppliers, for both the unfiltered and all-filtered drug-like structures. The first striking observation is that there are no structures that are shared by more than 12 suppliers out of the 23 considered here. Second, nearly 90% of the structures, either unfiltered or filtered, are available from only 1, 2, or 3 suppliers (1 077 001 + 273 665 + 137 911 = 1 488 577 = 90% of 1 622 763). This raises the question of how many suppliers need to be considered to optimize access to most of the drug-like compounds?

Figure 3 shows the optimized cumulative count of all-filtered drug-like structures obtained when considering an increasing number of suppliers. Thirteen suppliers need to be considered before getting access to more than 90% of the all-filtered drug-like 607 223 structures. There is, however, still benefit in considering additional suppliers, as, for our analysis, the last 10 suppliers provide access to 34 035 new all-filtered drug-like structures.

CONCLUSION

In this paper, we have demonstrated how a range of in silico drug-like filters can be developed and used to profile large multimillion compound catalogues from suppliers. The analysis shows that out of the 2.7 million compounds available from the 23 compound suppliers, there are slightly more than 1.6 million unique structures of which 607 223 (37%) pass all the Vernalis drug-like filters. The filters have revealed a tendency toward high lipophilicity in many of the compound libraries. As expected, the proportion of unique structures found from considering an additional supplier falls as the number of suppliers considered increases. However for the set of suppliers considered, no structure was available from more than 12 suppliers. This combination of analyses emphasizes the importance of considering as many suppliers as feasible.

For our analyses, we have developed and implemented a range of drug-like filters, then validated, and further calibrated the filters against drug-like molecules extracted from the World Drug Index. The accuracy of the solubility and Caco-2 permeability prediction appears to be within 1 log unit, which is adequate for the types of compound selection decisions being made. A range of chemical tractability/

desirability filters have been developed, encoding the experience and prejudice of the medicinal chemists within our company. This list is necessarily subjective and incomplete and as the number of structures marked as non-drug-like on this criterion alone is quite large (252 027, see Table 8), the set of filters does need constant challenging and updating.

Extensions of the work presented in this article concern the modeling of other specific components of drug absorption, such as cytochrome-mediated metabolism[39] or PgP efflux.[40]

An important consideration is how these filters are to be used. The main use for the large compound libraries analyzed here is in the early stage of the discovery process, where novel hits are identified by virtual or experimental screening and initial SAR and physicochemical properties established for a compound series prior to choosing lead compounds. The transition from hit to lead is an important step, committing considerable effort (medicinal chemistry, assays, etc.) to a compound. The choice of which properties need to be optimized in a compound during lead optimization varies dramatically according to the compound series and target (and between research teams). There are considerable challenges in making decisions based on lead-like properties as the improvement of compounds during lead optimization depends on many different factors, most crucial of which is the chemical tractability of the compounds and the impact of computational and structural information in guiding where and how to modify the lead to improve particular properties. So, although it would be unusual to apply all the drug-like filters discussed here in selecting compounds to be screened as potential hits, this analysis does provide insight into the extent of the drug-like properties in the available compounds.

There are numerous other applications in drug discovery for the combination of in silico prediction methods described here. The calculations could be used to profile virtual libraries of compounds that can be made using available reagents with a minimal number of synthetic steps. In addition, the thresholds used in the various filters could be reduced to allow profiling of more lead-like properties to design sets of compounds that can be used for hit generation. Finally, the methods can be used iteratively in the lead optimization process to guide the design of improved properties into compounds.

The analyses performed here do not include other critical supplier characteristics when using such libraries as a source of potential hits, such as the purity, the price, the time for delivery, and the real availability of the compounds. These parameters need to be combined with the duplicate and drug-like analyses into a multisupplier database that can efficiently impact the drug design process.

ACKNOWLEDGMENT

We thank Jarmo Huuskonen from the University of Helsinki, Finland, for providing the solubility data set of 1280 compounds.

REFERENCES AND NOTES

(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) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *Mol. Divers.* **2002**, *5*, 199−208.
(3) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(4) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882−894.
(5) van de Waterbeemd, H. G. E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192−204.
(6) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.
(7) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, 2615−2623.
(8) Oprea, T. I.; Gottfries, J.; Sherbukhin, V.; Svensson, P.; Kuhler, T. C. Chemical information management in drug discovery: optimizing the computational and combinatorial chemistry interfaces. *J. Mol. Graph. Model.* **2000**, *18*, 512−524.
(9) Ajay, A. W. W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(10) Frimurer, T. M.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunak, S. Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.
(11) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening − an overview. *Drug Discov. Today.* **1998**, 160−178.
(12) Bergstrom, C. A. S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P. Current Absorption Classification of Oral Drugs Based on Molecular Surface Properties. *J. Med. Chem.* **2003**, *46*, 558−570.
(13) Wright, T. G. V. J.; Green, D. V.; Pickett, S. D. Optimizing the size and configuration of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 381−390.
(14) Alanine, A. N. M.; Roberts, E.; Thomas, A. W. Lead generation—enhancing the success of drug discovery by investing in the hit to lead process. *Comb. Chem. High Throughput Screen.* **2003**, 51−66.
(15) Voigt, J. H.; Bienfait B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(16) Mozziconacci, J.-C. A. E.; Baurin, N.; Marot, C.; Morin-Allory, L. Preparation of a molecular database from a set of 2 million compounds for virtual screening applications: gathering, structural analysis and filtering. *Electronic Communication, 9th Electronic Conference of Computational Chemistry ECCC9, April, 2003. http://www.univ-orleans.fr/SCIENCES/ICOA/eposter/eccc9/ECCC9i.htm.*
(17) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, U.S.A. http://www.mdl.com..
(18) Weininger, D. SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(19) Wall, L. C. T.; Orwant, J. In *Programming Perl*; O'Reilly, T., Ed.; O'Reilly & Associates: Sebastopol.
(20) Feuerstein, S. P. B. In *Oracle PL/SQL*; O'Reilly, T., Ed.; O'Reilly & Associates: Sebastopol, 1997.
(21) Wegner, J. K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−84.
(22) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−17.
(23) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. AQUADAC 3: Aqueous Functional Group Activity Coefficients: Application to the estimation of aqueous solubility. *Chemosphere* **1995**, *30*, 1619−1637.
(24) Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 741−52.
(25) Huuskonen, J.; Rantanen, J.; Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **2000**, *35*, 1081−8.
(26) Gao, H.; Shanmugasundaram, V.; Lee, P. Estimation of aqueous solubility of organic compounds with QSPR approach. *Pharm. Res.* **2002**, *19*, 497−503.
(27) Butina, D.; Gola, J. M. Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837−41.
(28) Physical Properties Database (PHYSPROP). Syracuse Research Corporation. http://esc.syrres.com/interkow/PhysProp.htm..
(29) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(30) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(31) 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.
(32) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219.
(33) Molecular Operating Environment (MOE). 2002.03. Chemical Computing Group Inc., Montreal, H3A 2R7 Canada, http://www.chemcomp.com.
(34) Klopman, G.; Zhu, H. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439−445.
(35) Bush, B.; Nachbar, R. B. Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 587−619.
(36) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(37) Artursson, P. P. K.; Luthman, K. Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Adv. Drug Delivery Rev.* **1996**, *22*, 67−84.
(38) Egan, W. J.; Lauri, G. Prediction of intestinal permeability. *Adv. Drug Deliv. Rev.* **2002**, *54*, 273−289.
(39) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.
(40) Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for the prediction of intestinal drug absorption. *J. Med. Chem.* **2001**, *44*, 1927−1937.

CI034260M